Klassifikationsverfahren der Diskriminanzanalyse

Eine vergleichende und integrierende Übersicht

— Diplomarbeit —

Überarbeitete Fassung

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Mathematik

eingereicht von

MICHAEL NOTHNAGEL
geboren am 22. Juli 1971
in Berlin

betreut durch
DR. JÖRG POLZEHL
und
PROF. DR. OLAF BUNKE

Berlin, den 05. Oktober 1999

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Diskriminanzanalyse ist eine Klasse von statistischen Verfahren, deren Ziel es ist, Personen oder Objekte aufgrund ihrer Merkmale und Eigenschaften einer von mehreren a-priori festgelegten Klassen, Populationen oder Kategorien zuzuordnen oder die für solch eine Zuordnung wichtigsten Merkmale zu finden. Die Zuordnungsregel wird von einer Stichprobe schon klassifizierter Objekte bzw. Personen abgeleitet. Dieser allgemeine Zugang ermöglicht die Anwendung von Diskriminanzanalyseverfahren in sehr verschiedenen Bereichen praktischer Problemstellungen, z.B. in der Medizin (Diagnostik), der Biologie (Systematik, automatisiertes Auszählen von Kolonien auf Kulturschalen), der industriellen Fertigung (Qualitätskontrolle), bei Sicherungssystemen, der Schrifterkennung oder der Militäraufklärung (Objekterkennung). Dabei ist der statistische Charakter der Diskriminanzanalyse zu beachten, der nur Rückschlüsse in Wahrscheinlichkeiten oder für größere Populationen als Ganzes zuläßt. Entscheidungen zu einzelnen Personen aufgrund einer Diskriminanzregel, z.B. in der Medizin, stehen daher immer unter einem ethischen Vorbehalt.

Das Ziel dieser Arbeit ist es, die einzelnen Verfahren mit ihren Annahmen, Eigenschaften und Modifikationen vorzustellen, um so einen vergleichenden und integrierenden Überblick zu geben. Die Motivation geht dabei weniger von der theoretischen Seite aus als von einer praktischen Fragestellung: *Welche Verfahren kann ich zur Analyse meiner Daten verwenden?* Die Arbeit legt daher den Schwerpunkt nicht auf die Herleitung der Verfahren, asymptotische Resultate oder theoretische Optimalitätskriterien, sondern anwendungsorientiert auf die Vorstellung der Voraussetzungen und wesentlichen Ideen der Verfahren, auf Stichproben mit begrenztem Umfang und Erfahrungen von vergleichenden Studien bei simulierten und realen Daten. Ebenfalls aus praktischen Gründen werden nur Verfahren vorgestellt, die bereits weithin verwendet werden oder von denen dies aufgrund ihrer Eigenschaften in nächster Zukunft zu erwarten ist; d.h. Verfahren, die in einem gewissen Sinne allgemein verfügbar und auf eine Vielzahl von Problemstellungen anwendbar sind. Ich hoffe, daß diese Arbeit durch ihren heuristischen Schwerpunkt auch für interessierte Naturwissenschaftler und Mediziner verständlich ist. Für weiterführende Einblicke sei auf die Referenzen und Quellen verwiesen.

Zum einfacheren Verständnis konzentriert sich die Arbeit auf den Fall zweier Klassen. Die meisten Verfahren sind ohne weiteres auf mehrere Klassen erweiterbar; oft durch wiederholten Rückgriff auf zwei Klassen. Kapitel 6 bietet dazu einen kurzen Überblick.

Die Arbeit ist inhaltlich fortlaufend gegliedert. Das Kapitel 2 behandelt die statistischen Grundlagen der Diskriminanzanalyse, die Art des Fehlers in diesem Kontext, Ansätze für Schätzungen der Diskriminanzregel und ihres Fehlers, das Problem der Modellwahl sowie kurz Fragen der Datenbehandlung. In den Kapiteln 3 und 4 werden schon klassische Verfahren, unterschieden nach parametrischen und nichtparametrischen, mit ihre Eigenschaften und Erweiterungen erläutert. Die mit einem „*“ gekennzeichneten Teilkapitel beschreiben Regularisierungen und Modifikationen der Verfahren und können für einen ersten Überblick übersprungen werden. Kapitel 5 enthält eine kurze Vorstellung neuerer Ideen für nichtparametrische Verfahren, vornehmlich aus den 1990er Jahren, die durch die Verfügbarkeit immer größerer Rechenleistungen die Substanz zu einer Verbesserung klassischer Verfahren oder zu neuartigen Verfahren besitzen. Das Kapitel 6 erläutert einige Ansätze, die vorgestellten Verfahren auf Probleme mit mehr als zwei Klassen anzuwenden. Zum Abschluß werden im Kapitel 7 einige theoretisch vorgestellte Verfahren auf simulierte und reale Daten angewendet, um verschiedene theoretische Aussagen zu veranschaulichen. Kapitel 8 faßt Empfehlungen für die praktische Anwendung zusammen.

Die Literaturstellen sind — in Mathematik und Biologie & Medizin unterschieden und alphabetisch sortiert — fortlaufend numeriert; die Bezugnahme auf sie erfolgt über ihre Nummern, also z.B. „[161]“ für ein Buch von McLachlan aus dem Jahre 1992. Die Analysen in Kapitel 7 wurden in der statistischen Programmiersprache S-PLUS gerechnet (siehe z.B. [218]); der Quellcode ist im Anhang A dokumentiert. Im Anhang B sind die benutzten Symbole in ihrer häufigsten Bedeutung aufgeführt; außerdem findet sich dort eine Liste häufig benutzter Abkürzungen und Verfahrenskürzel.

Diese Arbeit wurde in LaTeX 2_ε gesetzt ([136]).

Sehr herzlich möchte ich mich bei Dr. Jörg Polzehl[1] und bei Prof. Dr. Olaf Bunke[2] für die vielen Hinweise, Erläuterungen und die Geduld bei der Betreuung meiner Diplomarbeit bedanken. Weiterer Dank gilt Prof. Dr. Herbert Schuster[3] für die materielle Unterstützung meiner Arbeit.

Berlin, den 05. Oktober 1999 Michael Nothnagel

Dies ist eine nachträglich überarbeitete und korrigierte Fassung der eingereichten Diplomarbeit nach Hinweisen des Gutachters, Dr. Polzehl, vom 9. November 1999.

[1] Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS) Berlin

[2] Humboldt-Universität zu Berlin, Institut für Mathematik

[3] Max-Delbrück-Centrum für Molekulare Medizin Berlin-Buch und INFOGEN Medizinische Genetik GmbH

Kapitel 2

Grundlagen der Diskriminanzanalyse

In diesem Kapitel werden die wahrscheinlichkeitstheoretischen Grundlagen der Diskriminanzanalyse dargestellt und die grundlegenden Schätzmethoden für Regeln und deren Fehler vorgestellt. Im weiteren wird die besondere Art des Fehlers im Diskriminanzkontext untersucht. Nach Betrachtungen zu Datentypen und -transformationen und zur Modellwahl werden die Erfahrungen daraus in einige Leitsätze destilliert.

2.1 Regel & Risiko

Allgemeines Ziel der Diskriminanzanalyse[1] ist die Ableitung einer Zuordnungs- oder Allokationsregel. Objekte, Beobachtungen oder Personen sollen durch diese Regel genau einer von mehreren *vorher* (a-priori) bekannten Klassen oder Populationen zugeordnet werden. Sie sind dabei durch Merkmale wie Messungen, Eigenschaften oder Kenndaten charakterisiert, mit deren Hilfe die Zuordnung erfolgt. Die Diskriminanzanalyse bemüht sich weder um die *Errichtung* von Klassen im Nachhinein (a-posteriori), wie das bei der Clusteranalyse der Fall ist (siehe etwa [61]), noch liefert sie eine Wahrscheinlichkeitsverteilung als Ergebnis.

Eine Person oder ein Objekt ist durch einen Vektor $x = (x_1, \ldots, x_p)' \in \Omega \subseteq \mathbb{R}^p$ repräsentiert, der die Ausprägungen in den p Merkmalen enthält. Die allgemeine Form einer Diskriminanzregel lautet:

Definition 2.1 (Diskriminanzregel) *Es seien $k \geq 2$ Populationen $\Pi_1, \ldots, \Pi_k$ auf Wahrscheinlichkeitsräumen mit einem gemeinsamen Grundraum und unterschiedlichen Verteilungen in den Objektmerkmalen, $\{(\Omega, \mathcal{A}, P_i)\}_{i=1,\ldots,k}$, gegeben. Eine Diskriminanzregel ist eine Abbildung*

$$\delta : \Omega \to \{1, \ldots, k\}, \tag{2.1}$$

die ein Objekt aus Ω genau einer Population Π_i zuordnet.

Zu Grundlagen der Maß- und Wahrscheinlichkeitstheorie siehe [13], [12]. Die Klassenzugehörigkeit einer Person oder eines Objekts kodiert oft eine Variable z:

$$z(x) = i \iff x \in \Pi_i.$$

[1] Fachgebiete mit ähnlichen Zielstellungen oder Ansätzen sind z.B. *Machine Learning*, insbesondere *Supervised Learning*, und *Pattern Recognition*; siehe [181] und [162] für Referenzen.

Auch $z_{ij} = 1 \iff x_i \in \Pi_j$ und 0 sonst ist üblich. Der Wert der Variable $z(x)$ ist nur für die Objekte der Trainingsstichprobe bekannt, ansonsten ist es eine Zufallsvariable.

Die eindeutige Zuordnung eines Objekts zu einer Population durch eine Regel δ bewirkt eine Zerlegung des Merkmalsraumes Ω in disjunkte Teilmengen $\Omega_1, \ldots, \Omega_k$, so daß δ alle Beobachtungen aus Ω_i der Klasse Π_i zuordnet. Es ist dabei zu beachten, daß Ω_i und $\operatorname{supp} \Pi_i$ im allgemeinen *nicht* übereinstimmen!

Man kann zwei Zielstellungen mit einer Diskriminanzanalyse verfolgen:

- *Vorhersage*: Mit der Diskriminanzregel sollen zukünftige, unklassifizierte Beobachtungen einer der Klassen oder Populationen zugeordnet werden (Klassifikation).
- *Interpretation*: Durch die Diskriminanzanalyse sollen trennende Merkmale zwischen den Populationen gefunden und nach Möglichkeit eine Auswahl getroffen werden (Dimensionserniedrigung, Diskrimination).

Hand [105] gibt eine leicht verständliche Einführung in die Diskriminanzanalyse, während McLachlan [161] einen strenger mathematischen Überblick über den Erkenntnisstand bis 1992 bietet.

Beispiel 2.1 *Betrachtet man Alter und systolischen Blutdruck bei Personen, um deren Gefährdung für einen Schlaganfall zu fassen, hat der theoretische Merkmalsraum mit $p = 2$ die Gestalt $\Omega = (0, \infty) \times (0, \infty)$, wobei die ∞ sicherlich biologisch beschränkt sind. Interessiert man sich für die Populationen Π_1={Personen mit geringer Gefahr} und Π_2={Personen hoher Gefahr}, d.h. $k = 2$, und stellt die (stark vereinfachende) Regel*

$$\delta : [\textit{Alter} \geq \textit{60 Jahre und Blutdruck} \geq \textit{160 mmHg}] \Rightarrow [\textit{höhere Gefahr}]$$

auf, erhält man als Zerlegung

$$\begin{aligned} \Omega_1 &= (0, 60) \times (0, 160) \cup (0, 60) \times [160, \infty) \cup [60, \infty) \times (0, 160) \\ \Omega_2 &= [60, \infty) \times [160, \infty), \end{aligned}$$

d.h. alle Personen, die mit ihren Meßdaten in Ω_1 liegen, werden durch δ als gering gefährdet eingestuft.

Die Populationen $\Pi_1, \ldots, \Pi_k$ können einen unterschiedlichen Anteil am Merkmalsgrundraum Ω besitzen:

Definition 2.2 (A-priori-Verteilung) *Die Wahrscheinlichkeiten $\pi_1, \ldots, \pi_k$ für das Auftreten einer Beobachtung aus den einzelnen Populationen $\Pi_1, \ldots, \Pi_k$ heißen A-priori-Verteilung. Es gilt: $\sum_{i=1}^{k} \pi_i = 1$.*

So läßt sich beispielsweise die Wahrscheinlichkeit für das Auftreten einer Krankheit in der Bevölkerung modellieren. Wichtig ist der Bezugsrahmen für die Analyse, z.B. ob man die relative Krankheitshäufigkeit in der Gesamtbevölkerung betrachtet oder nur in einer Teilpopulation, die aufgrund von möglichen Symptomen einen Arzt aufsucht.

Im Normalfall werden sich die Supports der Populationsverteilungen überlappen ($\exists i \neq j : \operatorname{supp} \Pi_i \cap \operatorname{supp} \Pi_j \neq \emptyset$), d.h. Objekte aus verschiedenen Populationen werden in derselben Region auftreten.[2]

[2] Ansätze aus der *Machine Learning*-Theorie gehen im Gegensatz dazu von einem disjunkten Support der einzelnen Klassen aus ($\forall i \neq j : \operatorname{supp} \Pi_i \cap \operatorname{supp} \Pi_j = \emptyset$).

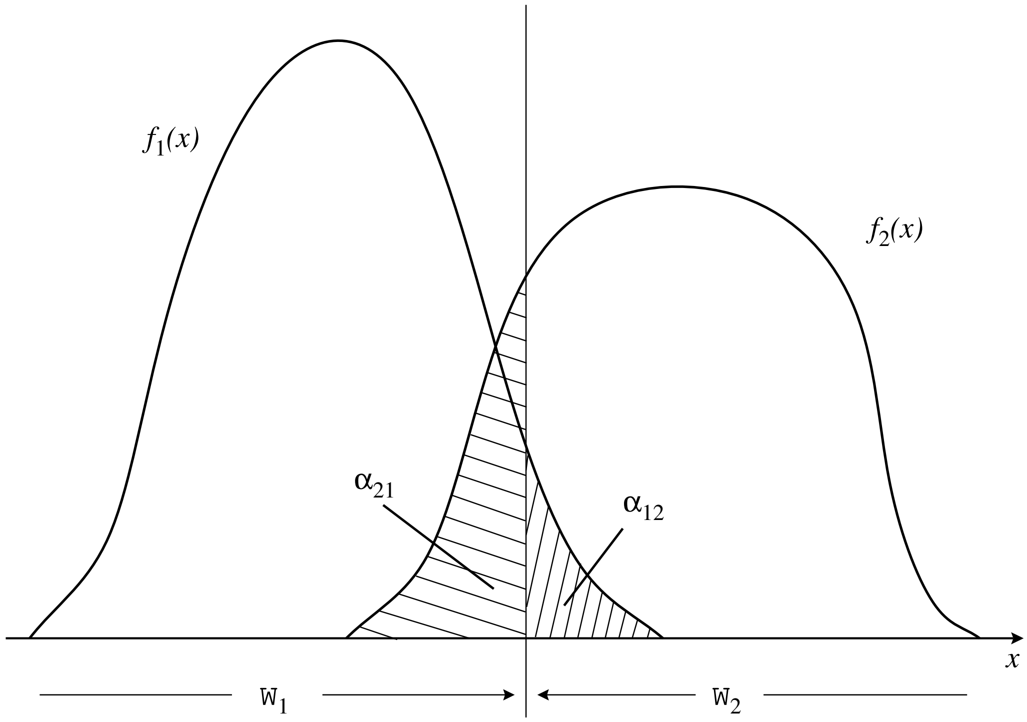

Abbildung 2.1: Beispiel einer Zerlegung von Ω für $k = 2$ und $p = 1$. Die schraffierten Flächen stellen als Integrale der klassenspezifischen Dichten im Zuordnungsgebiet der jeweils anderen Klasse die Fehlklassifikationswahrscheinlichkeiten dar. (Abbildung in Anlehnung an [131])

Beispiel 2.2 *Das Beispiel 2.1 weitergedacht, wird es bei einem Patientenkollektiv mit und einer Kontrollgruppe ohne Schlaganfälle in beiden Gruppen Personen mit ähnlichen Blutdruck- oder Blutfettwerten geben, so daß nicht ohne weiteres von einem gewissen, „normalen“ Meßwert auf die Gefährdung geschlossen werden kann.*

Die aus der Überlappung resultierenden Wahrscheinlichkeiten für die Fehlzuordnung eines Objekts durch eine Regel sind wie folgt bestimmt:

Definition 2.3 (Klassifikationswahrscheinlichkeiten) *Die Wahrscheinlichkeit α_{ij}, ein Objekt aus der Population Π_i durch eine Regel δ gemäß (2.1) der Population Π_j zuzuordnen $(\Pi_i \to \Pi_j)$, ist definiert als*

$$\alpha_{ij} := \int_{\Omega_j} dP_i(x) \quad (\Pi_i \to \Pi_j) \qquad (i, j = 1, \ldots, k). \tag{2.2}$$

Für $i \neq j$ heißen die α_{ij} Fehlklassifikationswahrscheinlichkeiten (FKW).

Für kleine α_{ij} separiert die Regel die Populationen relativ stark, kann sie also gut unterscheiden, während sie für größere im Mittel einen beträchtlichen Anteil an Fehlzuordnungen aufweisen wird (siehe Abbildung 2.1).

Beispiel 2.3 (Medizinische Diagnostik) *Falls Π_1 die Population der Erkrankten und Π_2 die der Gesunden bezeichnet, tragen die α_{ij} spezielle Namen: α_{11} heißt Sensitivität, α_{22} Spezifität; α_{12} ist die Wahrscheinlichkeit für einen falsch negativen Befund, α_{21} die für einen falsch positiven ([161], Kap. 1.3).*

Fehlklassifikationen können zusätzlich mit *Kosten* c_{ij} $(\Pi_i \to \Pi_j)$ belegt sein, um Zuordnungsfehler unterschiedlich zu bewerten oder realen Kosten einer Allokation Rechnung zu tragen ([220]). Die Gesamtkosten $\mathcal{R}_i$ für die (Fehl-) Klassifikation von Beobachtungen aus Π_i in die einzelnen Klassen ergeben sich dann als Summe der Produkte aus Kosten und der Wahrscheinlichkeit einer Fehlallokation, die diese Kosten verursacht:

$$\mathcal{R}_i(\delta) := \sum_{j=1}^{k} c_{ij}\alpha_{ij} \quad (i = 1, \ldots, k). \tag{2.3}$$

Meist werden die Kosten für eine korrekte Allokation $(\Pi_i \to \Pi_i)$ auf Null gesetzt $(c_{ii} = 0)$. Gilt $c_{ij} = c_{ji}$, spricht man von symmetrischen Kosten.

Beispiel 2.4 (Qualitätskontrolle) *Für die Endkontrolle über die Qualität eines Produktes mittels einer Diskriminanzregel könnte man die Herstellungskosten für ein ausgesondertes, aber qualitativ genügendes Produkt den Kosten einer Garantieerfüllung gegenüberstellen.*

Beispiel 2.5 (Medizinische Ergebnisse) *Kosten können auch benutzt werden, um falsch positive bzw. falsch negative Analyseergebnisse mit großer Sicherheit auszuschließen; dabei jeweils unter Inkaufnahme einer Vergrößerung des anderen Fehlers. Möchte man etwa unter allen Umständen sicherstellen, daß Mitteilungen an Patienten über einen positiven Gentest oder einen kritischen histologischen Befund der Realität entsprechen, so wird man die Kosten für eine Allokation eines unbedenklichen Befundes zu der eines bedenklichen hoch ansetzen. Man nimmt dafür aber in Kauf, daß dadurch mehr bedenkliche Befunde als unbedenklich eingestuft werden.*

Summiert man nun die klassenspezifischen Kosten gewichtet nach dem Anteil der Populationen, so erhält man das sogenannte BAYESsche Risiko[3] der Regel δ:

Definition 2.4 (BAYESsches Risiko)

$$\mathcal{R}(\delta) := \sum_{i=1}^{k} \pi_i \mathcal{R}_i(\delta)\,. \tag{2.4}$$

Eine Diskriminanzregel, die diese mittleren Klassifikationskosten zu gegebener A-priori-Verteilung minimiert, heißt BAYES*sche Regel*. Werden korrekte Allokationen nicht bestraft und gilt $c_{ij} \equiv c \quad (i, j = 1, \ldots, k \quad j \neq i)$, so wird nur die gewichtete relative *Anzahl* der Fehlklassifikationen bestimmt.

Kommentar 2.1 (BAYESsches Risiko für zwei Klassen) *Im Fall $k = 2$ vereinfacht sich (2.4) zu*

$$\mathcal{R}(\delta) := \pi_1 \underbrace{c_{12}\alpha_{12}}_{\mathcal{R}_1(\delta)} + \pi_2 \underbrace{c_{21}\alpha_{21}}_{\mathcal{R}_2(\delta)}\,. \tag{2.5}$$

Wie sieht nun die BAYESsche Regel aus?

[3] in der Literatur auch als *mittlere Klassifikationskosten*, *erwartete Kosten*, *expected overall loss* oder *expected cost of misclassification (ECM)* bezeichnet

Theorem 2.1 (BAYESsche Regel) *Es seien $f_1, \ldots, f_k$ als die Dichten zu den Verteilungen $P_1, \ldots, P_k$ gegeben. Falls für eine Diskriminanzregel δ die Zerlegung von Ω die Gestalt*

$$\Omega_i = \left\{ x : \sum_{l=1}^{k} \pi_l c_{li} f_l(x) \leq \sum_{l=1}^{k} \pi_l c_{lj} f_l(x) \quad (\forall j \neq i) \right\} \quad (i = 1, \ldots, k) \tag{2.6}$$

annimmt, so wird das BAYESsche Risiko

$$\mathcal{R}(\delta) = \sum_{i=1}^{k} \sum_{j=1}^{k} \pi_i c_{ij} \int\limits_{\Omega_j} f_i(x)\, dx \tag{2.7}$$

für diese Regel minimal. Diese optimale Diskriminanzregel heißt BAYESsche Regel und wird mit δ_0 bezeichnet (siehe [10], Kap. 6.7).

Falls die Verteilungen11 $P_1, \ldots, P_k$ bekannt sind, ist die BAYESsche Lösung also explizit bestimmt. Sind die zugehörigen Dichten stetig, so ist die BAYESsche Lösung bis auf Nullmengen eindeutig. Der BAYESsche Ansatz schließt durch eine geeignete Wahl der A-priori-Wahrscheinlichkeiten den NEYMAN-PEARSON-Ansatz[4] und den *Minimax*-Ansatz[5] mit ein. Die Menge der BAYESschen Regeln ist außerdem gleich der Menge der zulässigen Regeln (d.h. sie sind nicht mehr gleichmäßig über alle $x \in \Omega$ zu verbessern), falls alle A-priori-Wahrscheinlichkeiten größer als Null sind ([10], Kap. 6.7). Deswegen dient die BAYESsche Regel als Vergleichsmaßstab, an der sich formulierte Regeln messen lassen müssen.

Beispiel 2.6 *Im Falle zweier Klassen mit univariaten Normalverteilungen $N(-1, 1)$ und $N(+1, 1)$ als Populationsverteilungen ergibt sich das durchschnittliche Risiko der besten Diskriminanzregel bei $\pi_1 = \pi_2 = \frac{1}{2}$ und $c_{12} = c_{21} = 1$ als $\Phi(-1) = 0.1586$ ([89]). Damit wird* jede *Diskriminanzregel im Durchschnitt mindestens 15,9 % aller neuen Beobachtungen falsch klassifizieren.*

Kommentar 2.2 (BAYESsche Regel für zwei Klassen) *Für den in dieser Arbeit betrachteten Fall zweier Klassen ($k = 2$) wird die BAYESsche Regel (2.6) bei $c_{ii} = 0$ zu*

$$\Omega_1 = \{x : \pi_2 c_{21} f_2(x) \leq \pi_1 c_{12} f_1(x)\} \quad \text{und} \quad \Omega_2 = \Omega \setminus \Omega_1, \tag{2.8}$$

was äquivalent zu

$$\delta_0(x) = \begin{cases} 1: & \pi_1 c_{12} f_1(x) \geq \pi_2 c_{21} f_2(x) \\ 2: & \textit{sonst} \end{cases} \tag{2.9}$$

ist. Sind die Klassendichten f_1, f_2 für alle $x \in \Omega$ größer als Null, kann (2.9) auf eine prägnante Form gebracht werden:

$$\delta_0(x) = \begin{cases} 1: & \frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} \geq \frac{c_{21}}{c_{12}} \\ 2: & \textit{sonst} \end{cases} . \tag{2.10}$$

[4] d.h. die Beschränkung eines $\mathcal{R}_i(\delta)$ und die Minimierung der anderen

[5] d.h. die Minimierung von $\max_i \mathcal{R}_i(\delta)$

Der Ausdruck $\log\frac{c_{21}}{c_{12}}$ wird oft als *Schwellenwert* oder *Cut-off-Point* bezeichnet.

Die Form (2.10) zeigt, daß weniger die absolute Größe der Dichten als ihr Verhältnis zueinander für die Allokation eine Rolle spielt. Außerdem wird die Ambivalenz von Kosten und A-priori-Verteilung sichtbar: Unterschiedliche Anteile der Klassen am Grundraum Ω können durch reziproke Kosten aufgehoben werden.

Definition 2.5 (A-posteriori-Verteilung) *Die Wahrscheinlichkeit, daß ein beobachtetes Objekt x zur einer bestimmten Klasse gehört, wird als A-posteriori-Verteilung bezeichnet:*

$$\tau_i(x) := P(x \in \Pi_i | x) \quad (i = 1, \ldots, k). \tag{2.11}$$

Über das BAYESsche Theorem läßt sich diese Wahrscheinlichkeit folgendermaßen ausdrücken ([105], Kap. 1.3):

$$\tau_i(x) = \frac{\pi_i f_i(x)}{f(x)} \quad (i = 1, \ldots, k), \tag{2.12}$$

wobei

$$f(x) = \sum_{i=1}^{k} \pi_i f_i(x) \tag{2.13}$$

die Gesamtverteilung der Objekte im Grundraum darstellt. Es folgt sofort

$$\sum_{i=1}^{k} \tau_i(x) = 1; \tag{2.14}$$

speziell für $k = 2$ gilt also $\tau_1(x) + \tau_2(x) = 1$. Gilt $c_{ij} \equiv c \quad (i, j = 1, \ldots, k \quad j \neq i)$ und $c_{ii} = 0$, so ordnet δ_0 ein Objekt x gemäß $\arg\max_i \tau_i(x)$ der wahrscheinlichsten Klasse unter der Bedingung der gezogenen Stichprobe zu.

2.2 Ansätze für Schätzungen

Bei den meisten Diskriminanzproblemen sind weder die $\tau_i(x)$, die π_i noch die $f_i(x)$ bekannt und müssen deswegen wie das Risiko der Regel geschätzt werden.[6] Für die Ableitung einer Diskriminanzregel steht uns im Normalfall eine sogenannte *Trainingsstichprobe* $\mathcal{S}$, etwa vom Umfang n, zur Verfügung, die die Merkmalsausprägungen bereits klassifizierter Objekte enthält. $\mathcal{S}$ hat dann die Form

$$\mathcal{S} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} & z_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & z_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & z_n \end{pmatrix} = (X', z), \tag{2.15}$$

worin die z_i die Klassenzugehörigkeit kodieren:

$$z_j = i \iff x_j \in \Pi_i \quad (j = 1, \ldots, n). \tag{2.16}$$

[6] Die π_i sind dabei nur schätzbar, wenn eine Stichprobe ohne Ansehen der Klassenzugehörigkeit gezogen wurde (siehe Kapitel 2.2.2).

Die Klassifizierung der Objekte in $\mathcal{S}$ soll im weiteren korrekt sein. Fälle, in denen falschklassifizierte Objekte in der Trainingsstichprobe enthalten sind, werden hier nicht diskutiert (siehe dafür z.B. [96], [4]). Es sollen $n_1, \dots, n_k$ die Anzahlen der Objekte in $\mathcal{S}$ aus den einzelnen Populationen ($|\mathcal{S} \cap \Pi_i|$) bezeichnen, wobei $\sum_{i=1}^k n_i = n$ gilt. Mit $x_{i_1}, \dots, x_{i_{n_i}}$ seien im folgenden diejenigen Objekte in $\mathcal{S}$ bezeichnet, die aus der Population Π_i stammen. $x_{(1)}, \dots, x_{(p)}$ bezeichnen die einzelnen Merkmale eines Objekts.

Die Trainingsdaten können unterschiedlich erhoben worden sein:

- *Mixture Sampling*: Die Objekte wurden ohne Ansehen ihrer Klassenzugehörigkeit als Stichprobe gezogen (für (X', z)). Die Stichprobe enthält so Informationen über die Verhältnisse zwischen den Klassen.
- *Separate Sampling*: Für jede Klasse wurde separat eine Stichprobe erhoben (für $X'|z$); zusammen bilden sie die Trainingsstichprobe.
- *X-bedingtes Sampling*: Für einen oder mehrere feste Werte werden wiederholt Stichproben im Mixture Sampling erhoben (für $z|X'$).

Zu Sampling-Designs siehe auch [198], Kap. 6.4.2.

Beispiel 2.7 (Epidemiologische Studiendesigns) *Mixture Sampling ist typisch für prospektive Studien in der Medizin, da hier ein Personenkollektiv über einen längeren Zeitraum beobachtet wird und sich die Klassenmerkmale, etwa die Entwicklung einer bestimmten Krankheit, erst später zeigen. Retrospektive Studien, wie z.B. Fall-Kontroll-Studien, unterteilen die Personengruppe bereits nach dem Vorhandensein der Erkrankung und ziehen aus diesen getrennten Populationen ihre Stichprobe. Sie stellen also ein Beispiel für Separate Sampling dar. Für die Definition von Studiendesigns siehe z.B. [183].*

2.2.1 Schätzungen der Dichtequotienten

Viele nichtparametrische Verfahren schätzen direkt die A-posteriori-Verteilungen τ_i, während viele parametrische die π_i und f_i getrennt modellieren, um über das BAYESsche Theorem (2.12) die τ_i zu erhalten. Da letztendlich für die Entscheidung über die Zuordnung eines Objekts x nur die relativen Größen der A-posteriori-Dichten in x zählen, modelliert z.B. die Logistische Diskriminanzanalyse (Kapitel 3.4) nur deren Quotienten. Die Diskriminanzverfahren lassen sich grob in zwei Gruppen in Bezug auf den Schätzansatz einteilen: parametrischer und nichtparametrischer Zugang.[7]

Parametrischer Zugang

Es wird angenommen, daß die Verteilungen, die Quotienten der Verteilungen in den Populationen oder die Trennfläche(n) global einer a-priori angenommenen parametrischen Form genügen.[8] So kommt es darauf an, die Parameter für diese Form aus den Beobachtungen in $\mathcal{S}$ zu schätzen:

[7] Es existieren aber noch weitere Ansätze. So diskutieren H. LÄUTER & THIELE [147] Konvexkombinationen der A-posteriori-Verteilungen jeweils einer Klasse aus mehreren Verfahren oder Ansätzen: $\tilde{\tau}_i(x) = \sum_{j=1}^s l_j \hat{\tau}_i^{(j)}$, wobei $l_j \geq 0$ und $\sum_{j=1}^s l_j = 1$ gilt. Die $\hat{\tau}_i^{(j)}$ können z.B. aus unterschiedlichen Verfahren mit eigenen Stichproben stammen; die Verfahren erhalten so über die l_j verschiedene Gewichte. Die Zuordnung erfolgt dann mittels der $\tilde{\tau}_i$. Solche speziellen Ansätze werden aber aufgrund des Übersichtscharakters dieser Arbeit nicht weiter betrachtet.

[8] Manche Autoren stellen keine Forderungen an die Klassendichten, sondern z.B. an die Glattheit der Trennfläche, etwa in [155]. Dieser Ansatz soll hier nicht weiter verfolgt werden.

- *Plug-In-Schätzungen* ersetzen die gesuchten Parameter durch übliche Schätzungen, z.B. den Erwartungswert durch das Stichprobenmittel. Dieser Ansatz ist nicht immer gesichert, weil beispielsweise nicht immer eine geschlossene Form für die Parameterschätzung exisiert.

- *Likelihood-Schätzungen* bestimmen die gesuchten Parameter als diejenigen, die am besten die beobachtete Stichprobe erklären, d.h. sie mit größter Wahrscheinlichkeit hervorbringen. Dazu wird eine Plausibilitätsfunktion (Likelihood, Quasi-Likelihood) maximiert: Maximum-Likelihood (ML). Dieser Ansatz ist immer durchführbar, entweder in geschlossener Form (manche Plug-In- sind auch Likelihood-Schätzungen) oder durch numerische Approximation (z.B. Newton-Rhaphson, Quasi-Newton [195], EM-Algorithmus [150]). Außerdem besitzt er unter relativ geringen Voraussetzungen gute Eigenschaften wie Konsistenz, asymptotische Effizienz und Konsistenzbeharrung auch bei invalidem Modell ([161]).

  Plug-In- und Likelihood-Schätzungen lassen sich zum sogenannten *Schätzzugang*[9] zusammenfassen.

  Beispiel 2.8 (Normalverteilungen) *Betrachtet man zwei normalverteilte Populationen mit gemeinsamer Kovarianz, so liefern die Klassenmittel*

  $$\hat{\mu}_i := \bar{x}_i = \frac{1}{n_i}\sum_{l=1}^{n_i} x_{i_l} \quad (i = 1, 2)$$

  und die gemeinsame Kovarianz aller Beobachtungen aus der Stichprobe

  $$\hat{\Sigma} := S_{pooled} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2},$$

  mit

  $$S_i := \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (x_{i_j} - \bar{x}_i)(x_{i_j} - \bar{x}_i)' \quad (i = 1, 2)$$

  als Plug-In-Schätzungen die vollständig bestimmten Populationsverteilungen. Die $\bar{x}_i$ sind dabei auch ML-Schätzungen für die Mittelwerte.

- *Prädiktiver Zugang*[10]. Während beim Schätzzugang Information für die Schätzung nur aus der Stichprobe gewonnen wird, versucht der Prädiktive Zugang, weitere Information außerhalb der Stichprobe in die Schätzungen einfließen zu lassen:

  $$\hat{f}_i^{(P)}(x; \mathcal{S}) := \int f_i(x, \theta)\, p(\theta|\mathcal{S})\, d\theta \quad (i = 1, \ldots, k). \tag{2.17}$$

  $p(\theta|\mathcal{S})$ kann als A-posteriori-Dichte[11] von θ gesehen werden, d.h. die A-priori-Vorstellung $p(\theta)$ vom Parameter wird über die Stichprobe korrigiert. Dies liefert die sogenannte *prädiktive Dichte* $\hat{f}_i^{(P)}$. Man gewichtet so zwischen den Modellvorstellungen (die vielleicht auf fachspezifischen Erkenntnissen beruhen) und den beobachteten Daten. Dieser Ansatz ist durch die Integralbestimmung sehr rechenintensiv und nicht immer durchführbar. Siehe auch [185], [84], [82].

[9] auch: *Estimative Approach*

[10] auch: *Predictive Approach* oder BAYES-*Zugang*

[11] A-priori- und A-posteriori-Dichten beziehen sich hierbei auf die Verteilung des Parameters θ im Zuge BAYESscher Schätzungen, nicht auf den Diskriminanzanalyse-Kontext.

Beispiel 2.9 *Langjährige Niederschlagsmittel oder die Parteipräferenzen im Wahlverhalten bestimmter Personengruppen sind Beispiele für Informationen außerhalb der Stichprobe. Über die A-priori-Verteilung kann beispielsweise der Stimmenanteil der Parteien in einem Wahlbezirk im Mittel der letzten zehn Jahre in eine Schätzung einfließen.*

Für kleine n sind durch die beiden Ansätze sehr unterschiedliche Schätzungen möglich; für große n liefern sie ähnliche Resultate ([161]).

Ein Spezialfall ist das hier nicht behandelte *Sheshel*-Verfahren, das nur Annahmen über Erwartungswert und Varianz der Verteilungen zweier Populationen macht und das sowohl für stetige als auch für quantitative Variablen anwendbar ist (siehe [205], [133]).

Nichtparametrischer Zugang

Nichtparametrische Verfahren verzichten auf die Annahme einer a-priori globalen parametrischen Form. Stattdessen kommen verschiedene Ansätze zum Zuge, z.B.:

- *Kontingenztafeln* schätzen die Dichten $f_1, \ldots, f_k$ von diskreten Merkmalen in den einzelnen Zellen anhand der Beobachtungen, die beispielsweise in allen Merkmalsausprägungen mit der Zelle übereinstimmen (*Multinomialmodell*, Kapitel 4.1.1) oder auch nur in mindestens einer (*Unabhängigkeitsmodell*, Kapitel 4.1.1).

- Kerndichteschätzungen und k-Nearest-Neighbours-Methoden (Kapitel 4.2 und 4.3) führen *Lokale Dichteschätzungen* von $f_1, \ldots, f_k$ um den zu klassifizierenden Punkt mittels der Stichprobe durch.[12] Diese werden dann in die Bayessche Regel eingesetzt.

- CART (Kapitel 4.4) unterteilt den Merkmalsraum induktiv in verschachtelte disjunkte Untermengen, um ein *Unreinheitskriterium* bezüglich der Klassenzugehörigkeit der Beobachtungen in diesen Teilmengen zu minimieren.

- Eventuell nichtlineare *funktionelle Zusammenhänge* zwischen Merkmalsvariablen und der Klassenzugehörigkeit versuchen Neurale Netze zu modellieren (Kapitel 4.5).

Dabei adaptieren sich CART und Neuronale Netze selbst so, daß sie das Risiko möglichst minimieren; bei den lokalen Dichteschätzern existieren adaptive Versionen. Letztendlich müssen auch bei nichtparametrischen Verfahren Parameter geschätzt werden, etwa für Glättung (Kapitel 2.3) oder funktionellen Zusammenhang. Deren Schätzung erfolgt oft durch iterative Optimierung eines Kriteriums oder als Maximum-Likelihood-Schätzung (ML-Schätzung) Außerdem muß oft eine Auswahl von benutzten Funktionsklassen getroffen werden.

Es gilt aber, die andere inhaltliche und praktische Bedeutung der Parameter bei nichtparametrischen Verfahren zu beachten: Sie stellen *Verfahrensparameter* dar, während es sich beim parametrischen Ansatz um *Modellparameter* handelt.

[12] Nearest-Neighbours-Methoden werden in der Theorie oft als Dichteschätzung hergeleitet, modellieren in der Praxis aber eher den Dichtequotienten.

2.2.2 Schätzung von A-priori-Wahrscheinlichkeiten

Die A-priori-Wahrscheinlichkeiten $\pi_1, \dots, \pi_k$ sind oft durch Informationen in der Praxis gegeben oder werden unter Zusatzannahmen bestimmt. Soll beispielsweise eine Krankheit diagnostiziert werden, so ist π_1 mit der Prävalenz und π_2 mit $1 - \pi_1$ gegeben, wenn man Π_1 als die Population der Erkrankten und Π_2 als die der Gesunden annimmt ([183], S. 84). Oft liegen solche Informationen jedoch nicht vor, und die π_i müssen aus $\mathcal{S}$ geschätzt werden.

Wurde die gesamte Stichprobe im Mixture Sampling erhoben, so erhält man als ML-Schätzung ([161], Kap. 2.3):

$$\hat{\pi}_i := \frac{n_i}{n}. \tag{2.18}$$

Im Falle des Separate Sampling enthalten die n_i keine Informationen über die Anteile der Populationen. Hier ist eine Schätzung nur möglich, wenn weitere Beobachtungen, etwa m an der Zahl, unter Mixture Sampling gezogen wurden, aber unklassifiziert sind. Formt man aus den $x_1, \dots, x_n$ eine Klassifikationsregel δ_π (z.B. parametrisch über eine ML-Schätzung für die Gesamtverteilung[13] $f(x) = \sum_{i=1}^k \pi_i f_i(x)$, siehe [161], Kap. 2.7-2.10) und wendet sie auf die m neuen Beobachtungen an, so erhält man mit $m_1, \dots, m_k$ eine Schätzung für die Anteile der Klassen. Analog zu (2.18) erhält man mit $\frac{m_i}{m}$ eine Schätzung der A-priori-Verteilung. Diese ist jedoch verfälscht. McLachlan ([161], Kap. 2.3) gibt als erwartungstreuen Schätzer bei $k = 2$

$$\hat{\pi}_1^D := \frac{\frac{m_1}{m_2} - \alpha_{21}}{\alpha_{11} - \alpha_{21}} \quad \text{und} \quad \hat{\pi}_2^D = 1 - \hat{\pi}_1^D \tag{2.19}$$

an, wobei $\hat{\pi}_1^D$ bei Verlassen des Intervalls $[0, 1]$ auf 0 bzw. 1 gesetzt wird. Dieser Schätzer ist effizient, falls die parametrische Struktur erfüllt ist. Falls außerdem die π_i nicht zu unterschiedlich sind und n nicht zu klein bezüglich m, ist er auch relativ hoch effizient. Als Alternative gibt McLachlan ([161], Kap. 2.3) einen relativ asymptotisch effizienten Momenten-Schätzer an:

$$\hat{\pi}_1^M := \frac{(\bar{x}_1 - \bar{x}_2)' S_u^{-1} (\bar{x}_u - \bar{x}_2)}{(\bar{x}_1 - \bar{x}_2)' S_u^{-1} (\bar{x}_1 - \bar{x}_2)} \quad \text{und} \quad \hat{\pi}_2^M = 1 - \hat{\pi}_1^M. \tag{2.20}$$

Dabei sind $\bar{x}_u = \frac{1}{m} \sum_{j=n+1}^m x_j$ und $S_u = \frac{1}{m-1} \sum_{j=n+1}^m (x_j - \bar{x}_u)(x_j - \bar{x}_u)'$ Schätzungen für den Erwartungswert bzw. die Kovarianz der m neuen Beobachtungen.

2.2.3 Schätzung von Fehlerraten

Zum Vergleich von Diskriminanzregeln muß deren Risiko (2.4) geschätzt werden. Auch hier gibt es den parametrischen und den nichtparametrischen Weg. Alternative Kriterien, die etwa den Fehler bei der Dichtenapproximation schätzen (MISE, vgl. [218], Kap. 5.5) und dann auf das Risiko schließen, sind mit Vorsicht zu „genießen“, weil diese Kriterien nicht unbedingt an das Bayessche Risiko (2.4) gekoppelt sind und dadurch zu stark suboptimalen Regeln führen können (siehe Kapitel 2.3).

Parametrische Fehlerschätzung

In einigen Modellen lassen sich die Fehlklassifikationswahrscheinlichkeiten α_{ij} *parametrisch* explizit angeben oder asymptotisch annähern. Siotani et al. ([202],

[13] Die Identifizierbarkeit bei der Schätzung von Mischverteilungen ist oft nicht gegeben. Daher werden die Einzelverteilungen meist getrennt in den Einzelpopulationen geschätzt.

Kap. 9.6) etwa geben für die Lineare Diskriminanzregel (LDA, Kapitel 3.2) im Fall zweier normalverteilter Klassen mit gleicher Kovarianz ($\Sigma_1 = \Sigma_2 =: \Sigma$), symmetrischen Kosten ($c_{12} = c_{21}$) und gleichen A-priori-Wahrscheinlichkeiten ($\pi_1 = \pi_2$) asymptotische Resultate für α_{12} und α_{21} an:

$$\alpha_{12} \approx \alpha_{21} \approx \Phi\left(-\frac{\Delta}{2}\right),$$

wobei Φ die Standardnormalverteilung N(0,1) und

$$\Delta^2 := \|\mu_1 - \mu_2\|^2_{\Sigma^{-1}} = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \tag{2.21}$$

den Mahalanobis-Abstand zweier Populationen bezeichnet. Das Risiko kann so über $\mathcal{R}(\delta_{LDA}) \approx \Phi\left(-\frac{\Delta^2}{2}\right)$ angenähert werden. Das Problem dieser Resultate ist, daß die Risikoschätzung von den Modellannahmen abhängt. Bei Abweichungen davon kann die Schätzung beliebig schlecht werden und damit auch die Regel.

Resubstitutionsmethode

Der einfachste *nichtparametrische Zugang* ist die *Resubstitutionsmethode*. Sämtliche Beobachtungen aus $\mathcal{S}$ werden durch die gewonnene Regel erneut klassifiziert. Der Vergleich mit der bekannten Klassenzugehörigkeit liefert eine Schätzung der Fehlklassifikationskosten pro Klasse und damit des Risikos:

$$\hat{\mathcal{R}}_i(\delta, \mathcal{S}) = \sum_{j=1}^{k} c_{ij} \underbrace{\left[\frac{1}{n_i}\sum_{l=1}^{n_i} I_{\delta(x_{i_l})=j}\right]}_{\hat{\alpha}_{ij}} \quad (i = 1, \ldots, k) \tag{2.22}$$

$$\hat{\mathcal{R}}_{app}(\delta, \mathcal{S}) = \sum_{i=1}^{k} \pi_i \hat{\mathcal{R}}_i(\mathcal{S}), \tag{2.23}$$

wobei I die Indikator-Funktion mit

$$I_\eta := \begin{cases} 1: & \eta \quad \text{ist wahr} \\ 0: & \eta \quad \text{ist falsch} \end{cases} \tag{2.24}$$

darstellt. $\hat{\mathcal{R}}_{app}(\delta, \mathcal{S})$ heißt *Apparent Error Rate*[14] (*AER*, [124]) und ist eine Schätzung für das Bayessche Risiko (2.4). Da $\mathcal{S}$ schon zum Finden der Regel (minimales Risiko!) benutzt wurde, ist diese an $\mathcal{S}$ adaptiert (sogenanntes Over-fitting). Die $\hat{\mathcal{R}}_{app}(\delta, \mathcal{S})$ unterschätzt das Risiko von δ bei neuen Beobachtungen und ist damit verfälscht:

$$\hat{\mathcal{R}}_{app}(\delta, \mathcal{S}) \leq \mathcal{R}(\delta_0) \leq \mathcal{R}(\delta).$$

McLachlan [160] erhält z.B. für den Fall der Linearen Diskriminanzregel (Kapitel 3.2) unter den Voraussetzungen wie zuvor als asymptotischen Bias 2. Ordnung für die α_{ij} eine positive Größe:

$$\hat{\alpha}_{12}^{AER} \approx \Phi\left(-\frac{\Delta}{2}\right) - \underbrace{\phi\left(-\frac{\Delta}{2}\right)\left[\frac{1}{n_1}\left(\frac{\Delta}{4} + \frac{p-1}{\Delta}\right) + \frac{(p-1)\Delta}{2(n_1+n_2-2)}\right]}_{>0}$$

[14] auch: *Actual Error Rate*

(α_{21} analog mit n_2 statt n_1 im mittleren Term). ϕ bezeichnet die Dichte der Standardnormalverteilung. Unter gültigen Modellannahmen wird so das Risiko asymptotisch durch positiven Bias unterschätzt.

Es gibt zwei Lösungen dieses Problems:

- *Testdatensatz*: Durch Aufteilung der Stichprobe $\mathcal{S}$ in eine Trainingsmenge (Bestimmung der Regel) und eine Testmenge (Bestimmung des Risikos) umgeht man das Over-fitting durch Unabhängigkeit. Bei kleinen Stichprobenumfängen oder hochdimensionalen Problemen kann aber die verkleinerte Stichprobe zur Regelbestimmung das Risiko der Regel erhöhen.

- *Modifikation der Schätzung*: Durch Modifikation von $\hat{\mathcal{R}}_{app}(\delta, \mathcal{S})$ erhält man die Unverfälschtheit der Schätzung, so daß die gesamte Stichprobe zur Regelbestimmung benutzt werden kann. Dies wird aber mit einer größeren Variabilität der Risikoschätzung und einem erhöhten Rechenaufwand erkauft (siehe z.B. Efron [55], [56]).

Als gebräuchliche Modifikationen existieren Cross-validation und Bootstrap, die im folgenden kurz dargestellt werden sollen.

Cross-validation

Cross-validation (CV, Kreuzvalidierung) geht auf die *Leave-one-out*-Methode von Lachenbruch [143] zurück. Dabei wird jeweils eine Beobachtung x_j aus $\mathcal{S}$ weggelassen und mit dieser verkleinerten Stichprobe $\mathcal{S}_{-j}$ die Regel δ_{-j} bestimmt. Der Vergleich der bekannten Klassenzugehörigkeit von x_j mit $\delta_{-j}(x_j)$ ermöglicht wieder die Schätzung der α_{ij}:

$$\hat{\alpha}_{ij}^{(CV)} = \frac{1}{n_i} \sum_{l=1}^{n_i} I_{\delta_{-i_l}(x_{i_l})=j}$$

Für lineare Modelle ist gezeigt worden ([200]), daß Leave-one-out bei der Modellwahl konservativ[15] und inkonsistent ist, da einzelne Beobachtungen nur ungenügend die Variabilität der Stichprobe widerspiegeln. Durch Bildung mehrerer annähernd gleichgroßer Gruppen zum Herauslassen $\{\mathcal{X}_1, \ldots, \mathcal{X}_d\}$, $\bigcup_l \mathcal{X}_l = \mathcal{S}$, kann die Methode zum *Gruppierten Cross-validation*[16] verallgemeinert werden:

$$\hat{\alpha}_{ij}^{(d-CV)} = \frac{1}{n_i} \sum_{l=1}^{d} \sum_{x \in \mathcal{X}_l \cap \Pi_i} I_{\delta_{-\mathcal{X}_l}(x)=j} \tag{2.25}$$

$$\hat{\mathcal{R}}^{(CV)}(\delta, \mathcal{S}) = \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} c_{ij} \hat{\alpha}_{ij}^{(d-CV)}, \tag{2.26}$$

wobei $\mathcal{X}_l$ die l-te Gruppe der weggelassenen Beobachtungen darstellt.

Shao [199] hat gezeigt, daß durch das Weglassen von d Beobachtungen mit $\lim_{n\to\infty} \frac{d}{n} = 1$ und $\lim_{n\to\infty}(n - d) = \infty$ Cross-validation konsistent in der Modellwahl wird. Zhang [227] empfiehlt 5-10 Gruppen, Venables & Ripley [218] benutzen durchgängig 10 Gruppen. Die Form der Schätzung bedingt eine höhere

[15] d.h. überschätzt das Risiko

[16] auch: *d-fold Cross-validation*

Variabilität und einen, um den Faktor der Anzahl der Gruppen, erhöhten Rechenaufwand.

Annähernd gleiche Resultate liefert die *Jackknife*-Methode, die hier nicht weiter Gegenstand sein soll (siehe [200]).

Bootstrap

Die Bootstrap-Methode (BS) geht auf EFRON [55] zurück. Das Primärinteresse besteht in dem Ergebnis einer Statistik $T(f_1, \ldots, f_k)$ (im Fall der Diskriminanzanalyse die Risiko-Schätzung), für die oft keine theoretische Ableitung existiert oder die, bei praktischen Problemen, nicht jedesmal neu hergeleitet werden soll. Da die wahre Verteilung F jedoch unbekannt ist, versucht man, diese durch die empirische Verteilung $\hat{F}$ als bestmögliche Schätzung zu ersetzen, und hofft, daß die Eigenschaften von $T(\hat{f}_1, \ldots, \hat{f}_k)$ denen der Originalstatistik entsprechen. Dazu werden aus der neuen Grundgesamtheit $\mathcal{S}$ als Träger der empirischen Verteilung Stichproben mit Zurücklegen $\mathcal{S}^*$ vom Umfang n gezogen (sogenannte *Bootstrap-Stichproben*). Diese sollen Bias-Schätzungen $b_i^{(B)}$ liefern, mit denen die $\hat{\mathcal{R}}_i$ korrigiert werden können:

$$\hat{\mathcal{R}}_i^{(B)}(\mathcal{S}) = \hat{\mathcal{R}}_i(\mathcal{S}) - \hat{b_i}^{(B)}(\mathcal{S}) \tag{2.27}$$

$$\hat{\mathcal{R}}^{(B)}(\delta, \mathcal{S}) = \sum_{i=1}^{k} \pi_i \hat{\mathcal{R}}_i^{(B)}(\mathcal{S}). \tag{2.28}$$

Für jede Stichprobe $\mathcal{S}^*$ wird die Diskriminanzregel δ^* bestimmt und durch Anwendung auf $\mathcal{S}^*$ deren klassenspezifischen Kosten $\hat{\mathcal{R}}_i(\mathcal{S}^*)$ durch Resubstition analog der Apparent Error Rate (2.23) berechnet. Gemittelt über mehrere Wiederholungen (50 bis 350 oder mehr), liefert die Differenz $\hat{\mathcal{R}}_i(\delta^*, \mathcal{S}^*) - \hat{\mathcal{R}}_i(\delta^*, \mathcal{S})$ eine Bias-Schätzung für $\hat{\mathcal{R}}_i(\mathcal{S})$:

$$\hat{b_i}^{(B)}(\mathcal{S}) = \frac{1}{M} \sum_{m=1}^{M} \Big[\hat{\mathcal{R}}_i(\delta_m^*, \mathcal{S}_m^*) - \hat{\mathcal{R}}_i(\delta_m^*, \mathcal{S}) \Big].$$

Bootstrap hat eine geringere Variabilität, aber einen höheren Bias als Crossvalidation. Diese Unterschiede treten aber nur auf, wenn die Populationen eng beieinander liegen ([46]). Weiter verfeinerte Algorithmen sind z.B. Double-BS, Randomized-BS oder ein sogenanntes 0.632-BS[17]. Unter einigen Voraussetzungen ist die Bootstrap-Methode konsistent. Für Einzelheiten siehe SHAO & TU [200].

2.3 Klassifikationsfehler und Regularisierungen

Entgegen der intuitiven Meinung, daß möglichst genaue Approximationen der Klassendichten oder A-posteriori-Verteilungen auch bessere Klassifikationsergebnisse liefern, haben FRIEDMAN u.a.[18] die Herkunft der Klassifikationsfehler genauer untersucht. Sie sind dabei zu interessanten Ergebnissen gekommen, die im folgenden kurz dargestellt werden.

Bias und Varianz einer aus der Trainingsstichprobe $\mathcal{S}$ gewonnenen Regel $\hat{\delta}$ sind wie folgt definiert:

[17] EFRON [57] formuliert die Risikoschätzung als gewichtete Summe aus Resubstitutions- und Bootstrapschätzung mit Anteilen von 0.368 bzw. 0.632

[18] siehe z.B. Arbeiten von JAMES & HASTIE [129], BREIMAN [26], DIETTERICH & KONG [53]

Definition 2.6 (Bias und Varianz einer geschätzten Regel)

$$\mathrm{Bias}\,\hat{\delta}(x) \;:=\; \mathrm{E}_{\mathcal{S}}\left[z(x) - \mathrm{E}_{\mathcal{S}}\hat{\delta}_{\mathcal{S}}(x)\right] \tag{2.29}$$

$$\mathrm{Var}\,\hat{\delta}(x) \;:=\; \mathrm{E}_{\mathcal{S}}\left[\hat{\delta}_{\mathcal{S}}(x) - \mathrm{E}_{\mathcal{S}}\hat{\delta}_{\mathcal{S}}(x)\right]^2 \tag{2.30}$$

Die Varianz[19] reflektiert die Sensibilität einer Schätzung gegenüber der Stichprobe, der Bias[20] gegenüber der gesuchten Regel.[21] Stünden unbegrenzte Stichprobenumfänge für die Analyse zur Verfügung, ließe sich die Varianz einer Regel beliebig verringern. Da in der Praxis aber nur beschränkte Umfänge zu haben sind, ist die Varianz größer als Null und trägt zum Risiko bei. Theoretische Optimalitätskriterien unter der Bedingung $|\mathcal{S}| = \infty$ sind so nur begrenzt aussagefähig.

Wie wirken nun Bias und Varianz auf das Risiko? Für zwei Klassen ($k = 2$), $\pi_1 = \pi_2$ und symmetrische Kosten ($c_{12} = c_{21}$) hat FRIEDMAN [74] das Risiko (2.5) untersucht. Für eine geschätzte Regel $\hat{\delta}$ reduziert es sich unter diesen Voraussetzungen in einem Punkt x zur Wahrscheinlichkeit einer Fehlklassifikation: $\mathcal{R}(\delta, x) = P\left(\hat{\delta}(x) \neq z(x)\right)$. Eine Dekomposition in einen reduziblen und einen irreduziblen Teil (in einem Punkt x, hier in den Bezeichnungen weggelassen) ist durch

$$P(\hat{\delta} \neq z) \;=\; |2\tau_2 - 1| P(\hat{\delta} \neq \delta_0) + P(\delta_0 \neq z) \tag{2.31}$$

gegeben, wobei δ_0 die BAYESsche Regel bezeichnet ([74], Kap. 5). Der letzte Term in (2.31) gibt das (minimale) Risiko der BAYESschen Regel an, das unabhängig von $\hat{\delta}$ ist; der mittlere trägt additiv zum Risiko bei, indem er den Fehler bei der Schätzung der optimalen Trennfläche repräsentiert (*Trennflächenfehler*). FRIEDMAN [74] bietet als Approximation dafür

$$P(\hat{\delta} \neq \delta_0) \approx \tilde{\Phi}\left[\frac{1}{\sqrt{\mathrm{Var}\hat{\delta}}}\,\mathrm{sgn}\left(\tau_2 - \frac{1}{2}\right)\left(\mathrm{E}\hat{\tau}_2 - \frac{1}{2}\right)\right] \tag{2.32}$$

an, wobei $\tilde{\Phi}(x) = 1 - \Phi(x)$ die Fläche der Standardnormalverteilung oberhalb von x repräsentiert. Der Ausdruck

$$b(\tau_2, \mathrm{E}\hat{\tau}_2) := \mathrm{sgn}\left(\frac{1}{2} - \tau_2\right)\left(\mathrm{E}\hat{\tau}_2 - \frac{1}{2}\right) \tag{2.33}$$

heißt *Boundary Bias*. Damit ist das Risiko (2.31) nur über das *Vorzeichen* von $\left(\frac{1}{2} - \tau_2\right)$ abhängig, nicht vom absoluten Wert des Terms. Solange der Boundary Bias negativ ist, wird mit wachsendem Abstand zur optimalen Trennfläche ($|\mathrm{E}\hat{\tau}_2 - \frac{1}{2}| \uparrow$) das Risiko also *kleiner*! Ist der Boundary Bias fest, wird bei negativem Vorzeichen mit kleinerer Varianz auch das Risiko kleiner, während bei positivem Vorzeichen das Risiko mit kleiner werdender Varianz wächst (siehe Abbildung 2.2).

Die Ursache für dieses dem „gesunden Menschenverstand“ widersprechende Verhalten ist die andere Art des Fehlers in der Diskriminanzanalyse im Vergleich zum Regressionskontext, an dem sich der Verstand meist „schult“: In einer Beobachtung wird statt eines quantitativen Verlustes, etwa gemessen an der Abweichung $|f - \hat{f}|$ bezüglich einer L_p-Norm, ein qualitativer Verlust (0-1 — entweder richtig oder falsch klassifiziert) betrachtet. So kann der Bias eines Modells ruhig groß sein, solange man

[19] Maß für den zufälligen Anteil am Risiko

[20] zu deutsch: Verfälschung, Neigung, Hang; Maß für den Anteil des Schätzfehlers am Risiko

[21] $\mathrm{E}_{\mathcal{S}}[.]$ bedeutet die Mittelung über alle möglichen Stichproben.

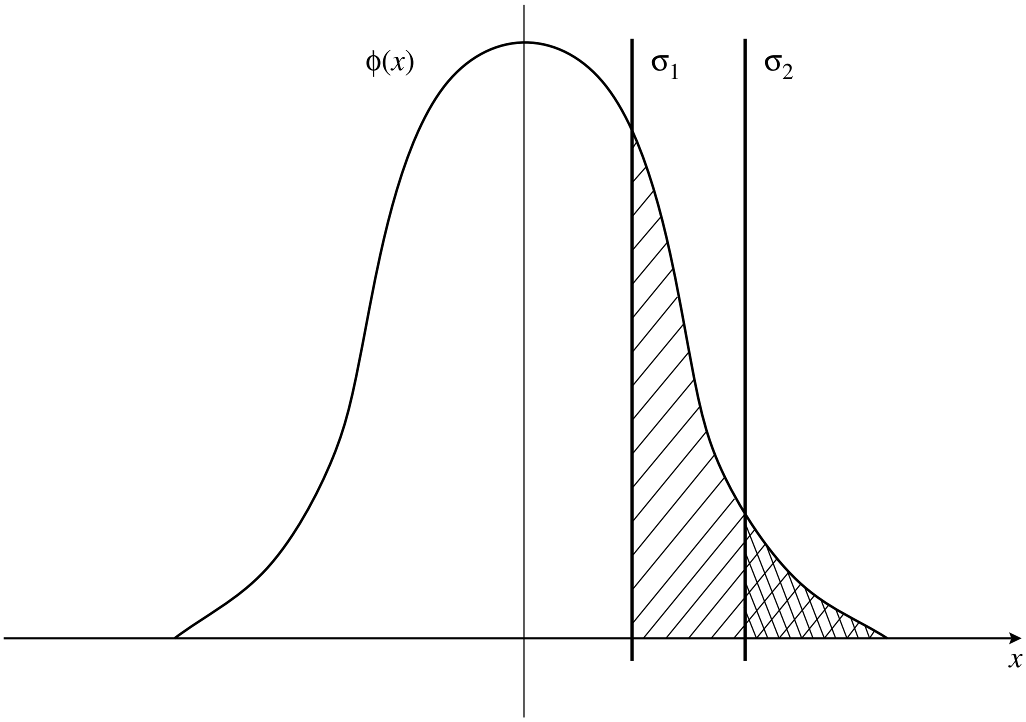

Abbildung 2.2: Fehlerapproximation durch FRIEDMAN [74]: Bei negativem Boundary Bias verringert sich gemäß der Approximation (2.32) mit abnehmender Varianz der Trennflächenfehler $P(\hat{\delta} \neq \delta_0)$ und damit das Risiko. Die Abbildung skizziert dies für zwei verschiedene Varianzen $\sigma_1 > \sigma_2$: Die schraffierten Flächen stellen den Wert von $\tilde{\Phi}(x)$ für die σ_j dar.

sich auf der „richtigen" Seite der Trennfläche (d.h. negativer Boundary Bias) befindet: Die korrekte Klassifikation durch die Regel wird deren Risiko gering halten. Das ist die tiefere Ursache dafür, daß Methoden, die zwar für die Approximation des Dichtequotienten in der Regression aufgrund eines sehr hohen Bias durch starkes Glätten (Over-smoothing) ungeeignet sind, trotzdem sehr gute Klassifikationsergebnisse liefern können. So werden die optimalen Glättungsparameter, wie die Bandweite bei Kernschätzungen (Kapitel 4.2) oder die Anzahl der betrachteten Nachbarn bei Nearest Neighbours (Kapitel 4.3), in der Diskriminanzanalyse oft größer ausfallen als bei einer reinen Kurvenapproximation.

Regularisierungen bewirken eine Varianzreduktion bei einer gleichzeitig möglichen Biaserhöhung. Dies kann durch zusätzliche Annahmen über Parameter geschehen, z.B. die Annahme gleicher Kovarianzen in beiden Klassen oder die der Unabhängigkeit von Merkmalen — also das Nullsetzen aller Nichtdiagonalelemente in Kovarianz- oder Glättungsmatrizen. Das ist z.B. bei schlecht gestellten[22] oder ärmlich gestellten Problemen[23] erforderlich, um die Parameteranzahl zu verringern; allgemein, wenn die n_i bezüglich p zu klein sind, so daß die Varianz groß oder die Parameter sogar nicht identifizierbar sind. Auch eine stärkere Glättung, d.h. die Mittelung über größere Gebiete oder mehr Werte, verfolgt diesen Zweck. Ein weiterer Weg ist die Bestrafung zu komplexer und variabler Strukturen in der Modell-

[22] *ill-posed*, d.h. $\forall i \in \{1, \ldots, k\} : n_i \leq p$
[23] *poor-posed*, $n_i < p$

wahl mittels der Addition von Straftermen in Minimierungsproblemen. Komplexe Strukturen können zwar theoretisch sehr komplizierte[24] Zusammenhänge modellieren, ihnen fehlt aber bei kleinen Stichproben schlicht genügend Information, um diese auch mit kleiner Varianz schätzen zu können. Die Bias-Varianz-Abwägung wird oft durch Parameter gesteuert, etwa in der Regularisierten Diskriminanzanalyse (Kapitel 3.3.3).

Die Regularisierungstechniken funktionieren in der Diskriminanzanalyse im allgemeinen viel besser als in der Regressionsanalyse. Sogar der gesamte parametrische Ansatz (Kapitel 3) kann als eine Form der Glättung durch Regularisierung verstanden werden, falls z.B. Nichtnormalität oder diskrete Daten vorliegen. Der ideale Kompromiß zwischen Varianz und Bias zur Minimierung des BAYESschen Risikos ist selten im vornherein klar. Die Abwägung zwischen Bias und Varianz ist deswegen am besten durch eine kreuzvalidierte oder Bootstrap-Schätzung des Risikos (2.4) zu erreichen.

Beispiel 2.10 (Parameteranzahl im Normalen Modell) *Bei k normalverteilten Populationen sind bei der Linearen und der Quadratischen Diskriminanzanalyse (siehe Kapitel 3.2 und 3.3) aufgrund der angenommenen Heteroskedastizität[25] bzw. bei Homoskedastizität[26] folgende Anzahlen an Parametern zu schätzen ([161]):*

$$k + \frac{1}{2}(p+1)p \quad \textit{bzw.} \quad k + k\frac{1}{2}(p+1)p. \tag{2.34}$$

Bei $k = 2$ und $p = 10$ sind das 120 bzw. 65 Parameter. Durch die Einschränkung auf eine gemeinsame Kovarianzmatrix vermindert sich so die Anzahl der Parameter um ungefähr die Hälfte.

2.4 Die Datengrundlage

2.4.1 Datentypen

Die Art der Merkmale bestimmt grundlegend die anwendbaren Diskriminanzverfahren. Messungen an Objekten werden immer über Skalen möglicher Werte durchgeführt. Nach Art der zugelassenen Werte auf einer Skala lassen sich zwei Hauptgruppen unterscheiden:

Stetige Merkmale

Stetige Merkmale[27] besitzen eine Ordnung, Abstände zwischen Werten sind interpretierbar, und die Menge der möglichen Werte ist dicht. Auch Quotienten von Werten machen in dieser Skala einen Sinn. Viele physikalische und geographische Meßwerte, aber auch Börsenkurse sind dadurch repräsentiert.

Beispiel 2.11 *Beispiele für stetige Merkmale sind Körpergrößen, Lebensalter, Produktpreise, Stromstärken, der Deutsche Aktien-Index (DAX) oder der Quotient aus Jahresgewinn und Umsatz eines Unternehmens.*

[24] engl.: sophisticated
[25] unterschiedliche Kovarianzmatrizen in den Populationen
[26] gleiche Kovarianzmatrizen in den Populationen
[27] auch: *reell* oder *metrisch*

Diskrete Merkmale

Diskrete Merkmale besitzen demgegenüber nicht die Eigenschaft beliebig dichter Wertemengen. Die möglichen Werte liegen auf einem Gitter (mit vielleicht unterschiedlichen Abständen); Werte zwischen den Gitterpunkten der Skala sind undefiniert.

Es gibt einige Spezialfälle, die noch gesondert unterschieden werden. Sind die Abstände (Differenzen) zwischen Skalenwerten nicht interpretierbar, so spricht man von *ordinalen Merkmalen*. Diese besitzen noch eine Rangfolge, d.h. sind geordnet, aber man kann quantitativ nicht mehr sagen, wo der Unterschied zwischen zwei Ausprägungen liegt. *Nominale Merkmale* verlieren sogar noch die Ordnung, so daß eine Unterscheidung (Benennung) der Ausprägungen der Merkmale, jedoch kein Vergleich zwischen ihnen möglich ist. Ordinale und nominale Merkmale werden oft zu *kategoriellen* oder *qualitativen Merkmalen* zusammengefaßt, denen die quantitativen, d.h. solche mit interpretierbarer Differenz, gegenübergestellt werden.

Binäre Merkmale stellen eine Extremform der diskreten Merkmale dar, da hier nur zwei Ausprägungen (Skalenwerte) gestattet sind. Die binäre Struktur kann zu einer Vereinfachung von Diskriminanzregeln führen. Ihre große Bedeutung rührt von der Ja-Nein-Kodierung her (Eins für Ja, Null für Nein). Qualitative und logische Einflüsse können so numerisch gefaßt und eventuell mit metrischen Daten vorbehaltenen Verfahren analysiert werden.

Beispiel 2.12 *Qualitätskategorien, Zensuren und Wasserhärtegrade sind ordinale Merkmale: klar geordnet, aber mit nur qualitativ, nicht quantitativ faßbaren Differenzen. Blütenfarben, erlernter Beruf, Namen einer Zellkulturlinie, Nationalität oder Land des Firmensitzes haben verschiedenartige Ausprägungen, die aber nicht vergleichbar sind im Sinne von „größer“ oder „mehr“; es sind nominale Merkmale. Geschlecht einer Person und Ja-Nein-Kodierungen sind Beispiele für binäre Merkmale.*

Die Anzahl von Personen oder Dingen ist ein zweischneidiges Beispiel. Eine kleine Anzahl, z.B. 10 Fahrräder, ist klar diskret und besitzt eine diskrete Wahrscheinlichkeitsverteilung. Für Größenordnungen um eine Million dagegen sind fünf Objekte mehr oder weniger nicht wichtig; hier kann eine stetige Beschreibung der Verteilung sinnvoller sein (siehe Kapitel 2.4.2).

Zu Datentypen siehe auch [54], [183], [111].

2.4.2 Behandlung von Daten

Die Behandlung der Daten vor oder während der Analyse kann verschiedene Zielstellungen beinhalten: Voraussetzungen von Verfahren erfüllen, Regeln stabiler machen, die Dimensionalität des Problems erniedrigen, spezielle Richtungen im Datenraum betonen. Man unterscheidet dann oft zwischen den ursprünglichen Merkmalen oder Variablen (*variables*) und davon abgeleiteten Variablen (*features*). Die Bezeichnung „ursprünglich“ ist dabei eher künstlich, denn bei der Erhebung eines Wertes ist a-priori schon über eine Skala entschieden worden, z.B. ob Entfernungen in Metern, Millimetern oder Ångström zu messen oder ob kategorielle Daten mit zwei oder fünf Ausprägungen zu fassen sind.

Transformationen

Transformationen sind angebracht, um entweder die funktionelle Form der Regel zu unterstützen, die Handhabung der Daten zu vereinfachen oder eine gewünschte

Verteilung der Merkmale zu erreichen, die so die Voraussetzungen oder Optimalitätskriterien eines Verfahrens erfüllen.

Beispiele für *Verteilungstransformationen* sind ($c = \text{const.}$):

- *Reziproke Transformation*: z.B. $g(x) = \frac{1}{x+c}$

  Bei Verteilungen mit starker Schiefe oder zeitabhängigen Variablen, etwa Lebensdauern oder Entwicklungszeiten, wodurch sehr lange Zeiträume nahe Null transformiert werden.

- *Wurzel-Transformation*: z.B. $g(x) = \sqrt{x+c}$ oder $g(x) = \sqrt{cx}$

  Bei Häufigkeiten, seltenen Ereignissen (POISSON-Verteilung) und der χ^2-Verteilung, die so annähernd auf Normalität transformiert werden ($\sqrt{x+3/8}$ bzw. $\sqrt{2x}$, [111])

- *Logarithmische Transformation*: z.B. $g(x) = \log x + c$, $g(x) = \log cx$ oder $g(x) = \log \frac{x}{1-x}$

  Bei lognormaler Verteilung (z.B. bei der Modellierung von Gendefekten, [183]), zur Handhabung großer Zahlen oder zur Projektion des Intervalls $[0, 1]$ auf die reelle Achse.

- BOX-COX-*Transformation*: z.B. $g(x) = \begin{cases} \frac{(x+c)^\lambda - 1}{\lambda} & : \text{falls } \lambda \neq 0 \\ \log(x+c) & : \text{falls } \lambda = 0 \end{cases}$

  Zur Transformation auf Normalverteilung (siehe Kapitel 3.2 und 3.3). Den Parameter c erhält man mittels einer ML-Schätzung im Test der Quantile der transformierten Variable gegen die der Normalverteilung ([202] Kap. 1.9.3, [161] Kap. 6.3.2).

- *Winkel-Transformation*: z.B. $g(x) = \arcsin\sqrt{\frac{x+c}{n+d}}$

  Zur Umwandlung von Prozentwerten und binomialverteilten Größen auf Normalität (etwa mit $\sqrt{n+1/2}\arcsin\sqrt{\frac{x+3/8}{n+3/4}}$ varianzstabilisierend, [111]).

Für weitere Transformationen auf Normalität siehe auch [198], Kap. 4.2.

Die *Linearisierung von Funktionszusammenhängen* ist durch monotone, eineindeutige Transformationen von Variablen erreichbar. Das kann für Neuronale Netze (Kapitel 4.5) wichtig sein; für Standard-CART (Kapitel 4.4) ist es durch die Skaleninvarianz unwichtig. Meist wird man die vermutete Umkehrfunktion benutzen. Beispiele sind: $\sqrt{x}$, x^2, e^x, $\log x$, x^{-1} und natürlich Abwandlungen davon.

Die Vielzahl an Transformationen ist praktisch kaum beschränkt. Eine Auswahl kann über Verteilungstests und graphische Methoden, etwa Q-Q-Plots, getroffen werden. Für Details und Anregungen siehe [183], [111] Kap. 4.5, [31], [36], [218].

Mehrdimensionale Skalierung

Die *Mehrdimensionale Skalierung* (MDS) versucht, aus einer gegebenen Distanzmatrix von Objekten deren Lage im Raum zu schätzen, etwa um eine graphische Darstellung zu ermöglichen oder um Daten mit ordinaler oder nominaler Skala in solche mit einer metrischen zu überführen, d.h. den Datentyp zu verändern.

LINDTNER [152] listet für die verschiedenen Datentypen Beispiele von Ähnlichkeits- und Distanzmaßen auf. Für metrische Daten wird oft die MINKOWSKI-Metrik

$$\rho(x,y) = \left(\sum_{j=1}^{p} |x_{(j)} - y_{(j)}|^{\lambda} \right)^{\frac{1}{\lambda}} \tag{2.35}$$

benutzt, die als Spezialfälle die EUKLIDische Metrik ($\lambda = 2$), die L_1- ($\lambda = 1$) und die Supremum-Metrik L_∞ ($\lambda = \infty$) enthält, oder der MAHALANOBIS-Abstand (2.21). Bei binären Variablen werden über Kontingenztafeln verschiedene Koeffizienten gewonnen, z.B. *Simple Matching* für den relativen Anteil an Übereinstimmungen von zwei Merkmalen, um unterschiedliche Korrelationen zu beschreiben. Nominale Variablen werden in binäre überführt; ordinale werden rangnormiert auf $[0,1]$ durch $\frac{r_{ij}-1}{R_i-1}$, wobei R_i den maximalen Rang der i-ten Variable und r_{ij} den Rang des j-ten Objekts in der i-ten Variable angeben. Über eine Eigenwertbestimmung der Distanzmatrix können dann die Datenpunkte in den Raum projeziert (metrische MDS bei L_2-Metrik) oder die Punkte im Raum iterativ gesucht werden (*Kruskal-Shepard-*Algorithmus u.a.), die der Distanzmatrix entsprechen (nichtmetrische MDS, siehe [152], [156], [19]).

Kombinationen von Variablen

Für die meisten Verfahren ist eine geringere Dimension des Merkmalsraumes ein möglicher Weg zu schnelleren und genaueren Ergebnissen. Die Kombination von Variablen ist ein Weg, dieses Ziel zu erreichen und dabei Abhängigkeiten zwischen den Merkmalen in die Analyse mit einzubeziehen. Sie vergrößert aber zunächst die Variablenanzahl. Falls einige wenige hochdiskriminierende Variablen dabei gefunden werden und diese die originalen Variablen nach der anschließenden Variablen- oder Modellwahl (siehe Kapitel 2.5.1) (teilweise) ersetzen, können sie das Risiko der Diskriminanzregel senken.

Die *Kanonische Korrelationsanalyse* (CCA) versucht für zwei Gruppen metrischer Variablen ($x \in \mathbb{R}^{q_1}$, $y \in \mathbb{R}^{q_2}$), solche Linearkombinationen $\eta' x$ und $\varphi' y$ zu finden, so daß η und φ die größtmögliche Korrelation besitzen. Die CCA kann so als Verallgemeinerung der linearen Regression ($q_2 := 1$) gesehen werden. Die Maximierung wird über Eigenwertbestimmungen erzielt (siehe [156], Kap. 10). Die dabei erhaltenen kanonischen Korrelationsvariablen $\eta' x$ und $\varphi' y$ sind unkorreliert und auf Varianz 1 standardisiert. Eine Gruppe von Variablen kann die Klassenzugehörigkeit definieren, so daß man einen Anhaltspunkt für wichtige Variablen erhält — allerdings nur bei linearem Zusammenhang. Das Pendant *Korrespondenzanalyse* tut dasselbe für ordinale und nominale Daten. Dabei werden die originalen Daten in binäre Variablen umgewandelt, die dann in eine modifizierte CCA einfließen (siehe [152], [156] Kap. 10, [15]). Ein übliches Verfahren ist dabei das *Optimal Scoring* bzw. *Optimal Scaling*.

Sind metrische und binäre Variablen stärker korreliert oder spielt eine differenzierte Variable erst bei Erfüllung einer Bedingung eine wichtige Rolle, kann ein *Produkt* aus beiden Variablen die Dimension verringern und nebenbei Probleme bei der Verfahrensanwendung aufgrund gemischter Daten vermeiden.

Beispiel 2.13 *Die Höhe der Blutfettwerte korreliert mit dem Auftreten von Herzinfarkten (siehe z.B. [237], [230], [231]). Frauen sind bis zum Eintritt in das Klimakterium*[28] *besser als Männer vor einem Herzinfarkt geschützt; danach gleicht sich ihr*

[28] Wechseljahre

Risiko dem der Männer schnell an. Eine abgeleitete Variable als Produkt zweier binärer (Frau ja/nein[29], Lebensalter über 45 Jahren ja/nein, d.h. binärisiert) enthält bedingte fachspezifische Information.

Weitere Ansätze sind die *Faktoranalyse* und die *Hauptkomponentenanalyse* (Suche einer größtmöglichen Korrelation innerhalb von Gruppen anstatt zwischen ihnen wie bei der CCA) , die aber nicht unbedingt die besten diskriminatorischen Achsen liefern (siehe [10], [131], [156], [161] Kap. 6.6).

Änderungen des Datentyps

- *diskret* $\rightarrow$ *stetig*: Diskrete Daten können versuchsweise in stetigen Daten vorbehaltenen Verfahren verwendet werden, vor allem bei großen Anzahlen. Oft machen aber eine Approximation, etwa der Binomial- durch eine Normalverteilung, oder eine Transformation Sinn, z.B. über MDS oder logarithmische und Winkel-Transformationen (siehe oben).

- *qualitativ* $\rightarrow$ *binär*: Ein Merkmal mit a Ausprägungen wird durch a binäre Variablen ersetzt. Dabei werden genau die i-te Variable Eins und die anderen Null gesetzt, falls das qualitative Merkmal seine i-te Ausprägung annimmt. Das ist Voraussetzung für die Anwendung der MDS, kann aber auch die Form von Regeln vereinfachen. Das Problem ist dabei die Dimensionalität, da beispielsweise für *eine* qualitative Variable mit 5 Ausprägungen 5 neue binäre Variablen eingeführt werden müssen, sich p also um 4 erhöht.

- *qualitativ* $\rightarrow$ *metrisch*: Unter Ignorierung ihrer Natur wird qualitativen Ausprägungen ein metrischer Wert zugeordnet (Kodierung), so daß sie in Verfahren mit metrischen Daten anwendbar oder auch leichter und effizienter speicherbar sind. Das ist sehr oft üblich in der Medizin, Soziologie und Psychologie, in denen viele qualitative Merkmale erfaßt werden.

- *metrisch* $\rightarrow$ *diskret*: Durch Diskretisierung metrischer Variablen, die die Bedingungen eines Verfahrens nicht erfüllen, erreicht man vielleicht die Erfüllung der (schwächeren) Voraussetzungen eines anderen, nichtmetrischen Verfahrens, oder vermeidet Probleme durch gemischte Daten. Als nichtstatistische Motivation kann eine einfache Interpretierbarkeit gelten.

- *metrisch* $\rightarrow$ *binär*: Metrische Variablen können durch das Setzen eines Schwellenwerts in binäre umgewandelt werden, die so eine qualifizierende Information von fachlicher Seite enthalten können.

  Beispiel 2.14 (Binärisierung) *Setzt man für den Bluthochdruck bei einem Menschen einen (willkürlichen) Schwellenwert von 160 mmHg (systolisch) an, so enthält die resultierende Variable*

$$X_{HT} = \begin{cases} 1: & \textit{systolischer Blutdruck} \geq 160\,mmHg \\ 0: & \textit{systolischer Blutdruck} < 160\,mmHg \end{cases}$$

  bewertende medizinische Information, die mit einer rein mathematischen Analyse nicht greifbar ist (aber diese vielleicht auch verfälscht, siehe Kapitel 2.5.2).

[29] kodiert über 1 für ja, 0 für nein

2.5 Die Modellwahl

Die Wahl des mathematischen Modells für die Diskriminanzanalyse hängt von der Art und dem Inhalt der Daten ab. Dabei gilt die Suche nicht *dem* Optimum, sondern einer Menge von akzeptablen Modellen, aus der dann ein geeignetes Modell, vielleicht ein einfaches oder interpretierbares, ausgewählt wird (siehe Kapitel 2.6).

2.5.1 Verfahren der Modellwahl

Für die Wahl eines Modells $\mathcal{M}$ — also der Menge an eingeschlossenen und vielleicht vorher transformierten Variablen — und die Entscheidung für eine Regel steht neben eventuell vorhandener A-priori-Information nur die Stichprobe zur Verfügung, die zufällig ist. Damit ist die unter Einschluß aller Variablen in $\mathcal{M}$ gewonnene Regel $\hat{\delta}_{\mathcal{M}}$ ebenso wie die Schätzung ihres Risikos $\hat{\mathcal{R}}(\hat{\delta}_{\mathcal{M}})$ zufällig. Setzt man für ein festes Verfahren das optimale Modell $\mathcal{M}^*$ als

$$\mathcal{M}^* := \arg\min_{\mathcal{M}} \hat{\mathcal{R}}(\hat{\delta}_{\mathcal{M}})$$

an, so ist auch dieses optimale Modell zufällig und von der Stichprobe abhängig: $\mathcal{M}^*|\mathcal{S}$. $\mathcal{M}^*$ kann so als Realisierung einer Zufallsgröße angesehen werden, die Modellwahl selbst als Schätzproblem. Der Vergleich zwischen verschiedenen Modellen (und verschiedenen Regeln) ist daher schwierig, weil keine eindeutigen Aussagen möglich sind. Darum ist es sinnvoll, nicht blind das Modell mit dem kleinsten (geschätzten!) Risiko auszuwählen, sondern eine Reihe von Modellen mit ähnlich geringem Risiko (*Kandidaten-Modelle*) einer weiteren Prüfung zu unterziehen und dann die Entscheidung nach Kriterien wie Interpretierbarkeit, Schnelligkeit u.ä. zu treffen (siehe Kapitel 2.6).

Fehlerrate und Variablenanzahl

Warum soll überhaupt eine Variablenauswahl getroffen werden? Da die Verteilungen bzw. Verteilungsquotienten nicht bekannt sind und aus endlichen Stichproben geschätzt werden müssen, existiert ein *Peaking*-Phänomen der Fehlerrate mit zunehmender Dimension p ([128]): Zunächst fällt die Fehlerrate (Bias-Reduktion durch mehr eingeschlossene Merkmale), um dann nach einem Minimum wieder anzusteigen (Varianzerhöhung durch zuviele Parameter bei festem Stichproben-Umfang). Es existiert ein Optimum bzw. ein optimaler Bereich an betrachteten Merkmalen für ein gegebenes Verfahren und eine gegebene Stichprobe. Die Variablenauswahl ist so ein Mittel der Bias-Varianz-Abwägung.

Ein Synonym für die Varianzerhöhung ist der sogenannte „Fluch der Dimensionalität" ([14]): Bei vielen Verfahren nimmt mit wachsender Dimension p die Anzahl der zu schätzenden Parameter sehr schnell zu oder die Datendichte in lokalen Umgebungen um den betrachteten Punkt (wichtig für nichtparametrische Verfahren) sehr schnell ab. Daraus folgt, daß für festes n die Varianz der Schätzungen der einzelnen Parameter immer größer wird und diese im Extremfall nicht mehr identifizierbar sind.[30] Aufgrund des besonderen Fehlerkriteriums in der Diskriminanzanalyse (siehe

[30] Scott ([196], Kap. 7) beschreibt den „Fluch der Dimensionalität" als Spärlichkeit der Daten in hohen Dimensionen: „Wenn die Nachbarschaften ‚lokal' sind, dann sind sie fast sicher ‚leer', und falls sie nicht ‚leer' sind, sind sie nicht ‚lokal'." Lokale Dichteschätzungen benötigen so in höheren Dimensionen ($p \geq 5$) sehr große Stichproben, wie er an mehreren Beispielen demonstriert. Als weiteren Aspekt sieht er Korrelationen zwischen Variablen als Ursache für die zunehmenden Schwierigkeiten mit wachsendem p.

Kapitel 2.3) wird man deswegen versuchen, die Bias-Varianz-Abwägung in Richtung eines stärkeren Bias zu verschieben, um die Varianz klein zu halten. Mittel dazu sind regularisierende Annahmen, etwa die Unabhängigkeit von Merkmalen, oder die Bestrafung zu komplexer Strukturen (mit vielen Parametern) in der Modellwahl; dabei moderiert über möglichst genaue Schätzungen des Risikos $\mathcal{R}$ via Cross-validation oder Testdatensätze. Eine weitere Erscheinung des „Fluches" in hohen Dimensionen ist, daß die Abstände zwischen sämtlichen Objekten der Stichprobe groß und annähernd gleich sind, was einen Verlust an Information über die Verteilungen bedeutet. Für eine ausführliche Diskussion siehe Friedman [72].

Parametrische Tests für eine Variablenauswahl?

In den parametrischen Verfahren lassen sich aufgrund der Verteilungsannahmen oft spezifische Teststatistiken für den Beitrag einzelner Variablen zur Klassifikationsgüte ableiten. Im Falle der Linearen Diskriminanzanalyse (Kapitel 3.2) nimmt Krishnaiah [137] an, daß nur die ersten $q < p$ Variablen wichtig für die Diskrimination sind, stellt als Nullhypothese die Gleichheit der um Korrelationen bereinigten Mittelwerte der letzten $p - q$ Variablen auf und leitet einen Test gegen die $F_{p-q,n-p-1}$-Verteilung ab. Schaafsma [185] robustifiziert diese Auswahlprozedur. Die Nullhypothese unterstellt jedoch eine Reihenfolge der Variablen, die willkürlich festgelegt wurde. Für die Logistische Diskriminanzanalyse (Kapitel 3.4) schlägt Anderson [9] eine schrittweise Auswahlprozedur vor. Da eine ML-Schätzung benutzt wird, testet er die Signifikanz jeder Variablen zur Erhöhung der Likelihood-Funktion durch einen asymptotischen χ^2-Test, ausgehend vom Null-Modell $\mathcal{M}_0$ mit nur einer Konstante ohne jede Variable. Bei Signifikanz wird die Variable $x_{(1)}$ eingeschlossen ($\mathcal{M}_1$), und die restlichen Variablen werden durch einzelne Hinzunahme erneut überprüft ($\mathcal{M}_2$). So erhält man eine Kaskade von Modellen: $\mathcal{M}_1, \mathcal{M}_2, \ldots$. Das Verfahren bricht bei Nichtsignifikanz ab.

Da diese Tests wie die Fehlerverteilungen von den Modellannahmen abhängen, können sie bei Abweichungen in die Irre führen. Hauptproblem bei sequentiellen Testverfahren ist das multiple Testproblem: Die Irrtumswahrscheinlichkeit des Gesamtproblems ist nicht kontrollierbar. Ihr Problem ist ihr eingeschränkter Blick, da sie nicht streng an das eigentlich interessierende Risiko (2.4) gekoppelt sind und immer eine Verfälschung in das Verfahren einbringen können. Das sequentielle Vorgehen impliziert immer eine Ordnung in den Variablen, die ebenfalls willkürlich ist. Sie sind daher eher fragwürdig. Das Leitmotiv ist daher:

> Als allgemeines *Vergleichskriterium* für alle Verfahren sollte eine möglichst genaue Schätzung des eigentlich interessierenden Risikos (2.4) über Resampling-Methoden (Cross–validation, Bootstrap) oder — bei genügend großen Stichproben-Umfängen — über Testdatensätze dienen (siehe Kapitel 2.2.3).

Um die Berechnung zu umgehen, sind auch andere Kriterien in Gebrauch, etwa die Divergenz oder der Chernoff-Abstand ([105], [141]), die aber nicht unbedingt an das Bayes-Risiko gekoppelt sind und daher vermieden werden sollten.

Weg zum Optimum

Ein weiteres Problem ist der *Weg* zur optimalen Menge an Variablen. Die Schätzung des Risikos für alle Teilmengen (*Umfassende Suche* oder *Exhaustive search* durch

$2^p - 1$ Mengen) garantiert das Auffinden der optimalen Teilmenge(n). Für $p = 10$ sind das 1.023, für $p = 15$ schon 32.767 und für $p = 20$ sogar 1.048.575 Mengen, so daß mit zunehmendem p dieser Ansatz an zeitliche Grenzen stößt. Beim heutigen Stand der Rechentechnik ist es aber ohne weiteres möglich, bei $p \leq 20$ für alle $2^p - 1$ Variablenteilmengen das Risiko in annehmbarer Zeit zu bestimmen. Insbesondere bei wirtschaftlichen Anwendungen stehen die Kosten einer Woche Rechenzeit in einem minimalen Verhältnis zu dem Gewinn an Genauigkeit und eventuell geringeren laufenden Kosten aufgrund eines geringeren Meßaufwandes.

Sequentielle Tests (siehe oben) an den irgendwie geordneten Variablen (*Forward Selection* und *Backward Selection*) neigen durch ihre Abbruchkriterien zum Einschluß bzw. Ausschluß zuvieler Variablen und durch die fehlende Kopplung an das BAYESsche Risiko zum Verfehlen des Optimums ([214]).

Für sehr große p können zwei Ansätze zur Verringerung der zu evaluierenden Teilmengen betrachtet werden:

- Falls eine Monotoniebedingung $\mathcal{S}_2 \subset \mathcal{S}_1 \subseteq \mathcal{S} \Rightarrow \mathcal{R}(\mathcal{S}_2) > \mathcal{R}(\mathcal{S}_1)$ erfüllt und außerdem ein $\mathcal{S}_3 \not\subseteq \mathcal{S}_1$ mit $\mathcal{R}(\mathcal{S}_3) < \mathcal{R}(\mathcal{S}_1)$ bekannt ist, müssen keine Teilmengen von $\mathcal{S}_1$ mehr evaluiert werden. Dieser *Branch-and-Bound*-Algorithmus ([168]) über einer Baumstruktur kann sehr viel Rechenzeit sparen.

- Durch Kombination von Forward und Backward Selection, etwa durch *Delete-d-Add-a-Variables*-Prozeduren[31] ([105]), bei der Schätzung des Risikos und Wiederholungen mit verschiedenen Startmengen kann die ausgedehnte Suche vermieden, die implizierte Reihenfolge der sequentiellen Verfahren aufgebrochen und so deren Nachteile gemildert werden.

Beide Ansätze können das Optimum aber nicht garantieren ([214]).

Datentransformationen wie das Bilden von Linearkombinationen oder von Produkten binärer mit metrischen Variablen können die Dimension außerhalb obiger Verfahren nach einer Variablenauswahl verringern (siehe Kapitel 2.4.2). Über graphische Methoden, etwa mittels der Befehle `biplot`, `brush` oder `spin` in der Programmiersprache S-PLUS ([218]), kann man sich ein Bild von den Daten machen und „per Hand“ Merkmale herauslassen (*Explorative Analyse*). Doch auch dies ist relativ willkürlich und kann den entgegengesetzten „Erfolg“ haben (siehe z.B. [198], Kap. 5, 6.10 und 10.1.6). Über die Verwendung mehrerer Verfahren, von Regularisierungen und adaptiven Verfahren hat man die Möglichkeit, die Stichprobe in den wohl meistens vorkommenden unsicheren Fällen über das zu benutzende Modell — d.h. das Modell mit dem geringsten Risiko durch die beste Bias-Varianz-Abwägung — entscheiden zu lassen. Solche Methoden sind im folgenden meist in den mit einem „*“ gekennzeichneten Kapiteln beschrieben.

BUNKE & FISCHER [33] fassen zusammen, daß die letztendliche Wahl einer Diskriminanzregel von der Wahl der Merkmale, des Wahrscheinlichkeitsmodells, der Schätzprozeduren und der Konstruktion der Regel abhängt. In den meisten praktischen Anwendungen sollte die umfassende Suche über viele Verfahren und Variablenmengen einige brauchbare Modelle liefern (siehe Kapitel 2.5.2). Bei großen Dimensionen p scheinen ein zielgerichtetes „Spielen“ mit den Daten und viele Versuche zum Ziel zu führen.

[31] Zunächst werden d Variablen aus der Menge herausgenommen, danach a hinzugefügt. Falls $d < a$ gilt, erhält man damit eine tendenzielle Vorwärts-Selektion, im anderen Fall eine Rückwärts-Selektion.

2.5.2 Die „Berliner Statistische Schule“

Die „Berliner Statistische Schule“[32] zieht aus den Unwägbarkeiten im Kapitel 2.5.1 folgende Schlüsse:

1. Für jede Verteilung existiert immer ein optimales Verfahren. Die Verteilung ist aber unbekannt, so daß das *Verfahren* geschätzt werden muß.

2. Neben Erfahrungen aus einer langjährigen oder wiederholten Anwendung ist die Stichprobe die oft einzig verläßliche Quelle an Information über die Verteilung. Deren Umfang ist beschränkt und damit die Gültigkeit theoretischer Optimalitätskriterien, so daß eine Bias-Varianz-Abwägung für die Schätzung des Verfahrens dessen Risiko senken kann (siehe Kapitel 2.3).

3. An Möglichkeiten einer Abwägung sind gegeben: a-priori die Wahl des Verfahrens, einschränkende Annahmen über die Parameter bzw. Verteilungen (inklusive prädiktiver Zugänge), die Variablenauswahl und die Bestrafung komplexer Strukturen in Optimierungsproblemen. Geeignete Transformationen von Variablen können die Annahmen oder Optimalitätskriterien von Verfahren erfüllen helfen und damit deren Risiko senken.

4. Im Vornherein ist sehr selten die optimale Bias-Varianz-Abwägung bekannt, die zum minimalen Risiko der Diskriminanzregel führt. Optimal wäre eine Suche des optimalen Verfahrens über eine Vielzahl an Verfahren, eine Vielzahl von Regularisierungen und alle möglichen Teilmengen an originalen und abgeleiteten Variablen. Dies ist aber praktisch nicht möglich. Stattdessen sollte in einer möglichst großen Klasse von Verfahren und (transformierten) Variablen gesucht werden.

5. Hinweise, unter bestimmten Verteilungsannahmen dieses oder jenes Verfahren zu benutzen und also andere Verfahren außen vor zu lassen, bergen die Gefahr einer Einschränkung der betrachteten Verfahrensklasse auf nur noch unzulässige Verfahren in sich, d.h. das prinzipielle Auffinden nur stark suboptimaler Lösungen. Analoges gilt für sequentielle Auswahlverfahren der Variablen (siehe Kapitel 2.5.1).

6. Das primäre Interesse bei einer Diskriminanzregel besteht in ihrem minimalen Bayesschen Risiko. Für den Vergleich der Verfahren und Variablenmengen bildet dessen Schätzung via Cross-validation ein hinreichendes Gütekriterium, da sie ein Maß für das Ergebnis der Bias-Varianz-Abwägung liefert. Es existiert kein Aufschaukeln der Fehler, so daß auch große Verfahrensklassen und Variablenmengen aufgrund fester, moderater Stichprobenumfänge verglichen werden können (siehe Kapitel 2.2.3).

7. Die Rechnerleistungen sind in den vergangenen Jahrzehnten extrem angewachsen — und bis jetzt ist ein Ende dieses Trends noch nicht abzusehen —, so daß eine umfassende Suche über Verfahrensklassen und Variablenmengen möglich geworden ist: Das Zeitproblem für das Finden einer Diskriminanzregel verschwindet für die meisten Anwendungen in zunehmendem Maße. Die Größe der untersuchten Verfahrens- und Variablenmengen orientiert sich so an den rechentechnischen Analysemöglichkeiten.

[32] mündliches Zitat von O. Bunke in einem persönlichen Gespräch vom 30. Juni 1999

8. Die umfassende Suche wird selten das beste Verfahren finden, aber immer ein sehr gutes.

O. Bunke nennt dies die *Adaptive Diskriminanzanalyse*[33], da sowohl Verfahren als auch Variablenauswahl an die Daten angepaßt werden.

Diese Philosophie stammt von der Regressionsanalyse her. Arbeiten in diesem Geist sind zum Beispiel [32], [33], [35], [36], [37].

2.6 Fragen an eine Diskriminanzregel

Für die Verfahrensauswahl gibt es neben den Fehlklassifikationskosten und den Datenarten weitere Kriterien, die eine Rolle spielen. Unter den gefundenen guten Modellen mit vergleichbarem Risiko muß oft eine Auswahl nach weiteren Gesichtspunkten getroffen werden. Wichtige Fragen dabei sind:

- *Analyseziel* (vorher): Welche Entscheidung soll aufgrund der Analyse getroffen werden? Reicht ein Ja-Nein-Urteil (2-Klassen-Problem), oder ist eine größere Differenzierung notwendig? Sollen nur diskriminierende Faktoren gefunden werden, oder sollen aufgrund einer Regel neue Objekte oder Personen klassifiziert werden?

- *Genauigkeit*: Liegt die Regel in den geforderten Fehlergrenzen? Ist die Regel durch menschliche Fähigkeiten zu verbessern oder diesen unterlegen? Sind geringere Fehlerkosten durch die Natur der Daten überhaupt möglich, weil sich z.B. die Populationsverteilungen stark überlappen? Vor allem, wenn die Vorhersage das Analyseziel ist, sind möglichst genaue Verfahren wünschenswert.

  Hier ist auch das Zulassen von *Indifferenzbereichen* zu überlegen, d.h. Regionen, in denen die Regel nicht zuordnet (Reject-Option). Fälle aus unsicheren Regionen werden erneut betrachtet, um anhand weiterer Analysen eine Zuordnung zu treffen (sequentielles Vorgehen). Auch das kann das Risiko senken.

  Beispiel 2.15 *Für eine Entscheidung, ob eine genetisch bedingte Erkrankung oder ein Risikos dafür vorliegt, wird man bei Patienten, bei denen die phänotypischen Merkmale für eine Klassifizierung noch nicht ausreichen, einen Gentest — so verfügbar — durchführen. Siehe auch [243], [242], [238], [239], [240], [236], [235].*

- *Geschwindigkeit*: Wie lange benötigt die Regel für eine Diskrimination, und wie lange benötigt man für die Aufstellung der Regel? Ist die Regel iterativ anzupassen, wenn weitere Informationen vorliegen? Bei Echtzeit-Entscheidungen sind schnelle Verfahren auf Kosten einer erhöhten Fehlerrate vielleicht vorzuziehen. Bei zeitlich abgekoppelten Analysen kann der Datenumfang, z.B. bei astronomischen oder metereologischen Daten, ein schnelleres Verfahren erzwingen.

- *Interpretierbarkeit*: Die Interpretierbarkeit einer Regel kann ein wichtiger Faktor sein. Inhaltlich können durch fachspezifisches Wissen fehlerhafte Modellannahmen entdeckt werden (und umgekehrt!). Vor allem, wenn das Analyseziel

[33] persönliche Mitteilung vom 30. Juni 1999

im Finden von diskriminatorischen Faktoren besteht, sind Verfahren mit interpretierbarer „klarer" Struktur vorzuziehen.

Als zweiter Punkt zählt aber auch die „Vermarktung", weil mathematische Ergebnisse oft vor externen Auftraggebern oder leitenden Angestellten als Laien präsentiert und gerechtfertigt werden müssen oder das Modell in einen fachspezifischen theoretischen Kontext passen soll.[34]

- *Stabilität I*: Was passiert bei Abweichungen von den Modellannahmen? Falls nicht bekannt ist, ob zukünftige Beobachtungen den Modellannahmen noch genügen, sollte die Wahl vielleicht auf ein robustes Verfahren fallen, um ein unkalkulierbares Verhalten im Sinne eines stark erhöhten Risikos der Regel zu vermeiden.
- *Stabilität II*: Wie stark ist die Regel von der vorliegenden Stichprobe abhängig? Das Problem der Variabilität von Regeln kann zu deren Ausschluß führen.
- *Speicherbedarf*: Während bei parametrischen Verfahren nur die relativ wenigen Parameter für die Anwendung der Diskriminanzregel auf eine neue Beobachtung verfügbar gehalten werden müssen, setzen einige nichtparametrische die gesamte Trainingsstichprobe voraus.
- *Kosten der Datengrundlage*: Sind die Kosten für die Erhebung einzelner Merkmale unterschiedlich, wird man bei annähernd gleichguten Verfahren das mit den kostengünstigeren Variablen wählen; ebenso bei hochkorrelierten Variablen.

Die letztendliche Wahl einer Diskriminanzregel ist so ein Kompromiß aus den obigen Anforderungen.

[34] „Das unabweisbare Bedürfnis, mit einer Erkenntnis glaubwürdig zu wirken, stellt sich spätestens in dem Augenblick ein, wo sie praktisch werden und menschliche Lebensvollzüge bestimmen soll. Der Besitz der Wahrheit ist dann gar nicht entscheidend, sondern allein noch die Frage, ob diese Wahrheit auch akzeptabel ist, also vermittlungsfähig. Glaubwürdigkeit muß als zusätzliche Qualität zur Wahrheit hinzukommen, ist ihr nicht etwa schon wesentlich." ([246], S. 80)

Kapitel 3

Parametrische Verfahren

Parametrische Verfahren zeichnen sich dadurch aus, daß sie Annahmen über die globale parametrische Form der klassenspezifischen Dichten, die Form der Trennfläche oder die Form des Dichtequotienten machen. Ihre Anwendung kann sinnvoll sein, wenn diese Annahmen korrekt sind oder wenn ein verursachter Bias die Reduktion der Varianz nicht übersteigt.

3.1 Die FISHERsche Diskriminanzanalyse

Die FISHERsche Diskriminanzanalyse hat ihren Ursprung in einer Arbeit von FISHER [63] aus dem Jahre 1936. Grundidee ist die Bestimmung einer *linearen* Trennfläche $a'x$ zwischen den beiden Klassen (also Geraden im Zweidimensionalen und Hyperebenen im Mehrdimensionalen). Die Fläche soll dabei so gewählt werden, daß der quadratische Abstand zwischen den Erwartungswerten μ_i der Klassen relativ zur angenommenen gemeinsamen Variabilität Σ beider Klassen in den einzelnen Merkmalen maximal wird. Das wird für zwei Klassen Π_1 und Π_2 durch die sogenannte FISHER*sche Diskriminanzfunktion*

$$a' = (\mu_1 - \mu_2)'\Sigma^{-1} \tag{3.1}$$

erreicht ([131], Kap. 11). Der mittlere Punkt $\frac{1}{2}(\mu_1 + \mu_2)$ zwischen den Erwartungswerten, den die Trennfläche $a'x$ dabei durchläuft, liefert den Schwellenwert $m = (\mu_1 - \mu_2)'\Sigma^{-1}\frac{1}{2}(\mu_1 + \mu_2)$ für die Diskriminanzregel. Je nach dem, ob für ein zu klassifizierendes Objekt x der Wert von $a'x$ größer oder kleiner als m ist, wird es Π_1 oder Π_2 zugeordnet:

$$\delta_{\text{FISHER}}(x) = \begin{cases} 1: & (\mu_1 - \mu_2)'\Sigma^{-1}x \geq m \\ 2: & (\mu_1 - \mu_2)'\Sigma^{-1}x < m \end{cases} . \tag{3.2}$$

Die μ_i können durch die klassenspezifischen Stichprobenmittel

$$\hat{\mu}_i = \bar{x}_i := \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i_j} \quad (i = 1, 2) \tag{3.3}$$

und die gemeinsame Kovarianz Σ durch die gepoolte Stichprobenkovarianz

$$\hat{\Sigma} = S_{pooled} := \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{3.4}$$

mit

$$\hat{\Sigma}_i = S_i := \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{i_j} - \bar{x}_i)(x_{i_j} - \bar{x}_i)' \quad (i = 1, 2) \tag{3.5}$$

geschätzt und in die Regel eingesetzt werden. Die Schätzungen sind erwartungstreu, suffizient ([149], Kap. 1) und konsistent.

Die FISHERsche Regel macht keine Annahmen über die Form der Dichten, sondern setzt nur metrische Werte, gleiche Kovarianzen in den Klassen und eine lineare Trennfläche voraus. Sie impliziert jedoch, daß die Klassenverteilungen vollständig durch die Erwartungswerte und Kovarianzen beschrieben werden, wofür die Normalverteilung ein typisches Beispiel ist. Sie läßt weder den Einfluß von Kosten noch von verschiedenen A-priori-Wahrscheinlichkeiten zu.

3.2 Die Lineare Diskriminanzanalyse

3.2.1 Das Verfahren

Die FISHERsche Diskriminanzregel läßt sich im Sinne von Kosten schnell verallgemeinern. Nimmt man multivariate Normalität der beiden klassenspezifischen Verteilungen bei gleichen Kovarianzmatrizen $\Sigma = \Sigma_1 = \Sigma_2$ an,

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)'\Sigma^{-1}(x - \mu_i)\right) \quad (i = 1, 2), \tag{3.6}$$

so liefert die BAYESsche Regel (2.10) das Optimum im Sinne des Risikos (2.4) als

$$\begin{aligned} \delta_0(x) = 1 \quad &\Longleftrightarrow \quad \tfrac{f_1(x)}{f_2(x)} \geq \tfrac{\pi_2 c_{21}}{\pi_1 c_{12}} \\ &\Longleftrightarrow \quad \exp\left(-\tfrac{1}{2}\|x - \mu_1\|_{\Sigma^{-1}} + \tfrac{1}{2}\|x - \mu_2\|_{\Sigma^{-1}}\right) \geq \tfrac{\pi_2 c_{21}}{\pi_1 c_{12}} \\ &\Longleftrightarrow \quad (\mu_1 - \mu_2)\Sigma^{-1}x - \tfrac{1}{2}(\mu_1 - \mu_2)\Sigma^{-1}(\mu_1 + \mu_2) \geq \log\left(\tfrac{\pi_2 c_{21}}{\pi_1 c_{12}}\right) \end{aligned}$$

und $\delta_0(x) = 2$ sonst (siehe [131], Kap. 11, oder [10], Kap. 6.8). Falls die Quotienten $\frac{c_{21}}{c_{12}}$ und $\frac{\pi_2}{\pi_1}$ reziprok[1] sind, ergibt sich die FISHERsche Regel (3.2). Die Trennfläche zwischen den beiden Populationen ist linear, da die Trennfunktion linear ist; man spricht deswegen von der *Linearen Diskriminanzanalyse* (LDA). In der älteren Literatur über multivariate Analyse wird dieses Verfahren auch einfach nur als Diskriminanzanalyse bezeichnet. Für ein Beispiel siehe Abbildung 3.1 auf S. 43.

Durch die Schätzungen (3.3) und (3.4) erhält man aus der Trainingstichprobe mittels Plug-In die geschätzte lineare Diskriminanzregel als

$$\hat{\delta}_{LDA}(x) = \begin{cases} 1: & (\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) \geq \log\left(\frac{\pi_2 c_{21}}{\pi_1 c_{12}}\right) \\ 2: & \text{sonst} \end{cases}, \tag{3.7}$$

d.h. der MAHALANOBIS-Abstand Δ^2 (2.21) wird als Metrik benutzt, um die Beobachtung x der am nächsten liegenden Klasse bezüglich des Mittelwertes zuzuordnen (mit der Konstante $-\log\left(\frac{\pi_2 c_{21}}{\pi_1 c_{21}}\right)$, die für Kosten und Anteile korrigiert). Der Term links vom „$\geq$“ wird oft als ANDERSONs *Klassifikationsstatistik W* bezeichnet. Die Ausrichtung der Trennfläche ist dieselbe wie bei FISHER, sie ist nur durch die Konstante verschoben.

[1] d.h. Produkt gleich Eins

Bei $\pi_1 c_{12} = \pi_2 c_{21}$ ist die LDA äquivalent ([161], Kap. 5.4) zu

$$\hat{\delta}_{LDA}(x) = \arg\min_i (x - \bar{x}_i)' S_{pooled}^{-1} (x - \bar{x}_i). \tag{3.8}$$

3.2.2 Eigenschaften

Vorteile der Linearen Diskriminanzanalyse sind ihre einfache Struktur und Interpretierbarkeit. Im Fall $k = 2$ ist die Trennfläche eine Gerade ($p = 2$) oder eine Hyperebene ($p > 2$) im Merkmalsraum. Die LDA ist einfach anzuwenden; einmal geschätzt, muß nur der Wert einer linearen Funktion bestimmt werden. Weiterhin ist die LDA unter homoskedastischer Normalverteilung zulässig, d.h. nicht mehr gleichmäßig über alle Parameter zu verbessern ([149], Kap. 1), und asymptotisch optimal ([161]). Falls der Quotient $\log\left(\frac{f_1(x)}{f_2(x)}\right)$ linear bleibt, ist sie auch robust gegenüber Abweichungen von der Normalität ([161]; siehe auch Kapitel 3.4). Sie toleriert Verteilungen mit milder Schiefe, besonders, falls die Klassendichten f_1, f_2 dieselbe Form haben; ebenso Verteilungen, die leichtere Enden als die Normalverteilung haben.

Die LDA ist nicht robust gegenüber stärkerer Schiefe, gepaart mit schwereren Enden als denen der Normalverteilung. Interaktionen zwischen den Variablen verschlechtern das Risiko. Für größere p wird die Regel instabil: Bei festen Stichprobenumfängen, unbekanntem Σ und beschränktem MAHALANOBIS-Abstand (2.21) tendiert die Fehlerrate für $p \uparrow n_1 + n_2 - 2$ gegen $\frac{1}{2}$ ([149], Kap. 1). Insgesamt sollten Variablen, wenn möglich, vorher auf Normalität transformiert werden (siehe Kapitel 2.4.2).

Durch die Voraussetzung der Normalverteilung ist die LDA eigentlich stetigen Variablen vorbehalten. Man kann aber versuchen, auch diskrete Variablen einzuschließen in der Hoffnung, ein gutes Ergebnis zu erzielen; es gibt dafür Beispiele. Für rein diskrete Daten ist die LDA bei ähnlichen Interaktionen zwischen den Merkmalen in beiden Klassen robust. Bei Interaktionen erster Ordnung ist die QDA (Kapitel 3.3) vorzuziehen ([161], Kap. 5.6). KRZANOWSKI [139] schließt aufgrund von Simulationen, daß geringe Korrelationen zwischen diskreten und stetigen Merkmalen in der einen Population, aber hohe in der anderen bzw. ein Wechsel des Vorzeichens zwischen beiden ein Indikator unvorteilhafter Bedingungen für die LDA sein können. PRESS & WILSON [173] empfehlen bei Nichtnormalität aufgrund zweier Studien die Logistische Diskriminanzanalyse (Kapitel 3.4). Bei einer Mischung von binären und normalverteilten Merkmalen schlägt KRZANOWSKI [138] ein spezielles Lokationsmodell vor, siehe Kapitel 4.1.2. Hier ist die Interaktionsstruktur zwischen den stetigen Daten wichtiger als zwischen den binären. Bei rein binären Daten und fehlender Monotonie des Quotienten $\log\left(\frac{f_1(x)}{f_2(x)}\right)$ wird die LDA schlecht ([198], Kap. 6.3).

Für gemischt binäre und stetige Daten $(x^{(1)}, x^{(2)})$ schlägt KRZANOWSKI [140] eine lineare Transformation der stetigen Daten vor, um eine möglichst geringe Korreliertheit zwischen beiden Datengruppen und homogene Mittelwerte in allen Zellen für eine Klasse zu erhalten. Das wird durch $a'x$ erreicht, wobei sich a durch das Minimieren von $\frac{a'\Sigma_C a}{a'\Sigma^{(2)} a}$ und $\frac{a'\Sigma_C a}{a'\Sigma_B a}$ bestimmt (mit $\Sigma^{(2)}$ als der Gesamt-Kovarianz in den stetigen Merkmalen, Σ_C als der Kovarianz zwischen den Zellen einer Klasse und Σ_B als der Kovarianz zwischen den Klassen in gleichen Zellen). Die transformierten stetigen Daten schätzen dann ohne die binären die Diskriminationsregel (3.7).

Die LDA hat nicht nur als eigenständiges Diskriminanzverfahren Bedeutung; aufgrund ihrer Robustheit und Einfachheit gibt es vielfältige Versuche der Kombination mit anderen Verfahren (siehe die Verweise am Ende dieses Teilkapitels).

3.2.3 Regularisierungen*

Da die LDA zulässig ist, sind Verbesserungen im Sinne des Risikos nicht über alle Parameter, sondern nur über Teilmengen zu haben. Durch einschränkende Annahmen (Regularisierungen) kann die LDA dort teilweise verbessert werden.

Eine starke Regularisierung ist die Annahme von unabhängigen Variablen mit gleicher Varianz in beiden Populationen, d.h.

$$\Sigma = \sigma^2 I_p, \tag{3.9}$$

wobei I_p die p-dimensionale Einheitsmatrix darstellt. Durch Ignorieren von Σ wird so die normale EUKLIDische Metrik als Abstandsmaß benutzt. Dieses Vorgehen ist oft in Pattern-Recognition-Anwendungen üblich. Bei großen p kann es zu besseren Ergebnissen führen, falls sphärische Neutralität[2] vorliegt oder der Quotient aus MAHALANOBIS- und EUKLIDischem Abstand klein ist. Durch die starke Verringerung der Parameterzahl auf Eins bleibt Σ auch für große p noch identifizierbar; die Varianz verringert sich bei möglicher Zunahme des Bias.

Etwas schwächer ist die Annahme der Unkorreliertheit der Variablen, so daß die Kovarianzmatrix die Gestalt

$$\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2) \tag{3.10}$$

annimmt. Dadurch reduziert sich die Anzahl der für $\hat{\Sigma}$ zu schätzenden Parameter von $\frac{(p+1)p}{2}$ auf p. So stehen bei festem n für jeden zu schätzenden Parameter mehr Beobachtungen zur Verfügung, wodurch die Varianz der Schätzung $\hat{\Sigma}$ geringer wird, der Bias allerdings durch die Annahme der Unkorreliertheit wachsen kann. Bei Hinzunahme weiterer Variablen steigt die Parameteranzahl außerdem nur linear, während beim klassischen Ansatz ein Zuwachs von quadratischer Ordnung resultiert. Σ bleibt daher bei Diagonalgestalt mit wachsendem p länger identifizierbar und ist für kleine n alternativlos. Die Diagonalgestalt von Σ ist äquivalent zu einer vorherigen Standardisierung der Varianz der einzelnen Merkmale auf Eins und der anschließenden Benutzung des EUKLIDischen Abstandes.

Durch beide obigen Annahmen wird die Bias-Varianz-Abwägung (siehe Kapitel 2.3) schon durch den Schätzansatz vollzogen.

Ridge- und kombiniertes Verfahren nach J. LÄUTER

J. LÄUTER [149] hat Parameterrestriktionen zur Stabilisierung der LDA untersucht. Für verschiedene Modelle, z.B. daß sämtliche Einflußvariablen von einer nicht beobachtbaren, *latenten* Variable abhängen (Ein-Faktor-Modell) und Erweiterungen davon, erhält er zulässige, stabilisierte und skaleninvariante Verbesserungen der LDA. Diese liefern jedoch nur für extrem kleine Stichproben ($n \approx 5$) bessere Ergebnisse im Sinne der Gesamtfehlerrate; er empfiehlt sie nur zur Glättung und nicht zur breiten Anwendung in der Diskriminanzanalyse. Dort sieht er eher Platz für ein Ridge[3]- und ein kombiniertes Verfahren:

Die Idee der *Ridge*-Methode ist wieder eine Glättung. Man nimmt an, daß die Inverse der Kovarianzmatrix, Σ^{-1}, zufällig um ein Vielfaches einer Matrix Σ_0^{-1} als Erwartungswert gemäß einer WISHART-Verteilung ([167]) verteilt ist: $\Sigma^{-1} \sim W_p((\nu\Sigma_0)^{-1}, \nu)$. Benutzt man anstatt S_{pooled} nun $S_{pooled} + \nu\Sigma_0$ in der Regel (3.7),

[2] keine unterschiedlichen Varianzen für einzelne Merkmale

[3] zu deutsch: Kamm, Grat (Das „Erheben“ der Diagonale in der Kovarianzmatrix ist damit gemeint.)

erhält man eine verfälschte und geglättete Schätzung ([149], Kap. 4). Probleme sind die Wahl von ν und Σ_0. Hier empfiehlt J. LÄUTER [149]

$$\nu := \frac{p(n-2)}{(n-4)(n+p-3)} \quad \text{und} \quad \Sigma_0 := \operatorname{diag}\left[S(\operatorname{diag} S^{-1})S\right], \tag{3.11}$$

mit $S := S_{pooled}$. Für $n \to \infty$ nähert sich diese *Ridge*-Regel R der normalen Plug-In-Regel an, für wachsendes p wird der Ridge-Anteil wichtiger. Bei Ein-Faktor-Modellen ist diese Regel besser, bei zufälligen Strukturen in den Variablen schlechter als die LDA. Ersetzt man S_{pooled} durch Σ_0, erhält man die *Mehr-Faktor*-Regel MF (siehe [149], Kap. 7.1). Die Regel R läßt sich verallgemeinern, wenn man den Ridge-Anteil noch mit einer weiteren Konstante als Steuerparameter multipliziert: $c\nu\Sigma_0$. Durch Schätzen des c über die Leave-one-out-Methode (Kapitel 2.2.3) erhält man so eine adaptive Regel, die LÄUTER als Alternative zur klassischen Variablenauswahl der LDA vor allem bei Variablen mit hohem univariaten Trennvermögen sieht ([149], Kap. 7.2).

Eine kombinierte Regel COM entscheidet mittels der Apparent Error Rate (2.23) über die endgültige Wahl zwischen mehreren Verfahren ([149], Kap. 7.5). Falls

$$0.7 * \hat{\mathcal{R}}_{app}(\text{MF}) < \hat{\mathcal{R}}_{app}(\text{LDA}) < \hat{\mathcal{R}}_{app}(\text{R}) \tag{3.12}$$

gilt, wird die klassische LDA gewählt, da sie als hinreichend stabil und die Variablen als nicht zu stark korreliert angesehen werden. Anderenfalls wird das Ridge-Verfahren, eventuell mit einem Selektionsverfahren, das über Bilinearkombinationen von Variablen auch deren gegenseitige Glättung zuläßt, benutzt. In 20 Problemen mit unterschiedlichen Datenstrukturen erhielt LÄUTER für die Regeln R und COM durchgängig gleiche oder bessere Ergebnisse als für die klassische LDA.

Penalized Discriminant Analysis

Für sehr viele, hochkorrelierte Merkmale ($p \geq 200$), wie sie bei der Diskretisierung analoger Signale, etwa in der Bildverarbeitung oder Spracherkennung, entstehen, schlagen HASTIE et al. [114] die sogenannte *Penalized Discriminant Analysis* (PDA) vor. Anstatt wie beim Ridge-Ansatz global zu glätten, soll hier nur eine lokale Glättung stattfinden. Das ist technisch dadurch begründet, daß die gewöhnliche Kovarianzschätzung durch die hohe Parameteranzahl unrealistisch, instabil und schnell nicht identifizierbar ist ($p = 200 \Rightarrow 20.100$ Parameter!) sowie durch die Bevorzugung ungeglätteter Variablen durch die Invertierung von Σ. Inhaltlich macht dies ebenfalls Sinn, weil z.B. Graustufen eines Bildes in bestimmten Gebieten hochkorreliert sind und man mit dem gewöhnlichen Ansatz diese Information verschenkt. Deswegen empfehlen die Autoren, die Schätzung der gemeinsamen Kovarianzmatrix durch eine regularisierte Form zu ersetzen:

$$\hat{\Sigma} = S_{pooled} + \lambda\Omega. \tag{3.13}$$

Dabei ist Ω eine Strafmatrix, die z.B. den lokalen Kontrast, etwa Differenzen im zweiten Moment, bestraft. Zur Bestimmung von Ω zeigen sie die Äquivalenz mit Optimalitätskriterien von Optimal Scoring und Kanonischer Korrelationsanalyse (siehe Kapitel 2.4.2).

Weitere Ansätze

Für gemischt normale und binäre Daten hat KRZANOWSKI [138] ein spezielles Lokationsmodell entwickelt (siehe Kapitel 4.1.2). HASTIE & TIBSHIRANI [118] kombinieren das Nearest-Neighbours-Verfahren mit der LDA, um eine adaptive Version zu erhalten (siehe Kapitel 4.3). Loh & Vanichsetakul [154] versuchen, durch Benutzung der LDA den CART-Algorithmus zu beschleunigen (siehe Kapitel 4.4.3). Verallgemeinerungen der LDA, die aber zu nichtlinearen Trennflächen führen und z.T. nichtparametrische Verfahren benutzen, sind im Kapitel 5.2.2 dargestellt. Rigby [179] modifiziert die Schätzung von $\frac{f_1(x)}{f_2(x)}$, so daß sie weniger extrem und nach seinen Worten „mehr vertrauenswürdig" ist. Huber [125] versucht, die Schätzung der Hyperebene durch Benutzung des Medians anstelle des Mittelwertes gegen Ausreißer zu robustifizieren.

3.3 Die Quadratische Diskriminanzanalyse

3.3.1 Das Verfahren

Die Lineare Diskriminanzanalyse (LDA, Kapitel 3.2) basiert auf der Annahme von Normalverteilungen mit identischen Kovarianzmatrizen für die beiden Klassen. Sie kann durch das Zulassen von $\Sigma_1 \neq \Sigma_2$ für Normalverteilungen verallgemeinert werden. Die Bayessche Regel (2.10) liefert die optimale Regel als

$$\delta_0(x) = 1 \iff -\frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x - C \geq \log\left(\frac{\pi_2 c_{21}}{\pi_1 c_{12}}\right)$$

und $\delta_0(x) = 2$ sonst, wobei

$$C := \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)$$

gilt ([131], Kap. 11). Diese quadratische Funktion definiert eine ebensolche Trennfläche; dieses Verfahren wird deshalb als *Quadratische Diskriminanzanalyse* (QDA) bezeichnet. Mit den üblichen Schätzungen (3.3) und (3.5) lautet die Regel:

$$\hat{\delta}_{QDA}(x) = \begin{cases} 1: & -\frac{1}{2}x'(S_1^{-1} - S_2^{-1})x + (\bar{x}_1'S_1^{-1} - \bar{x}_2'S_2^{-1})x - \hat{C} \geq \log\left(\frac{\pi_2 c_{21}}{\pi_1 c_{12}}\right) \\ 2: & \text{sonst} \end{cases} \tag{3.14}$$

mit

$$\hat{C} := \frac{1}{2}\log\left(\frac{|S_1|}{|S_2|}\right) + \frac{1}{2}(\bar{x}_1'S^{-1}\bar{x}_1 - \bar{x}_2'S^{-1}\bar{x}_2). \tag{3.15}$$

Ein Beispiel für die QDA ist in Abbildung 3.2 auf S. 44 dargestellt.

3.3.2 Eigenschaften

Die QDA ist asymptotisch optimal unter der heteroskedastischen Normalverteilung ([161]). Die Trennfläche ist nicht mehr linear, aber immer noch einfach interpretierbar: Für $k = 2$ und $p \geq 2$ etwa nimmt sie eine elliptische, parabolische oder hyperbolische Form an. Die QDA ist robust bei geringer Schiefe der Klassendichten; auch starke Kurtosis oder Verteilungen mit leichteren Enden als denen der Normalverteilung sind kein Problem. Bei symmetrischen Verteilungen mit schwereren Enden als bei der Normalverteilung können sich einzelne $\mathcal{R}_i$ (2.3) erhöhen, das Gesamtrisiko bleibt aber kaum beeinflußt. Die QDA ist aufgrund der zu schätzenden Parameteranzahl eher für große n_i geeignet ([161], Kap. 5.6), bei kleinen wird sie aufgrund der Varianz der Schätzung schlechter als die LDA (Kapitel 3.2). Seber ([198], Kap. 6.3) gibt als Empfehlungen: für kleine Stichprobenumfänge oder kleine Unterschiede in den Σ_i Benutzung der LDA, für große n_i, $p > 6$ und große Unterschiede Benutzung der QDA.

Bei schiefen Verteilungen mit schwereren Enden als denen der Normalverteilung wird die QDA schlecht. Falls die n_i klein bezüglich p sind, wird die Schätzung S_i sehr variabel und damit die geschätzte Trennfläche. Für $n_i \leq p$ ist sie nicht mehr identifizierbar. Insgesamt sollten Variablen, wenn möglich, vorher auf Normalität transformiert werden (siehe Kapitel 2.4.2). Bei Homoskedastitzität verliert die QDA durch die Mehrheit an Parametern gegenüber der LDA an Effizienz ([161], Kap. 5.3). Bei binären Daten wird die QDA schlecht ([198], Kap. 6.3).

3.3.3 Regularisierungen*

Eine Form der Regularisierung ist die Annahme gleicher Kovarianzmatrizen, also eine Verfälschung der Schätzungen $\hat{\Sigma}_1$ und $\hat{\Sigma}_2$ durch $\hat{\Sigma}_{pooled}$ bei vorliegender Heteroskedastizität. Man erhält so die Lineare Diskriminanzanalyse (Kapitel 3.2). Es existieren aber auch gleitende Übergänge zwischen Quadratischer und Linearer Analyse:

Ähnliche Kovarinazmatrizen

Der erste grundsätzliche Zugang ist die Annahme von Ähnlichkeiten zwischen den Kovarianzen der beiden Gruppen. Nimmt man etwa Proportionalität der Σ_i an,

$$\Sigma_2 = \kappa^2 \Sigma_1, \tag{3.16}$$

sind $\frac{1}{2}p(p-1)$ Parameter weniger zu schätzen. Die ML-Schätzung unter p-variater Normalverteilung lautet:

$$\begin{aligned}
\hat{\Sigma}_1 &= \frac{1}{n}\left(n_1\hat{\Sigma}_1 + \frac{n_2}{\hat{\kappa}^2}\hat{\Sigma}_2\right) \\
\hat{\kappa}^2 &= \left[\mathrm{tr}\hat{\Sigma}_1^{-1}\hat{\Sigma}_2\right]^{\frac{1}{2}}.
\end{aligned}$$

Die iterative Lösung (Start mit $\kappa = 1$) existiert und konvergiert eindeutig ([161], Kap. 5.4). Vergröbernd kann man auch annehmen, daß nur die Gruppen-Korrelationsmatrizen gleich sind, d.h.

$$\Sigma_2 = K\Sigma_1 K \quad (K = \mathrm{diag}(\kappa_1, \ldots, \kappa_p)). \tag{3.17}$$

Die ML-Schätzung unter p-variater Normalverteilung lautet dann:

$$\begin{aligned}
\hat{\Sigma}_1 &= \frac{1}{n}\left(n_1\hat{\Sigma}_1 + K^{-1}\hat{\Sigma}_2 K^{-1}\right) \\
\hat{\kappa}_\nu &= \sum_{j=1}^{p} \frac{1}{\hat{\kappa}_j}\left(\hat{\Sigma}_1^{-1}\right)_{j\nu}\left(\hat{\Sigma}_2\right)_{j\nu} \quad (\nu = 1, \ldots, p).
\end{aligned}$$

Auch hier konvergiert die iterative Lösung (Start mit $\kappa_j = 1,\ j = 1, \ldots, p$), aber eventuell sehr langsam ([161], Kap. 5.4). Beide Verfahren sind skaleninvariant.

Common Principal-Component Model (CPC)

Ein weiterer Ansatz der Parameterreduktion ist die Annahme von gemeinsamen Hauptkomponenten: Die Σ_i besitzen dieselben prinzipiellen Achsen mit Unterschieden in Größe und Rang. In diesem CPC-Modell ([65]) sind die Kovarianzmatrizen also durch dieselbe orthogonale Matrix $A \in \mathcal{M}_{p\times p}$ diagonalisierbar:

$$A\Sigma_i A = \Lambda_i \quad (i = 1, 2) \tag{3.18}$$

mit $\Lambda_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$. Die ML-Schätzung, wiederum unter p-variater Normalverteilung, lautet:

$$a_l'\tilde{S}_{lm}a_m = 0 \qquad (l, m = 1, \ldots, p \quad l \neq m) \tag{3.19}$$

$$\tilde{S}_{lm} = \sum_{i=1}^{2}(n_i - 1)\frac{\lambda_{il} - \lambda_{im}}{\lambda_{il}\lambda_{im}}S_i \tag{3.20}$$

mit der Nebenbedingung $a_l'a_m = \delta_{l,m}$[4], wobei $A = (a_1, \ldots, a_p)$ gelte ([161], Kap. 5.4). Für andere Verteilungen als die Normale existieren weitere Schätzungen, außerdem sind Verallgemeinerungen möglich (Partial CPC, Common Space Model). Das CPC-Modell ist gut bei großem p und mehreren Klassen, jedoch nicht skaleninvariant, so daß die Reihenfolge der Eigenvektoren wichtig ist.

SIMCA & DASCO

Eine weitere Regularisierung sind das SIMCA[5]- und das DASCO[6]-Modell. Hierbei werden die Variablen in jeder Klasse durch ein eigenes Hauptkomponenten-Modell repräsentiert. Eine Klassifizierung erfolgt dann aufgrund der relativen Distanz zu den Klassenmodellen. Es seien $\lambda_{i1}, \ldots, \lambda_{ip}$ die Eigenwerte und $\psi_{i1}, \ldots, \psi_{ip}$ die auf Länge Eins normierten Eigenvektoren von S_i. Über die kreuzvalidierte Gesamtfehlerrate werden davon die ersten p_i Vektoren im Modell gehalten, so daß sich die SIMCA-Regel als Regel des kleinsten Abstandes ergibt:

$$\delta_{SIMCA}(x) = \arg\min_i \frac{d_i^2(x)}{\frac{(n_i-1)\sum_{l=p_i+1}^{p}\lambda_{il}}{(n_i-p_i-1)}}\,, \tag{3.21}$$

wobei $d_i^2(x) = \sum_{l=p_i+1}^{p}\left[(x - \bar{x}_i)'\psi_{il}\right]^2$ die Summe der Quadrate der weggelassenen Eigenvektoren darstellt. Die S_i^{-1} werden so als $\sum_{l=1}^{p}\frac{\psi_{il}\psi_{il}'}{\lambda_{il}}$ geschätzt, wobei für $l \leq p_i$ die $\hat{\lambda}_{il}$ auf ∞ gesetzt und sonst als $\frac{n_i-1}{n_i-p_i-1}\sum_{l=p_i+1}^{p}\lambda_{il}$ geschätzt werden ([161], Kap. 5.4). Probleme sind die Ignoranz gegenüber Klassendifferenzen in den primären Unterräumen, die von den ersten p_i Eigenvektoren aufgespannt werden. Eine Verbesserung liefert der DASCO-Algorithmus. Hierbei werden die letzten $p-p_i$ Eigenwerte gemittelt; außerdem fließen die ersten p Eigenwerte ebenfalls mit ein, so daß die DASCO-Schätzungen der S_i^{-1}, die dann in (3.14) eingesetzt werden, folgende Form haben:

$$S_i^{-1} := \sum_{l=1}^{p_i}\frac{\psi_{il}\psi_{il}'}{\lambda_{il}} + \sum_{l=p_i+1}^{p}\frac{\psi_{il}\psi_{il}'}{\bar{\lambda}_i}, \tag{3.22}$$

[4] KRONECKER-Symbol: $\delta_{x,y} = 1 \iff x = y$, sonst 0

[5] Soft Independent Modeling of Class Analogy

[6] Discriminant Analysis with Shrunken Covariances

wobei $\bar{\lambda}_i := \frac{1}{p-p_i}\sum_{l=p_i+1}^{p}\lambda_{il}$.

Regularisierte Diskriminanzanalyse

Die *Regularisierte Diskriminanzanalyse* (RDA) als wichtige Form der Regularisierung entstammt einer Arbeit von FRIEDMAN [70]. Durch zwei verschiedene Kontrollparameter kann die Bias-Varianz-Abwägung in zwei Richtungen betrieben werden, indem die „Schrumpfung" der Schätzungen der Σ_i beeinflußt wird: $\lambda \in [0,1]$ kontrolliert, inwieweit die $\hat{\Sigma}_i$ der gepoolten Stichprobenvarianz S_{pooled} angeglichen werden (Kompromiß mit der LDA), $\gamma \in [0,1]$ den Grad der Angleichung an ein Vielfaches der Einheitsmatrix (Ridge-Ansatz). Die Schätzungen haben dann folgende Form ($i = 1, 2$):

$$\hat{\Sigma}_i(\lambda) = \frac{(1-\lambda)(n_i-1)S_i + \lambda(n-k)S_{pooled}}{(1-\lambda)(n_i-1)+\lambda(n-k)} \tag{3.23}$$

$$\hat{\Sigma}_i(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_i(\lambda) + \gamma c_i I_p \quad \left(c_i := \frac{1}{p}\mathrm{tr}\hat{\Sigma}_i(\lambda)\right). \tag{3.24}$$

Dadurch bietet sich eine reiche Klasse an Regularisierungsgraden. In dem Parameter-Quadrat $[0,1] \times [0,1]$ ist eine Vielzahl von Modellen denkbar. Über eine Schätzung von λ und γ aus der Stichprobe mittels der kreuzvalidierten Gesamtfehlerrate erhält man darüber hinaus ein adaptives Verfahren; die Daten „bestimmen" selbst, wieviel Glättung nötig ist. Spezialfälle sind:

	$\lambda = 0$	$\lambda = 1$
$\gamma = 0$	Klassische QDA	Klassische LDA
$\gamma = 1$	LDA mit $S_i^{-1} := c_i I_p$ in (3.8)	LDA mit $S_{pooled}^{-1} := I_p$ in (3.8)

Durch die Parametrisierung γ ist die RDA nicht skaleninvariant.

Als Schätzung empfiehlt FRIEDMAN [70] die Minimierung der kreuzvalidierten Gesamtfehlerrate

$$A^{(CV)}(\lambda,\gamma) = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n} z_{ij} I_{z_{ij} \neq \hat{\delta}^{-x_j}_{RDA(\lambda,\gamma)}(x_j)}. \tag{3.25}$$

Dieses zweidimensionale Optimierungsproblem löst er durch ein Netz von Gitterpunkten auf $[0,1] \times [0,1]$ mit 25-50 Punkten pro Dimension, für die jeweils $A^{(CV)}$ berechnet wird. Durch iterative Formeln kann der Rechenaufwand verringert werden (siehe [70]). Für den häufigen Fall, daß das Minimum über größere Bereiche erreicht wird, schlägt er die Wahl des maximalen λ und dann des maximalen γ vor. In seinen Simulationen erhält er durch dieses Vorgehen stets Schätzungen der beiden Parameter, die immer nahe dem Optimum sind.

Die RDA ist sehr viel besser als die LDA oder die QDA, wenn $n < p$, p groß im Vergleich zu n oder die Σ_i sich ähnlich sind, also in unvorteilhaften Situationen für die klassischen Verfahren. Sie ist aber auch kaum schlechter als die beiden in Situationen, die jeweils optimal für jene sind (p-variate Normalverteilung). Die kreuzvalidierte Fehlerrate unterschätzt die des Gesamtverfahrens aufgrund eines negativen Bias analog zur Apparent Error Rate (2.23) um ca. 20 % ([70]); da aber nur geringe Korrelationen zwischen beiden bestehen, ist sie als Optimalitätskriterium geeignet. Vor allem für hohe Dimensionen ($p \geq 20$) erfolgt so eine Glättung durch die RDA, womit man dem „Fluch der Dimensionalität" (Kapitel 2.5.1) entgeht. Die adaptive Glättung unterscheidet die RDA auch von den anderen Regularisierungen, bei denen durch die *Wahl* des Verfahrens der Grad der Glättung bestimmt wird.

Weitere Ansätze

Falls die Mittelwerte μ_1 und μ_2 nah beieinander liegen oder gleich sind oder die Klassenverteilungen schwere Enden besitzen, schlägt LACHENBRUCH [145] die sogenannte *Absolut-Lineare Diskriminanzanalyse* vor, allerdings nur für den Fall $p = 1$. Von sämtlichen Werten wird der Absolutbetrag genommen und darauf eine Lineare Diskriminanzregel errichtet. Falls sich die Klassenvarianzen nicht zu sehr unterscheiden (Faktor ≤ 5), liefert sie in seinen Simulationen ([145]) gute Ergebnisse. Bei schweren Enden wird sie besser als die QDA.

MCLACHLAN ([161], Kap. 5.3) trägt mehrere Ansätze robuster Schätzungen zusammen. Eine Form sind die *M-Schätzungen*, die über modifizierte Schätzungen der Σ_i Beobachtungen aus dem „Hauptkörper" der Daten (d.h. näher zum Erwartungswert) mehr Gewicht zuordnen und so vermutlich „atypische" Daten geringer zu gewichten, um damit die Regel gegenüber Ausreißern in der Stichprobe zu robustifizieren.

3.4 Die Logistische Diskriminanzanalyse

3.4.1 Das Verfahren

Während Lineare und Quadratische Diskriminanzanalyse (Kapitel 3.2 und 3.3) noch bestimmte Formen für die einzelnen Klassendichten f_i voraussetzen, macht die *Logistische Diskriminanzanalyse* (LGA) nur noch Annahmen über den Quotienten $\frac{f_1(x)}{f_2(x)}$. Das ist sinnvoll, weil für die Diskriminanzentscheidung nicht die absolute Größe der Dichten ausschlaggebend ist, sondern ihr Verhältnis zueinander. Die LGA geht auf ANDERSON [6] u.a. zurück.

Die fundamentale Annahme der LGA ist die Linearität des natürlichen Logarithmus des Dichtequotienten, so daß er in folgender Form darstellbar ist:

$$\log\left(\frac{f_1(x)}{f_2(x)}\right) = \beta_0 + \beta' x \quad (\beta' = (\beta_1, \ldots, \beta_p)) \,. \tag{3.26}$$

Kommentar 3.1 *Diese Linearitätsbedingung wird von einer Vielzahl von Verteilungen erfüllt. Beispiele sind:*

- *Multivariate Normalverteilungen mit gleichen Kovarianzmatrizen*
- *Verteilungen unabhängiger binärer Variablen*
- *Multivariate diskrete Verteilungen nach dem log-linearen Modell mit gleichen Interaktionstermen zweiter und höherer Ordnung*
- *Gemeinsame Verteilungen metrischer und diskreter Variablen mit den obigen Verteilungen*
- *Bestimmte abgeschnitte (truncated) Versionen der obigen Verteilungen*
- *Verteilungen der Exponentialfamilie*

(siehe [6], [9], [198] Kap. 6.4.1).

Die LGA kann so als ein Weg der Verallgemeinerung der LDA gesehen werden.

Aus der Linearitätsbedingung (3.26) folgt über das BAYESsche Theorem (2.12) sofort die Form des A-posteriori-Dichte-Quotienten:

$$\frac{\tau_1(x)}{\tau_2(x)} = \frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} = \exp\left(\log\frac{\pi_1}{\pi_2} + \beta_0 + \beta' x\right). \tag{3.27}$$

Mit dieser Bedingung und $\tau_2(x) = 1 - \tau_1(x)$ bei $k = 2$ ist leicht

$$\tau_1(x) = \frac{\exp\left(\log\frac{\pi_1}{\pi_2} + \beta_0 + \beta' x\right)}{1 + \exp\left(\log\frac{\pi_1}{\pi_2} + \beta_0 + \beta' x\right)} \tag{3.28}$$

$$\tau_2(x) = \frac{1}{1 + \exp\left(\log\frac{\pi_1}{\pi_2} + \beta_0 + \beta' x\right)} \tag{3.29}$$

zu erhalten. Diese Form war namensgebend für die LGA. Die logistische Diskriminanzregel lautet so:

$$\hat{\delta}_{LGA}(x) = \begin{cases} 1: & \log\frac{\pi_1}{\pi_2} + \beta_0 + \beta' x \geq \log\frac{c_{21}}{c_{12}} \\ 2: & \text{sonst} \end{cases} \tag{3.30}$$

Die Bestimmung der Parameter erfolgt über den Maximum-Likelihood-Ansatz. Für die unterschiedlichen Sampling-Schemata muß jeweils die gleiche Likelihood-Funktion

$$\mathcal{L}(x; \beta_0^*, \beta) = \prod_x [\tau_1(x)]^{n_1(x)} [\tau_2(x)]^{n_2(x)}$$

maximiert werden, wobei $\beta_0^* = \beta_0 + \log\frac{\pi_1}{\pi_2}$ gilt und $n_i(x)$ die Anzahl derjenigen $y \in \mathcal{S}$ ist, die aus Π_i stammen und für die $x = y$ gilt. Die Optimierung erfolgt über Quasi-Newton-Methoden ([9]). Die 1. und 2. Ableitungen sind mit

$$\begin{aligned} \frac{\partial \log \mathcal{L}(x; \beta_0^*, \beta)}{\partial \beta_j} &= \sum_x [n_1(x) - n(x)\tau_1(x)]\, x_j \quad (j = 0, 1, \ldots, p) \\ \frac{\partial^2 \log \mathcal{L}(x; \beta_0^*, \beta)}{\partial \beta_j \partial \beta_l} &= -\sum_x n(x)\tau_1(x)\tau_2(x) x_j x_l \quad (j, l = 0, 1, \ldots, p) \end{aligned}$$

gegeben. Anderson [6] schlägt als vertrauenswürdigen Startwert Null für alle β_j vor.

Unter homoskedastischer Normalität erhält man ([54], Kap. 10.6.2):

$$\beta' = \Sigma^{-1}(\mu_1 - \mu_2) \tag{3.31}$$

$$\beta_0 = \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) + \log\frac{\pi_1}{\pi_2}. \tag{3.32}$$

3.4.2 Eigenschaften

Die Hauptvorteile der LGA sind die relativ geringen Verteilungsannahmen, die Anwendbarkeit auf stetige, diskrete und gemischte Daten und die einfache Benutzung, da nach der Schätzung nur noch eine lineare Funktion berechnet werden muß. Das bedingt auch eine lineare Trennfläche, die einfach zu interpretieren ist. Weiterhin müssen nur $p + 1$ Parameter geschätzt werden. Für beide Datentypen wird die gleiche iterative ML-Schätzung benutzt. Da nur der Quotient modelliert wird, sind viele verschiedene Klassenverteilungen möglich. Durch die teilweisen Verteilungsannahmen können Tests zur Variablenselektion ([9]) und Erweiterungen wie Updating abgeleitet werden. Bei Benutzung dieser Tests ist jedoch Vorsicht geboten: Bei Abweichungen von den Modellannahmen können diese Tests in die Irre führen (siehe Kapitel 2.5.1). Kleine Stichproben-Umfänge können durch starken Bias problematisch für die LGA sein.

Für zwei Fälle hat die Likelihood-Funktion $\mathcal{L}$ kein eindeutiges Maximum. Die Trainingsstichprobe kann erstens linear vollständig separierbar[7] sein. ANDERSON [9] sieht aufgrund der Separiertheit der Klassen darin kein Problem; jede separierende Trennfläche wird eine gute Diskriminanzregel sein. Als zweiter Fall kann bei diskreten Daten das Problem auftreten, daß eine Klassendichte in allen Zellen einer Variable gleich Null in der Trainingsstichprobe ist, für die andere Dichte dagegen nicht. Bei einer neuen Beobachtung, die in eine dieser Zellen fällt, würde das sofort eine Zuordnung zur letzteren Klasse mit Wahrscheinlichkeit Eins erzwingen ([9]). Um die Variable trotzdem als sicherlich wichtigen Prädiktor zu erhalten, diskutiert ANDERSON [7] einige Lösungsansätze.

Laut ANDERSON [6] ist die Linearitätsbedingung (3.26) für diskrete Variablen mit mehreren Ausprägungen schwer zu erfüllen. Er empfiehlt deren Umwandlung in binäre Variablen. Durch die Verwendbarkeit von gemischten Daten hat die LGA weite Anwendung in den Sozialwissenschaften gefunden.

Unter homoskedastischer Normalität und kleinem MAHALANOBIS-Abstand (d.h. $\Delta^2 \le 1.5$) ist die LGA annähernd so effizient wie die LDA (Kapitel 3.2), sonst (d.h. $\Delta^2 \ge 2.5$) fällt sie ab ([161]). Bei ungleichen Kovarianzmatrizen wird die LGA leicht besser als die LDA. Bei klarer Nichtnormalität oder Heteroskedastizität ist die LGA der LDA eindeutig vorzuziehen, obwohl sie eine ähnliche diskriminatorische Kraft hat ([198] Kap. 6.4.4, [90]). PRESS & WILSON [173] erhalten in zwei Studien unter Nichtnormalität geringere AERs (2.23) der LGA gegenüber der LDA, aber nicht grundsätzlich andere Resultate. KRZANOWSKI [138] erhält für Mischungen von binären und stetigen Merkmalen meist identische Resultate.

3.4.3 Erweiterungen*

AITCHISON & DUNSMORE [2] benutzen statt der Likelihood-Schätzung den Prädiktiven Ansatz (2.17). Sie schätzen die Aposteriori-Dichte (3.28) als

$$\hat{\tau}_1^{(P)}(x) = \int \tau_1(x;\tilde{\beta})\phi_{\hat{\beta},\hat{V}}(\tilde{\beta})\,d\tilde{\beta} \tag{3.33}$$

mit $\tilde{\beta} = (\beta_0, \beta)'$, der ML-Schätzung $\hat{\beta}$ von $\tilde{\beta}$ mit der Varianz $\hat{V}$ und ϕ als Dichte der $(p+1)$-variaten Normalverteilung.

ANDERSON [8] modelliert den Logarithmus des Quotienten in (3.26) als quadratische Funktion,

$$\log\left(\frac{f_1(x)}{f_2(x)}\right) = \beta_0 + \beta' x + x'\Gamma x \quad (\Gamma \text{ symmetrisch}), \tag{3.34}$$

wie er beispielsweise bei normalverteilten Klassen mit verschiedenen Kovarianzmatrizen auftritt. Die Trennfläche wird so quadratisch; dieser Ansatz wird darum als *Quadratisch-Logistische Diskriminanzanalyse* bezeichnet. Modelliert man die Σ_i vollständig (d.h. $\Gamma = \Sigma_1^{-1} - \Sigma_2^{-1}$), so steigt die Anzahl der Parameter jedoch sofort von $p+1$ auf $(k-1)(p+1)(1+\frac{p}{2})$ an. Deswegen schlägt er verschiedene Approximationen vor, etwa über die größten Eigenwerte von Γ:

[7] d.h. alle Objekte aus Π_1 liegen auf der einen Seite einer Hyperebene im Merkmalsraum und alle Objekte aus Π_2 auf der anderen

$$\hat{\Gamma} := \sum_{j=1}^{\check{p}} \lambda_j l_j l_j' \quad (\check{p} \leq p),$$

wobei die λ_j die Eigenwerte mit den zugehörigen normierten Eigenvektoren l_j darstellen. Für $\check{p} = 1$ erhöht sich dadurch die Parameteranzahl auf $2p + 1$. Für $p \leq 20$ und $k = 2$ funktioniert dieser Ansatz gut, für binäre Daten mit log-linearer Verteilung jedoch nicht ([8]).

Eine weitere Verallgemeinerung ist die Annahme, daß (3.26) lokal als Summe glatter Funktionen geschätzt wird:

$$\log\left(\frac{f_1(x)}{f_2(x)}\right) = \sum_{j=1}^{p} s_j\left(x_{(j)}\right), \tag{3.35}$$

Man benutzt dabei die Erweiterung des Linearen Modells zum Verallgemeinerten Additiven Modell ([115]), um über *Scatterplot Smoother*, den *MARS*-Algorithmus von Friedman [71] oder andere flexible nichtparametrische Regressionsverfahren den Quotienten zu schätzen, so daß die Trennfläche nicht mehr linear wird (siehe Kapitel 5.2).

Cox & Ferry [43] robustifizieren die LGA gegen Ausreißer.

3.5 Nichtnormale Modelle

In Spezialgebieten, in denen die Klassenverteilungen genau bekannt sind oder deren Eigenschaften eher von der Normalverteilung wegführen, kann es Sinn machen, spezielle Verteilungen direkt anzusetzen, deren Parameter zu bestimmen und diese Schätzungen in die Bayessche Regel (2.6) einzusetzen. Einige Verteilungen sollen im folgenden kurz vorgestellt werden. Für Referenzen siehe [161], Kap. 7.7.

- Die Familie der *θ-generalisierten Normalverteilungen* [8] hat folgende allgemeine Gestalt (univariat bzw. multivariat):

$$\begin{aligned} f(x;\theta,\mu,a) &= \frac{1}{2a\Gamma(1+\frac{1}{\theta})} \mathrm{e}^{-\left|\frac{x-\mu}{a}\right|^\theta} \\ f(x;\theta,\mu,A) &= \left[(2\Gamma(1+\frac{1}{\theta}))^p|A|\right]^{-1} \exp\left\{-\sum_{j=1}^{p}\left|\left(A^{-1}\right)_j(x-\mu)\right|^\theta\right\}. \end{aligned}$$

  Dabei ist A eine nichtsinguläre Matrix und $(A)_j$ die j-te Spalte von A. Der Familienparameter $\theta > 0$ liefert für $\theta = 2$ die Normalverteilung mit $\Sigma = \Gamma(3/\theta)\Gamma(1/\theta)AA'$.

- Die *Exponentialverteilung* mit der Dichte

$$f(x;\mu) = \frac{1}{\mu}\mathrm{e}^{-\frac{x}{\mu}} I_{(0,\infty)}(x)$$

  wird oft für Lebensdauer- und Zuverlässigkeitsanalysen verwendet. Die optimale Diskriminanzregel ist hierbei linear in x. Über den Mittelwert $\bar{x}_i$ für die μ_i der Klassen kann so leicht eine Plug-In-Schätzung der Regel formuliert werden.

[8] auch: *r-normierte Exponentialverteilungen*

- Für die Beschreibung von BROWNschen Bewegungen (z.B. bei Aktienkurs-Modellierungen) oder Anwendungen in der Elektronik und Zuverlässigkeitstheorie werden oft *Invers-Normale Modelle* benutzt. Im Univariaten hat deren Dichte die Form

$$f(x;\theta) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \mathrm{e}^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}} I_{(0,\infty)}(x)$$

  mit $\theta = (\mu, \lambda)'$ und $\lambda = \frac{\mu^3}{\sigma^2}$ ($\mu > 0$). Für große λ nähert sich die Verteilung der Normalverteilung an, für kleine λ jedoch wird ihre positive Schiefe immer stärker. Sie repräsentiert gut positiv schiefe Verteilungen mit langen Enden (im Gegensatz zur Normalverteilung, die symmetrisch ist und recht schnell abfällt). Plug-In-Schätzungen (ML-Schätzungen) für $k = 2$ sind über

$$\begin{aligned} \hat{\mu}_i &= \bar{x}_i \\ \hat{\lambda}_i &= \left[\sum_{j=1}^{n} z_{ij} \frac{x_j^{-1} - \bar{x}_i^{-1}}{n_i}\right]^{-1} \end{aligned}$$

  gegeben.

- Ebenfalls zur Modellierung längerer Enden als denen der Normalverteilung wird die multivariate *t-Verteilung* mit m Freiheitsgraden benutzt, da bei sogenannten schweren Enden dieser Ansatz besser als die LDA (Kapitel 3.2) wird. Für $m \to \infty$ nähert sie sich der Normalverteilung an.

- MORRIS & LAYCOCK [166] haben für Richtungsdaten[9] (*directional data*) über die VON-MISES- und die FISHER-Verteilungen BAYESsche Diskriminanzregeln abgeleitet.

Weitere Modellierungen sind *Log-lineare Modelle* (Modellierung des Logarithmus der Klassendichte als lineare Funktion), LANCASTER-*Modelle* (Erweiterung des log-linearen Modells um Interaktionsterme 1. und 2. Ordnung) und *Latent-class-Modelle* (Modellierung der Klassendichten als unbekannte Mischung verschiedener Dichten), die hier nicht weiter beschrieben sein sollen (siehe [161], Kap. 7.3).

[9] Während „normale“ Daten durch Richtung und Abstand relativ zum Koordinatenursprung definiert sind, ist bei Richtungsdaten der Abstand uninteressant. Die Objekte sind so auf der p-dimensionalen Einheitskugel S_p repräsentiert; ihr p-dimensionaler Wertevektor I ist auf 1 normiert: $I'I = 1$. Die Dichten werden meist in sphärischen Polarkoordinaten ausgedrückt, also z.B. für p=2: $i_1 = \cos\alpha$, $i_2 = \sin\alpha$ ($\alpha \in [0, 2\pi)$). Für Details siehe [156], Kap. 15.

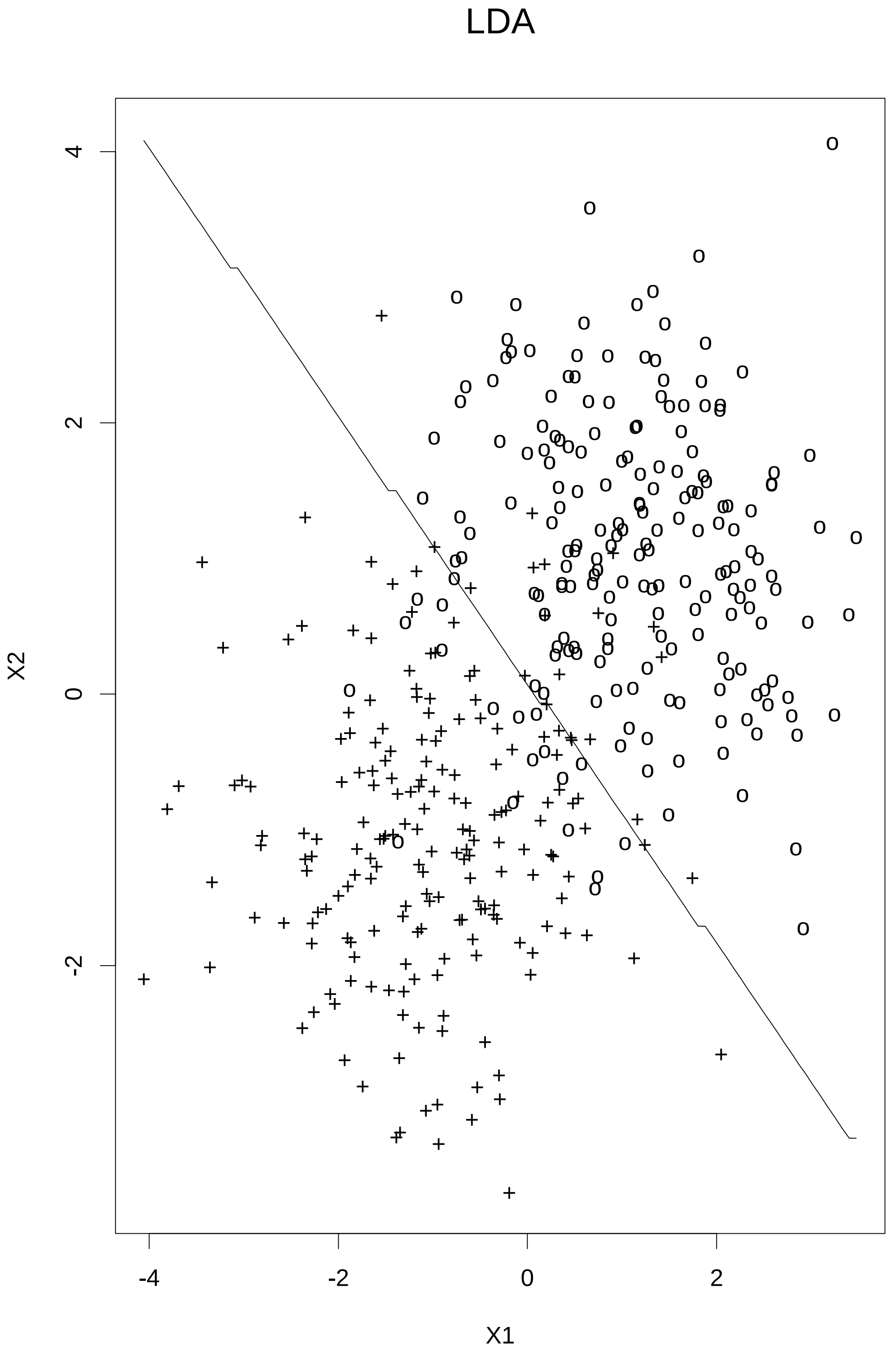

Abbildung 3.1: Beispiel einer linearen Trennfläche mittels der Linearen Diskriminanzanalyse. Die Populationen wurden aus $N((1,1)', I_2)$ bzw. $N((-1,-1)', I_2)$ generiert, bei $n_1 = n_2 = 100$ (siehe Kapitel 7.1, Beispiel 6). Objekte aus der Population Π_1 sind durch „+“ dargestellt, solche aus Π_2 durch „o“. Die über Cross-validation mit 10 Gruppen ermittelte Fehlklassifikationswahrscheinlichkeit beträgt 9,8 %. Die „Brüche“ in der Geraden sind von technischer Art bei der Darstellung.

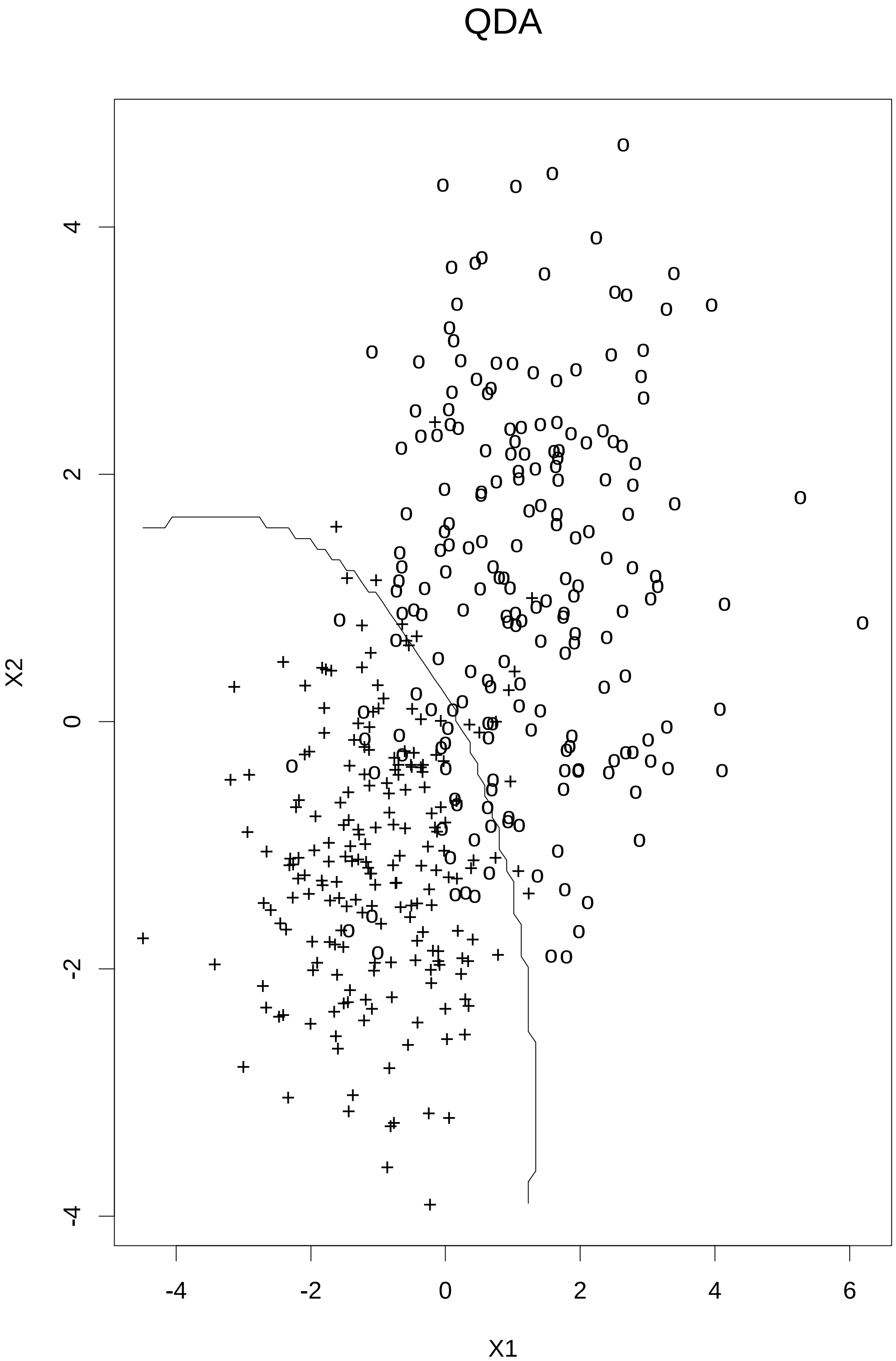

Abbildung 3.2: Beispiel einer quadratischen Trennfläche mittels der Quadratischen Diskriminanzanalyse. Die Populationen wurden aus den Verteilungen $N((1,1)', I_2)$ bzw. $N((-1,-1)', \mathrm{diag}(2,2))$ generiert, bei $n_1 = n_2 = 100$ (siehe Kapitel 7.1, Beispiel 7). Objekte aus der Population Π_1 sind durch „+" dargestellt, solche aus Π_2 durch „o". Die über Cross-validation mit 10 Gruppen ermittelte Fehlklassifikationswahrscheinlichkeit beträgt 10,8 %. Die „Brüche" in der Kurve sind von technischer Art bei der Darstellung.

Kapitel 4

Nichtparametrische Verfahren

Nichtparametrische Verfahren verzichten auf globale Annahmen über die Form von Dichten oder der Trennfläche. Deswegen werden sie oft als verteilungsfreie Methoden bezeichnet. Sie ersetzen jedoch die globalen Annahmen durch andere, z.B. die lokale Glattheit von Dichten oder A-posteriori-Verteilungen (d.h. diese ändern sich in kleinen Umgebungen nur geringfügig) oder das Vorhandensein eines funktionalen Zusammenhanges zwischen Merkmalen und Klassenzugehörigkeit. So finden bei nichtparametrischen Verfahren ebenfalls — lokal oder global — Verteilungs- oder Parameterschätzungen statt.

4.1 Modelle über Kontingenztafeln

4.1.1 Das Multinomial- und das Unabhängigkeitsmodell

Das einfachste Modell für diskrete Daten ist das *Multinomialmodell* (MNA). Die p diskreten Merkmale $X_1, \ldots, X_p$ sollen $a_1, \ldots, a_p$ Ausprägungen besitzen. Durch Kombination der Kategorien der einzelnen Merkmale miteinander erhält man eine p-dimensionale Tabelle (Kontingenztafel) mit $A := \prod_{i=1}^{p} a_i$ disjunkten Zellen.

Eine natürliche Schätzung (ML-Plug-in) für die Klassendichten f_i in den einzelnen Zellen ist sofort mit

$$\hat{f}_i^{(MNA)}(x)\Big|_{x \in c} := \hat{q}_{ic} := \frac{n_{ic}}{n_i} \quad (i = 1, 2) \tag{4.1}$$

gegeben, wobei $c \in \{1, \ldots, A\}$ den Zellenindex[1] angibt und n_{ic} die Anzahl der $x_j \in \mathcal{S}$, die aus der Population Π_i stammen und in der Zelle c liegen ([161], Kap. 9.2). Für die Dichteschätzung an einem Punkt kommen so nur Objekte der Stichprobe in Betracht, die in allen Merkmalen mit der Ausprägung des zu klassifizierenden Objekts übereinstimmen. Die daraus gemäß der BAYESschen Regel (2.10) entstehende Regel,

$$\hat{\delta}_{MNA}(x) = \begin{cases} 1: & \frac{\pi_1 \hat{f}_1^{(MNA)}(x)}{\pi_2 \hat{f}_2^{(MNA)}(x)} \geq \frac{c_{21}}{c_{12}} \\ 2: & \text{sonst} \end{cases}, \tag{4.2}$$

ist BAYES-konsistent und konvergiert mindestens exponentiell schnell mit wachsendem n gegen die BAYESsche Regel ([86], [87]).

[1] Ein einfaches Beispiel für einen Zellenindex etwa bei einer 2×5-Kontingenztafel (d.h. $A = 10$) ist: $x_c \equiv x_{s,t} \iff c = 5(s-1) + t$.

Das Problem dieses Ansatzes ist die Dimensionalität. Bei zwei Merkmalen mit jeweils 5 Ausprägungen ergeben sich 25 Zellen. Mit 250 gleichmäßig auf die Zellen verteilten Beobachtungen in $\mathcal{S}$ aus einer Klasse stehen so pro Zelle ca. 10 für eine Schätzung zur Verfügung. Bei drei zusätzlichen Merkmalen mit ebenfalls 5 Ausprägungen benötigt man für dieselbe Datendichte bereits 31.250 Beobachtungen. Infolgedessen werden Dichteschätzungen in vielen Zellen gleich Null sein (für beide Klassen) und eine Zuordnung unmöglich machen. Dadurch, daß nur Beobachtungen der Stichprobe betrachtet werden, die in allen Merkmalen der zu schätzenden Zelle entsprechen, wird der Bias der Schätzung gering gehalten. Für wachsendes p steigt aber die Varianz, also die Abhängigkeit von den zufällig in der Stichprobe vorhandenen Beobachtungen, und damit das Risiko der Regel (bei festem n) stark an — eine Form des „Fluches der Dimensionalität“ (siehe Kapitel 2.5.1). Die MNA ist so nur für kleine Dimensionen A von Merkmalen anwendbar.

Unabhängigkeitsmodell

Eine starke Form der Glättung des MNA ist das *Unabhängigkeitsmodell* (UMA). Die Variablen werden als unabhängig angenommen, so daß die Dichten in den einzelnen Zellen das Produkt der Randdichten sind. Man hofft, durch diese Regularisierung die Varianz auf Kosten eines erhöhten Bias zu verringern, um insgesamt ein geringeres Risiko zu erzielen. Die ML-Schätzung hat folgendes Aussehen ([161], Kap. 7.2):

$$\hat{f}_i^{(UMA)}(x) := \prod_{j=1}^{p} \frac{n_{ij}}{n_i} \quad (i = 1, 2), \tag{4.3}$$

mit n_{ij} als der Anzahl der $x \in \mathcal{S}$, die aus der Population Π_i stammen und im j-ten Merkmal mit der Ausprägung von x übereinstimmen. Somit sind alle Objekte aus der Stichprobe, die in mindestens einer Ausprägung mit dem zu klassifizierenden Objekt übereinstimmen, mit gleichem Gewicht an der Dichteschätzung beteiligt. Die Bayessche Regel (2.10) liefert dann:

$$\hat{\delta}_{UMA}(x) = \begin{cases} 1: & \frac{\pi_1 \hat{f}_1^{(UMA)}(x)}{\pi_2 \hat{f}_2^{(UMA)}(x)} \geq \frac{c_{21}}{c_{12}} \\ 2: & \text{sonst} \end{cases} . \tag{4.4}$$

Auch dieser Ansatz kann nicht das Verschiedensein der Schätzung von Null garantieren. Eine Ridge-Schätzung ([161]) vermeidet das:

$$\tilde{f}_i^{(UMA)}(x) := \prod_{j=1}^{p} \frac{n_{ij} + a_j^{-1}}{n_i + 1} \quad (i = 1, 2). \tag{4.5}$$

Die regularisierende Annahme der Unabhängigkeit der Variablen verringert die Anzahl der zu schätzenden Parameter erheblich; insbesondere steigt bei zunehmendem p diese nicht als Produkt der Anzahlen möglicher Ausprägungen der Variablen, sondern nur als deren Summe. Problem des Unabhängigkeitsmodells ist die möglicherweise zu starke Glättung (d.h. ein zu großer Bias), wenn z.B. starke Korrelationen zwischen Variablen vorliegen. Die *Nearest-Neighbours-* und *Kernschätzung*-Verfahren gehen denn auch von einer lokalen Glättung aus, indem sie nur Beobachtungen in der Stichprobe aus der Umgebung von x betrachten oder den Beobachtungen dort verschiedene Gewichte zuteilen (siehe Kapitel 4.3 und 4.2).

Weitere Ansätze

K. SCHULZE [194] betrachtet als Zwischenform das *Blockmodell*, das die Merkmale in disjunkte Blöcke unterteilt. Die Wahrscheinlichkeit eines Blockes wird durch die relativen Häufigkeiten nach dem MNA geschätzt; die Zellwahrscheinlichkeiten ergeben sich als Produkt dieser Blockwahrscheinlichkeiten. Das MNA und das UMA sind als Spezialfälle bei genau einem bzw. genau p Blöcken enthalten.

Hat man eine Vorstellung von den Zelldichten, etwa eine A-priori-Verteilung $\xi_i = (\xi_{i1}, \ldots, \xi_{iA})$, kann durch *Convex Smoothing*,

$$\tilde{q}_{ic} = (1-\omega_i)\hat{q}_{ic} + \omega_i \xi_{ic} \quad (\omega_i \in [0,1],\, i = 1,2), \tag{4.6}$$

d.h. der Abwägung zwischen der Information $\hat{q}_{ic}$ aus $\mathcal{S}$ und der A-priori-Information ξ_{ic}, eine Glättung des MNA erreicht werden ([161], Kap. 9.2). Für die ξ_{ic} ist aber auch jede beliebige Schätzung, etwa die Unabhängigkeitsschätzung, möglich. Die ω_i können z.B. als

$$\omega_i := \frac{C_i}{C_i + n_i \sum_{c=1}^{A} (\hat{q}_{ic} - \xi_{ic})^2}, \quad C_i := 1 - \sum_{c=1}^{A} \hat{q}_{ic}{}^2$$

oder mittels eines Glättungsparameters $h_i \in [0,1]$ als

$$\omega_i := \frac{A}{A-1}(1-h_i)$$

berechnet werden ([161]). Letztere Methode ergibt mit $\xi_i = (\frac{1}{A}, \ldots, \frac{1}{A})$ den Binomial-Kern von AITCHISON & AITKEN (siehe Kapitel 4.2). Für weitere (Pseudo-) BAYESsche Ansätze siehe z.B. auch BISHOP et al. [18], Kap. 12.

GRUND ([93], [94]) und GRUND & HALL ([95]) untersuchen die Annahme loglinearer Modelle und die Glättung mittels Kernschätzungen für die Schätzung der Zellwahrscheinlichkeiten (d.h. Möglichkeiten zur Varianzreduktion). Die letztere Arbeit betrachtet dabei hochdimensionale binäre Kontingenztafeln mit nur spärlichen Daten. Eine weitere Möglichkeit der Glättung bei binären Daten sind *Orthogonale Reihen* ([169], [161] Kap. 9.2).

O. BUNKE [32] schlägt eine adaptive Glättung der Zellwahrscheinlichkeiten vor. Dabei werden nicht die Zellwahrscheinlichkeiten direkt geglättet, sondern der Quotient der Schätzungen nach dem Multinomial- und nach dem Unabhängigkeitsmodell für eine Zelle durch die Quotienten benachbarter Zellen über das Nearest-Neighbours-Verfahren (Kapitel 4.3); siehe [32] für Einzelheiten.

4.1.2 Das Lokationsmodell

Für gemischt diskrete und stetige Daten wird oft das sogenannte *Lokationsmodell* (LOM) verwendet. Es seien die Variablen so geordnet, daß die ersten p_1 Variablen diskret und die nachfolgenden $p_2 := p - p_1$ stetig sind, so daß sich der Merkmalsvektor eines Objektes x in $\left(x^{(p_1)}, x^{(p_2)}\right)$ aufteilt. Man erhält auf diese Weise wieder eine Kontingenztafel in den p_1 Merkmalen, in deren Zellen die p_2 stetigen Merkmale getrennt modelliert werden.

Als einziges Beispiel in dieser Arbeit sei das Modell von KRZANOWSKI [138] vorgestellt. Er betrachtet den Fall, daß die diskreten Daten binär und die stetigen Daten in jeder Zelle normalverteilt sind mit gleicher Kovarianz Σ in allen Zellen und beiden Klassen:

$$X^{(p_2)}|X^{(p_1)} \sim \mathrm{N}(\mu_{ic}, \Sigma)\,. \tag{4.7}$$

Schätzt man den diskreten Anteil über das MNA, so ergeben sich die ML-Schätzungen der Klassendichten als Produkt beider Anteile:

$$\hat{f}_i^{(LOM)}(x)\Big|_{x\in c} := \frac{n_{ic}}{n_i}\,\phi_{\bar{x}_{ic},S_{pooled}}(x^{(p_2)}) \quad (i = 1, 2). \tag{4.8}$$

Dabei bezeichnen die n_{ic} die Anzahlen der $x_j \in \mathcal{S} \cap \Pi_i$, die in ihren diskreten Merkmalen dieselben Ausprägungen wie x besitzen. Die $\bar{x}_{ic}$ sind deren p_2-dimensionale Mittelwertsvektoren in den stetigen Merkmalen, S_{pooled} ist die gepoolte Kovarianz der stetigen Merkmale in allen Zellen analog (3.4), und ϕ bezeichnet die Dichte der p_2-variaten Normalverteilung. Die Dichten werden erneut in die BAYESsche Regel (2.10) eingesetzt und liefern nach ein paar Umformungen für $x^{(p_1)} \in c$:

$$\hat{\delta}_{LOM}(x) = \begin{cases} 1: & (\bar{x}_{1c}-\bar{x}_{2c})'S_{pooled}^{-1}x^{(p_2)} - \frac{1}{2}(\bar{x}_{1c}-\bar{x}_{2c})'S_{pooled}^{-1}(\bar{x}_{1c}+\bar{x}_{2c}) \\ & \geq \log\left(\frac{\pi_2 c_{21}}{\pi_1 c_{12}}\right) + \log\left(\frac{n_{2c}/n_2}{n_{1c}/n_1}\right) \\ 2: & \text{sonst} \end{cases}\,. \tag{4.9}$$

Für diesen Ansatz müssen $2^{p_1} + k2^{p_1}p_2 + \frac{1}{2}p_2(p_2+1)$ Parameter geschätzt werden. Das sind z.B. bei $k = p_2 = 2$ für $p_1 = 2$ genau 23 und für $p_1 = 4$ schon 83 Parameter. So benötigt der LOM-Ansatz sehr große Stichproben-Umfänge, vor allem für wachsendes p_1, da für jede Zelle und Klasse ein Normales Modell mit eigenen Parametern geschätzt werden muß. Dem entgegen werden für große p die obigen Schätzungen schlecht im Sinne des Risikos ([161], Kap. 7.4). Das LOM ist daher vor allem für kleine p anwendbar. Bei Interaktionen zwischen den binären und den stetigen Daten[2] liefert das LOM bessere Ergebnisse als die LDA (siehe Kapitel 3.2) ([138]).

Verallgemeinerungen sind durch klassenspezifische Kovarianzen Σ_i oder gar klassen- und zellspezifische Σ_{ic} denkbar, die aber aufgrund der dramatisch ansteigenden Parameteranzahl kaum noch vernünftig schätzbar sind. Es existiert auch ein prädiktiver Zugang, der aber in Simulationen keine geringeren Fehlerraten lieferte ([161], Kap. 7.4). KRZANOWSKI [140] schlägt eine lineare Transformation der stetigen Daten mittels der binären vor, so daß die LDA anwendbar wird (siehe Kapitel 3.2). Auch eine Kernschätzung der zellspezifischen Σ_{ic} ist denkbar.

4.2 Kernschätzungen

4.2.1 Das Verfahren

Kerndichte- oder kurz *Kernschätzungen*[3] sind ein weitverbreitetes Mittel der nichtparametrischen Dichteschätzung, das sich über die *Averaged Shifted Histograms* als Verallgemeinerung des Histogramms herleiten läßt.[4] Die Diskriminanzanalyse mittels Kernschätzungen bemüht sich in den einzelnen Klassen um eine lokale Dichteschätzung am zu klassifizierenden Punkt. Da durch die Stichprobe nur punktförmig Information über die Klassendichten zur Verfügung steht, findet eine Glättung dieser

[2] Interaktionen bedeuten hier, daß sich die Mittelwerte μ_{ic} in den Zellen unterscheiden. Ohne diese Interaktionen werden sie über alle Zellen identisch sein ($\mu_{ic} \equiv \mu_i\,\forall c$), so daß eine einfache LDA in den p_2 stetigen Merkmalen anwendbar ist.

[3] in englisch: *Kernel density estimation*

[4] siehe z.B. [201], [196], [127], [105], [107]

Information im Umkreis des Punktes durch eine Kernfunktion statt. Die allgemeine multivariate Form für die Kerndichteschätzung ([196], Kap. 6.3.2) lautet:

$$\hat{f}_i^{(K)}(x) := \frac{1}{n_i}\frac{1}{|H_i|}\sum_{j=1}^{n_i} K_p\left(H_i^{-1}(x - x_{i_j})\right) \quad (i = 1, 2). \tag{4.10}$$

Dabei sind K_p eine p-variate, reellwertige *Kernfunktion* (oder *Kern*) mit $\int K_p(x)dx = 1$ und die $H_i \in \mathcal{M}_{p \times p}$ klassenspezifische nichtsinguläre Matrizen von Glättungsparametern (*Bandweitenmatrizen*). Diese dienen sowohl der Skalierung als auch der Definition einer Umgebung für die Glättung. Oft werden die Kerne nichtnegativ, unimodal und symmetrisch um Null gewählt, d.h. $K(x) = K(-x)$. Die Schätzung der Dichte am Punkt x ist damit das arithmetische Mittel der n_i Beobachtungen aus $\mathcal{S} \cap \Pi_i$, die über den Kern nach einem gewissen Abstandsmaß um x herum gewichtet werden. H_i vermittelt die „Lokalität“ der Schätzung in den einzelnen Merkmalen: Erstens wird der Grad der Glättung bestimmt, zweitens aber auch die Richtung der Glättung im Merkmalsraum.

Die so geschätzten Dichten f_i werden in die BAYESsche Regel (2.10) eingesetzt (Plug-in) und liefern die Diskriminanzregel:

$$\hat{\delta}_{Kern}(x) = \begin{cases} 1: & \frac{\pi_1 \hat{f}_1^{(K)}}{\pi_2 \hat{f}_2^{(K)}} \geq \frac{c_{21}}{c_{12}} \\ 2: & \text{sonst} \end{cases} . \tag{4.11}$$

EPANECHNIKOV [60] u.a. empfehlen aus asymptotischen Gründen die Verwendung von Produktkernen, so daß sich der multivariate Kern K_p als Produkt univariater Kerne K_1 für jede Variable ergibt:

$$K_p^{(P)}(x) := \prod_{l=1}^{p} \frac{1}{h_l} K_1\left(\frac{x_{(l)}}{h_l}\right) \quad (h_l \geq 0) \tag{4.12}$$

($x = (x_{(1)}, \ldots, x_{(p)})'$, $H = \operatorname{diag}(h_1, \ldots, h_p)$). Das impliziert die Unabhängigkeit bzw. Unkorreliertheit der Variablen und, damit verbunden, eine starke Verringerung der (Verfahrens-) Parameteranzahl in H_i. Die Struktur von H_i wird eher gesetzt als geschätzt, dabei oft diagonal aufgrund einer besseren Rechenbarkeit. Wichtiger als die Struktur von H_i ist deren *Skalierung* (cH_i), da dies — bei beschränktem Support — über die Umgebung $\mathcal{U}(x)$, in der eine Mittelung erfolgt ($\mathrm{E}_{\mathcal{U}(x)}\, f(x)$), entscheidet oder das Gewicht der einzelnen Punkte bestimmt. Die Parameteranzahl in H_i ist nicht wichtig für eine Reduktion der Varianz. Die Skalierung moderiert die Bias-Varianz-Abwägung (siehe Kapitel 2.3), indem größere Umgebungen zu einer Verminderung der Varianz bei möglicher Biaserhöhung führen. Durch Verwendung eines einzigen, globalen Glättungsparameters h_i pro Klasse ($H_i = h_i\, I_p$) erreicht man die häufigste Form der Kernschätzung (*Global window width*) mit der Gestalt

$$\hat{f}_i^{(K)}(x) := \frac{1}{n_i}\frac{1}{h_i^p}\sum_{j=1}^{n_i}\prod_{l=1}^{p} K_1\left(\frac{x_{(l)} - x_{i_j(l)}}{h_i}\right) \quad (i = 1, 2) \tag{4.13}$$

besitzt. Große h_i bewirken, daß auch weiter von x entfernte Beobachtungen aus $\mathcal{S} \cap \Pi_i$ sich auf die Schätzung der Dichte auswirken, also eine stärkere Glättung erfolgt mit dem Ergebnis verminderter Varianz. Bei Verwendung nur eines h_i pro Klasse müssen die Merkmale im allgemeinen vorher auf ähnliche Größenordnungen skaliert werden. Eine einzige Bandweite für beide Klassen bewirkt dieselbe Variabilität für beide Dichteschätzungen, falls $n_1 = n_2$ gilt.

Kern	$K_1(u)$	*Bemerkung*
Multinomial	$I_{\{u=0\}}$	keine Glättung
Uniform (Gleichvert.)	$\frac{1}{2}I_{\{\|u\|\leq 1\}}$	
Dreieck (Triangle)	$(1-\|u\|)I_{\{\|u\|\leq 1\}}$	
Binomial [3]	$\begin{cases} \lambda & : \quad u=1 \\ 1-\lambda : & \quad u=0 \end{cases}$	$\lambda \in [\frac{1}{2}, 1]$, $x_i \in \mathcal{S}$
Epanechnikov [60]	$\frac{3}{4}(1-u^2)I_{\{\|u\|\leq 1\}}$	
Biweight (Quartic)	$\frac{15}{16}(1-u^2)^2 I_{\{\|u\|\leq 1\}}$	
Triweight	$\frac{35}{32}(1-u^2)^3 I_{\{\|u\|\leq 1\}}$	
Normal (Gauss)	$\phi_{0,1}(u)$	infiniter Support
Double Exponential	$\frac{1}{2}e^{-\|u\|}I_{\{\|u\|\leq 1\}}$	
Shifted Exponential	$e^{-(u+1)}I_{\{\|u\|\leq 1\}}$	
Polynomial (ein Bsp.)	$\frac{15}{32}(1-u^2)^2(7-u^2)I_{\{\|u\|\leq 1\}}$	

Tabelle 4.1: Univariate Kern-Funktionen. Es sind einige gebräuchliche Kernfunktionen für stetige und diskrete Merkmale aufgeführt (siehe [3], [161] Kap. 9.3, [105], [109], [110] Kap. 2.1, [196] Kap. 6, [127]).

Eine weitere Möglichkeit, die Anzahl der Parameter zu verringern, sind sogenannte *Rotationskerne*[5] von der Form

$$K_p^{(R)}(x) := c_{K_1,p}\, K_1\left((x'x)^{\frac{1}{2}}\right), \tag{4.14}$$

mit $c_{K_1,p}^{-1} = \int K_1\left((x'x)^{\frac{1}{2}}\right)dx$ als Normierungskonstante ([221], Kap. 4.2). In den univariaten Kern fließt so allein der Betrag des Vektors ein. Die Kernschätzung gibt so allen Richtungen im Raum das gleiche Gewicht, während Produktkerne Richtungen zwischen den Koordinatenachsen für größere Werte stärker betonen. Durch die Betragsbildung machen Rotationskerne nur Sinn, wenn die Absolutwerte der Merkmale in ähnlichen Größenordnungen liegen. Diese müssen daher eventuell vorher skaliert oder die Betragsbildung in den einzelnen Merkmalen gewichtet werden (z.B. über $x'Hx$).

Es gibt nicht *den* Kern und *die* Bandweite für alle möglichen Problemstellungen. Es stellt sich so die Frage nach der Wahl des Kernes und der Wahl der Glättungsparameter.

Wahl des Kernes

Es existiert eine Vielzahl möglicher Kernfunktionen. Einige gebräuchliche Kerne sind in der Tabelle 4.1 aufgeführt. Aus der nichtparametrischen Kurven- und Dichteschätzung sind einige Fehlerkriterien bekannt, mit denen die Güte einer Schätzung erfaßt werden kann. Tabelle 4.2 enthält einige oft benutzte. So sind unter einigen Bedingungen die Schätzungen $\hat{f}_i^{(K)}$ (4.10) asymptotisch erwartungstreu und punktweise im MSE (vgl. Tabelle 4.2) konsistent ([161], Kap. 9.3).

Im Zusammenhang mit der Diskriminanzanalyse sind diese Fehlermaße aber nur bedingt aussagekräftig, da durch die andere Art des Fehlers (qualitativer 0-1-Verlust anstelle eines Abstandes) auch grob verfälschte Dichteschätzungen gute Ergebnisse

[5] auch: *rotationssymmetrische Kerne* (engl. *spherically* oder *radially symmetric kernels*)

Kriterium	*Definition*
Mittlerer Quadratischer Fehler (MSE)	$\mathrm{MSE}_x(\hat{f}) = \mathrm{E}_{\mathcal{S}} \left[\hat{f}(x) - f(x)\right]^2$
Integrierter Quadratischer Fehler (ISE)	$\mathrm{ISE}(\hat{f}) = \int \left[\hat{f}(x) - f(x)\right]^2 dx$
Mittlerer Integrierter Quadratischer Fehler (MISE)	$\mathrm{MISE}(\hat{f}) = \mathrm{E}_{\mathcal{S}} \int \left[\hat{f}(x) - f(x)\right]^2 dx$
Asymptotischer MISE (AMISE)	$\mathrm{AMISE}(\hat{f}) = \lim_{h\to\infty} \mathrm{MISE}(\hat{f})$

Tabelle 4.2: Fehlerkriterien zur Kurvenapproximation. Es sind einige Fehlerkriterien aus der allgemeinen Problematik der Kurvenapproximation aufgeführt (siehe [201] Kap. 3.1, [110] Kap. 1.2). Diese sind im Kontext der Diskriminanzanlyse aber nur bedingt aussagefähig (siehe dazu Kapitel 2.3).

bei einer Klassifizierung liefern können (siehe Kapitel 2.3), zumal nicht die reine Kurvenapproximation, sondern die des Quotienten der Dichten interessiert.

Welchen Kern sollte man nun für welche Datenarten benutzen?

- Für *stetige Daten* sind Normal-Kerne aufgrund ihrer analytischen Eigenschaften[6] sehr populär. Sie haben allerdings einen infiniten Support, so daß trotz lokaler Betrachtung alle Punkte der Stichprobe evaluiert werden müssen. Oft benutzt werden auch stückweise lineare und Polynomial-Kerne (z.B. EPANECHNIKOV-, Biweight- und Triweight-Kerne) mit sehr einfachen Funktionen und finitem Support, so daß nicht sämtliche Punkte der Stichprobe evaluiert werden müssen. SCOTT ([196], Kap. 6) betrachtet Kerndichteschätzungen unabhängig vom Kontext der Diskriminanzanalyse als Kurvenapproximation und ermittelt den EPANECHNIKOV-Kern hinsichtlich des AMISE (vgl. Tabelle 4.2) als optimal, wobei die anderen in Tabelle 4.1 aufgeführten Kerne ihm in der Effizienz kaum nachstehen.[7] Letztlich ist weniger die Wahl des Kernes entscheidend (von SCOTT als symmetrischer Polynomial-Kern mit geringer Ordnung empfohlen) als die Wahl der Glättungsparameter (siehe [196], Kap.6.2).

- Für multivariate unabhängige *binäre Daten* (d.h. $x \in \{0,1\}^p$) haben AITCHISON & AITKEN [3] ihren Binomial-Kern vorgeschlagen. Die Dichteschätzung hat mit $\lambda_i := h_i$ die Form

$$\hat{f}_i^{(\mathrm{A\&A})}(x) := \frac{1}{n_i} \sum_{j=1}^{n_i} \lambda_i^{p-d_{ij}^2} (1-\lambda_i)^{d_{ij}^2} \quad (\lambda_i \in [\frac{1}{2}, 1]\ , i = 1, 2), \tag{4.15}$$

  wobei $d_{ij}^2 = \|x - x_{i_j}\|^2$ die Anzahl der Nichtübereinstimmungen zwischen x und $x_{i_j} \in \mathcal{S}$ angibt. Mit $\lambda_i = 1$ erhält man den Multinomial-Kern und damit die Multinomial-Regel (4.2), also keine Glättung. Dagegen wird mit $\lambda_i = \frac{1}{2}$ als stärkster Glättung sämtlichen Beobachtungen das gleiche Gewicht beigemessen. Den hier nicht behandelten Fall unvollständiger Daten diskutiert TITTERINGTON [206].

[6] z.B. unendlich oft stetig differenzierbar

[7] Die meisten der in Tabelle 4.2 aufgeführten Kerne besitzen eine relative Effizienz von über 90 % bezüglich des EPANECHNIKOV-Kerns, d.h. der EPANECHNIKOV-Kern benötigt für einen gleich geringen Approximationsfehler ca. 10 % weniger Beobachtungen als diese Kerne ([196], Kap. 6).

- AITCHISON & AITKEN [3] und HAND [105], Kap. 5.2.2, erweitern obigen Ansatz auch auf *nominale und ordinale Merkmale* ($K_1^{(\mathrm{n})}$ bzw. $K_1^{(\mathrm{o})}$), deren Ausprägungen durch $1, \ldots, a$ kodiert sind:

$$K_1^{(\mathrm{n})}(u) \;=\; \begin{cases} \lambda & : \quad u = 0 \\ \frac{1-\lambda}{a-1} & : \quad u \neq 0 \end{cases} \tag{4.16}$$

$$K_1^{(\mathrm{o})}(u) \;=\; \begin{cases} \lambda_1 & : \quad u = 0 \\ \lambda_2 & : \quad u = 1 \\ \vdots & \\ \lambda_a & : \quad u = a-1 \end{cases} \qquad \left(\sum_{j=1}^{a} \lambda_j = 1 \right), \tag{4.17}$$

  wobei für den ordinalen Fall auch andere Gewichtungen denkbar sind (siehe [3]). Andererseits können qualitative Merkmale auch binärisiert werden (siehe Kapitel 2.4.2).

- Für *gemischt stetige und binäre Daten* schlagen AITCHISON & AITKEN [3] einen Produkt-Kern vor, anteilig aus univariaten Kernen für die stetigen (z.B. Normal-Kern) und für die binären Variablen (z.B. Binomial-Kern).

In der Praxis sind Produkt- oder Rotationskerne empfohlen. Letztere sind in kleinen Dimensionen p praktisch genauso effizient hinsichtlich des AMISE wie Produktkerne ([221], Kap. 4.5). Für theoretische Studien können vollständig multivariate Kerne, d.h. mit nichtdiagonaler Bandweitenmatrix $H \neq \operatorname{diag}(h_1, \ldots, h_p)$ in (4.10), betrachtet werden, um so Korrelationen zwischen den Variablen Rechnung zu tragen. Ein Beispiel ist die p-variate Normalverteilung N(0,Σ) als Kernfunktion. Diese ist gleichzeitig die einzige Kernfunktion, die als Produkt- und Rotationskern beschrieben werden kann ([221], Kap. 4.5). Für Details siehe [196], Kap. 6.3.2.

Für größere Dimensionen ($p \geq 10$) beobachtet man das „Phänomen des leeren Raumes“ ([197]) als eine Form des „Fluches der Dimensionalität“ (siehe Kapitel 2.5.1). Betrachtet man etwa eine Standardnormalverteilung, so liegen für $p = 1$ ca. 90 % der Masse im Intervall $[-1.6, 1.6]$, während für $p = 10$ ca. 99 % (!) aller Punkte einen größeren Abstand als 1.6 zum Ursprung haben ([201], Kap. 4.5). Mit zunehmender Dimension werden fast alle Beobachtungen an den „Rändern der Verteilung“ liegen (im Sinne von nicht im Zentrum liegend) und die Abstände zwischen ihnen immer größer werden, da sie höchstwahrscheinlich in mindestens einem Merkmal entfernt von dessen Mittelwert liegen werden. Die Form der Enden des Kernes wird so wichtiger (auch Regionen mit sehr niedriger Dichte können nun extrem wichtig für die Dichteschätzung sein), und die benötigten Stichprobenumfänge für gleichbleibendes Risiko werden rasch wachsen. HALL [100] schlägt bei vermuteten schweren Enden der Klassendichte den Kern

$$K_1(u) = \mathrm{const} \cdot \mathrm{e}^{-\frac{1}{2}[\log(1+|u|)]^2} \quad (u \in (-\infty, \infty))$$

vor, da der Normal- oder der Double-Exponential-Kern zu dünne Enden besitzen.

Da man im vornherein nicht exakt weiß, welche Bedingungen vorliegen, sollte einige Kerne ausprobiert werden. Jede Vorauswahl kann immer auch Vergrößerung des systematischen Fehlers (Bias) bedeuten. Wichtiger ist jedoch die Wahl der Bandweite.

Wahl der Bandweite

Art und Stärke der Glättung bestimmen den Bias und die Varianz der Dichteschätzung. Für kleine h_i wird $\hat{f}_i^{(K)}$ viele kleine Spitzen (Peaks) um die x_i herum aufweisen — die Schätzung hängt zu sehr von der Stichprobe ab (große Varianz durch Undersmoothing). Für große h_i wird $\hat{f}_i^{(K)}$ sehr stark geglättet — der Bias erhöht sich (Over-smoothing).

Für einmalige Analysen und bei $p \leq 3$ ist eine nichtautomatische Auswahl der h_i oft ausreichend. HAND ([105], Kap. 2.3) gibt folgende Empfehlungen für die Bestimmung der h_i zu einem gegebenen Kern „per Hand“:

- *Graphische Darstellung* der Randdichten von $\hat{f}_i^{(K)}$ und so eine Bestimmung der h_i per „Augenmaß“
- Bestimmung des *durchschnittlichen Abstandes* zwischen Objekten der Stichprobe und ihren q nächsten Nachbarn und Benutzung dieses Wertes als h_i (oft $q = 10$)[8]
- Versuche mit *mehreren Werten* für h_i und Auswahl in Hinblick auf das minimale Risiko bei einem Testdatensatz.

SILVERMAN ([201], Kap. 3.4.5) bietet die *Test-Graph-Methode* an. Über eine graphische Repräsentation von $\hat{f}_i'' - \mathrm{E}\hat{f}_i''$ läßt sich die Glattheit von $\hat{f}$ recht deutlich veranschaulichen und eine geeignete, subjektive Wahl von h_i treffen.

Kann man bestimmte Verteilungsannahmen über die Dichte f_i machen, lassen sich im Sinne des AMISE optimale Bandweiten angeben. Für eine angenommene unterliegende Normalverteilung und einen GAUSS-Kern erhält man so

$$\hat{h}_{\text{opt}} = (4\pi)^{-\frac{1}{10}} \frac{3}{8} \pi^{-\frac{1}{2}} \hat{\sigma}\, n^{-\frac{1}{5}}$$

([201], Kap. 3.4.2). Mehr robust gegen Schiefe und langen Enden der Verteilung ist

$$\hat{h}_{\text{opt}} = 0.79\, R\, n^{-\frac{1}{5}},$$

wobei R den IQR[9] bezeichnet ([201], Kap. 3.4.2). Zusammenfassend gibt SILVERMAN ([201], Kap. 3.4.2)

$$\hat{h}_{\text{opt}} = 0.9 \min\left\{\hat{\sigma}, \frac{R}{1.34}\right\} n^{-\frac{1}{5}}$$

als eine adäquate Wahl für viele Verteilungen wie Normal-, t-, Log-normal- und Mischungen von Normalverteilungen an.

Eine automatisierte Auswahl, entweder über das Risiko oder indirekt über Fehler bei der Dichtenapproximation, ist über Cross-validation- oder Bootstrap-Schätzungen (Kapitel 2.2.3) möglich. Durch Zerlegung des ISE-Kriteriums ,

$$\int \left[\hat{f}_i - f_i\right]^2 = \underbrace{\int \hat{f}_i^{\,2} - 2\int \hat{f}_i f_i}_{=:\, R(\hat{f}_i)} + \int f_i^{\,2},$$

[8] Das garantiert, daß bei Normal-Kernen durchschnittlich q Punkte innerhalb einer Standardabweichung liegen.

[9] *Interquartil range*: Differenz zwischen dem oberen und unteren Quartil (75. bzw. 25. Perzentile) einer Verteilung (siehe [183], S. 166)

erhält man das äquivalente Kriterium $R(\hat{f}_i)$ zur Minimierung des Approximationsfehlers der Dichtenschätzung. Eine kreuzvalidierte Schätzung von $R(\hat{f}_i)$ ist durch

$$M_0(h_i) := \int \hat{f}_i^{\,2} - 2\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1}{n_i-1}\frac{1}{h_i}\sum_{x\in\mathcal{S}_{-j}} K_p\left(\frac{x_{i_j}-x}{h_i}\right)$$

gegeben ([201], Kap. 3.4.3). M_0 kann durch einfacher zu berechnende Kriterien ersetzt werden. Anstatt dieses *Least-Squares-CV*-Kriterium (LS-CV) zu minimieren, kann eine *Likelihood-CV* (L-CV) maximiert werden. So schlagen AITCHISON & AITKEN [3] für den Binomial-Kern (4.15) über Leave-one-out die Maximierung der Pseudo-Likelihood-Funktion

$$\prod_{j=1}^{n_i} \hat{f}_i^{(K)}\left(x_{i_j}|\mathcal{S}_{-i_j}\cap\Pi_i,\lambda\right)$$

vor und zeigen die Konsistenz des mit dem maximierenden $\hat{\lambda}_{\max}$ arbeitenden Kernes für die Dichteschätzung. HALL [97] modifiziert diese Likelihood-Funktion, um gegen leere Zellen oder solche mit nur einer Beobachtung zu schützen.

Beide CV-Verfahren sind gut für stetige Daten geeignet; L-CV ist aber empfindlich gegenüber Ausreißern und neigt je nach Kern zum Under- oder Over-smoothing, während LS-CV unter geringen Bedingungen asymptotisch optimal für den ISE ist ([127]). Bei diskreten Daten ist jedoch Vorsicht geboten ([201], Kap. 3.4). Eine ausführliche Diskussion und weitere Ansätze sind bei SILVERMAN ([201], Kap. 3.4) und SCOTT ([196], Kap. 6.5) zu finden. JONES et al. ([132]) bieten einen Überblick über verschiedene Methoden. Die in diesen Arbeiten abgeleiteten Bandweiten sind aber nur bei der Approximation der Klassendichten optimal, nicht unbedingt für das Risiko der abgeleiteten Diskriminanzregel (siehe Kapitel 2.3 und [101], [157], [94], [95].

TUTZ ([212], [213]) diskutiert für den Binomial-Kern (4.15) eine kreuzvalidierte Wahl von h_i direkt in Hinsicht auf das BAYESsche Risiko. Für die diskriminatorische Anwendung der Kernschätzungen sind solche Ansätze am ehesten zu empfehlen, da hier die h_i eher größer gewählt werden, um eine Varianzreduktion zu Lasten einer eventuellen Biaserhöhung zu erzielen (siehe Kapitel 2.3).

4.2.2 Eigenschaften

Als nichtparametrisches Verfahren machen Kernschätzungen keine expliziten Annahmen über eine globale Form der Dichten oder Trennflächen, implizieren aber, daß für benachbarte Punkte oder Zellen gewisse Glattheitseigenschaften vorliegen: im stetigen Falle z.B. zweifach stetig differenzierbare und auf möglichst niedrigem Niveau beschränkte Dichten, bei Kontingenztafeln die Annahme, daß sich die erwarteten Beobachtungen in benachbarten Zellen nicht zu stark unterscheiden. Das bedeutet, daß sich die Wahrscheinlichkeitsverteilung in kleinen Umgebungen nur geringfügig ändert und dort geglättet werden kann ([105], Kap. 5.5). Bei genügend großen Stichproben können so beliebig komplizierte, aber relativ glatte Klassendichten geschätzt werden. Durch die starke Glättung besitzen Kern-Regeln eine vergleichsweise geringe Varianz bei relativ hohem Bias. Die mittels Kernschätzung abgeleitete Diskriminanzregel ist unter relativ geringen Voraussetzungen an den Kern konsistent im Sinne des Risikos (2.4) ([161], Kap. 9.6).

Ein großer Nachteil der Kernschätzungen ist, daß die gesamte Stichprobe für die Klassifizierung neuer Objekte verfügbar sein muß, während bei parametrischen

Verfahren nur einzelne Parameter gehalten werden. Kernschätzungen sind außerdem in der Anwendung langsamer als parametrische Verfahren wie die LDA oder QDA (Kapitel 3.2 und 3.3), da die Bestimmung der Regel praktisch erst bei ihrer Anwendung stattfindet — und dabei für jeden Punkt neu. Falls globale Bandweiten benutzt werden, ist die Kern-Regel nicht skaleninvariant. Die Dichteschätzung erbt die Eigenschaften des Kernes.

In niedrigen Dimensionen an Merkmalen können Kernschätzungen für eine möglichst genaue Dichteapproximation sinnvoll sein, für $p > 5$ legen theoretische Argumente und praktische Erfahrungen nahe, daß Kernschätzungen sinnlos sind ([196], Kap. 7.2). Durch die andere Art des Fehlers (siehe Kapitel 2.3) können solche Schätzungen in der Diskriminanzanalyse jedoch sehr wohl für $p > 5$ erfolgreich sein. Friedman [73] macht außerdem geltend, daß multivariaten Daten im hochdimensionalen Raum oft auf einer Mannigfaltigkeit oder Hyperebene sehr viel geringerer Dimension liegen. Schiefe und schwere Enden in der Verteilung verschlechtern die Kernschätzungen in Hinsicht auf den AMISE. Die Kernschätzung via Produkt-Kern ist gut geeignet für stetige Daten mit höchstens leicht korrelierten Variablen ([161], Kap. 9.6). Besonders bei bezüglich p kleinen n_i neigen Kernschätzungen zum Over-fitting, d.h. einer Vergrößerung der Varianz ([107]).

Verschiedene Studien über die Klassifikationsgüte von Kern-Diskriminanzregeln wurden angestellt:

- *Stetige Daten*: Remme et al. [178] vergleichen die LDA und QDA (Kapitel 3.2 und 3.3) mit einem Produkt-Gauss-Kern (h_i über CV bestimmt) bei nur gering korrelierten Variablen ($p = 2, 6$). Für multivariat normale Daten erhalten sie erwartungsgemäß bei Homoskedastizität die LDA und bei Heteroskedastizität die QDA als Optimum im Sinne des Risikos. Im ersteren Fall ist die Kern-Regel für kleinen Populationsabstand gut, im zweiten der QDA ebenbürtig ($n_i = 35$) oder bei kleinen Stichproben ($n_i = 15$) sogar überlegen. Bei größeren Korrelationen zwischen den Variablen wird die Kern-Regel aber durch zu starken Bias aufgrund des Produktkernes zunehmend schlecht. Bei log-normalen Daten ist die Kern-Regel leicht besser als LDA und QDA; ebenso bei Mischungen aus Normalverteilungen, d.h. in eher „nichtparametrischen" Fällen, in denen die ideale Trennfläche weder linear noch quadratisch ist und die parametrischen Verfahren einen zu hohen Bias verursachen. Van Ness [215] stellt die Notwendigkeit einer Kopplung der Bandweiten an die Populationskovarianzen heraus.

- *Diskrete Daten*: Aitken [5] wendet die LGA (Kapitel 3.4), Nearest Neighbours (Kapitel 4.3), den Binomial-Kern u.a. auf zwei reale binäre Datensätze an und erhält den Binomial-Kern einmal als bestes und einmal als schlechtestes Verfahren (mit der LGA führend). Titterington et al. [210] untersuchen qualitative Daten von 1.000 Personen mit Kopfverletzungen, wovon 500 als Testdatensatz dienen, in unterschiedlicher Auswahl der Variablen ($p = 4, 6, 10$). Im Vergleich zu LDA, LGA, QDA, UMA (Kapitel 4.1.1) u.a. schneidet die Regel mit einem Binomial-Kern immer relativ schlecht ab, besonders für die größeren p. Ebenfalls schlechte Ergebnisse in höheren Dimensionen ($p = 10$) erhält Titterington [207] in einer weiteren Studie mit binären Daten, wobei die Kerne mit einem globalen h_i pro Klasse noch am besten abschneiden. Hand [107] vergleicht die LDA und eine Binomial-Kern-Regel in 6 multivariat binären Datensätzen. Bei großen p wird die Kern-Regel vergleichsweise schlecht, da sie durch zu große Flexibilität zum Over-fitting neigt.

Das kann aber auch auf eine zu kleine Wahl der Bandweite hindeuten. Der zusätzliche Vergleich zwischen dem Binomial-Kern und der Modifikation durch HALL [97] bringt keine diskriminatorischen Unterschiede, gleichwohl die über CV bestimmten Bandweiten bei letzterem in allen Fällen größer sind.

- *Gemischte Daten*: VLACHONIKOLIS & MARRIOTT [219] vergleichen für gemischt binäre und stetige Daten in zwei Datensätzen die LDA, LGA und das Lokationsmodell (Kapitel 4.1.2) gegen einen gemischten Binomial-Normal-Kern und erhalten letzteren als weniger effektiv, wobei LGA und LDA annähernd gleichgut klassifizieren. SCHMITZ et al. [191] untersuchen Herzinfarkt-Daten und erhalten für 3 binäre und 9 stetige Merkmale ähnliche Ergebnisse von LDA, LGA und Kern-Regel, nur die QDA fällt ab. Eine weitere Analyse dieser Daten durch SCHMITZ et al. [192] mit 3 binären und 3 stetigen Variablen ergibt ein vergleichbares Verhalten von LGA und LDA und die Empfehlung, das jeweils bessere Verfahren von LDA und QDA zu benutzen. Die Kern-Regel ist nie optimal, folgt aber meist dem Optimum nach (insbesondere bei „nichtparametrischen“ Strukturen, die nicht eindeutig ein parametrisches Verfahren favorisieren). In einer Simulationsstudie generieren SCHMITZ et al. [193] eine 4-dimensionale Normalverteilung, von der nachfolgend drei Dimensionen diskretisiert werden. Hier ergibt sich die Kern-Regel als durchschnittlich beste; und die Autoren empfehlen eine Wahl zwischen LDA und Kern-Regel.

Für wachsende n_i werden die Kern-Regeln durch Varianzreduktion rasch besser. Dabei sind sie paradoxerweise vor allem bei kleinen n_i anderen, parametrischen Methoden überlegen ([216]), was möglicherweise auf die ihnen innewohnende starke Glättung zurückzuführen ist, also dem Erzielen einer geringen Varianz auch bei kleinen Stichprobenumfängen. Eine Diskriminanzregel, die auf Kernschätzungen beruht, scheint so vor allem bei kleinen Dimensionen p, in „nichtparametrischen“ Situationen (z.B. Mischungen von Dichtefunktionen oder Wendepunkte in $\log \frac{f_1(x)}{f_2(x)}$) und bei relativ kleinen Stichprobenumfängen n_i in Frage zu kommen. Für weitere Details siehe [196], Kap. 6.6, und [201], Kap. 5.3.

4.2.3 Erweiterungen*

Die obigen Kernschätzungen mit globalen Bandweiten haben vor allem Probleme mit schiefen Verteilungen, die der Symmetrie der Kernfunktion widersprechen, und an den Enden von Verteilungen, an denen die Anzahl der Objekte in der Stichprobe gering wird. Das ist insbesondere bei höheren Dimensionen p wichtig (siehe oben): Ein Beharren auf dem gewählten h_i erhöht die Varianz, eine stärkere Glättung den Bias. Die nachfolgenden Modifikationen bemühen sich deswegen, die Glättungsparameter nicht nur von den Variablen, sondern auch von der Lage der Objekte im Raum abhängen zu lassen.

WAND et al. [222] untersuchen bei schiefen Verteilungen Transformationen der Variablen, die zu symmetrischen Verteilungen führen (z.B. BOX-COX-Transformationen, siehe Kapitel 2.4.2), um so Verbesserungen der Kernschätzungen mit einer globalen Bandweite zu erreichen. Die rücktransformierten geglätteten Werte werden dann als Dichteschätzung in die Regel (4.11) eingesetzt (siehe auch [225]).

HALL & WAND [104] greifen eine Schwäche der gewöhnlichen Kernschätzung auf: Die Klassendichten werden getrennt geschätzt und die gruppenspezifischen Bandweiten getrennt z.B. nach dem MISE-Kriterium bestimmt. Da man aber in der Diskriminanzanalyse weniger an der Modellierung der einzelnen Dichten als an ihrem

Verhältnis zueinander interessiert sind, schlagen sie im Fall $k = 2$ für symmetrische Kosten einen zur BAYESschen Regel (2.10) äquivalenten Ausdruck vor:

$$\frac{f_1(x)}{f_2(x)} \geq \frac{\pi_2}{\pi_1} \quad \Longleftrightarrow \quad g(x) := \pi_1 f_1(x) \; - \; (1 - \pi_1) f_2(x) \geq 0. \tag{4.18}$$

Die Zuordnung erfolgt so über

$$\hat{\delta}_{\mathrm{H\ \&\ W}}(x) = \left\{ \begin{array}{ll} 1: & g(x) \geq 0 \\ 2: & \text{sonst} \end{array} \right. . \tag{4.19}$$

Bestimmt man nun die Bandweiten h_1, h_2 simultan so, daß der MISE$(\hat{g})$ minimal wird, steht zur Schätzung dieser beiden jeweils die gesamte Stichprobe zur Verfügung. Außerdem wird der eigentlich interessierende Quotient als Ganzes in einer numerisch angenehmen Form modelliert.

CWIK & MIELNICZUK [44] schlagen einen Kernschätzer für Rangdaten vor.

Variable Kerne

BREIMAN et al. [29] kombinieren den Kernschätzungs- mit dem Nearest-Neighbours-Ansatz (Kapitel 4.3), indem der Glättungsparameter $h(\{x_j\}_j)$ vom Abstand d_{jk} des k-ten nächsten Punktes in der Stichprobe zu $x_{i_j} \in \mathcal{S}$ abhängt:

$$\hat{f}_i^{(vK)}(x) := \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{(\alpha_k d_{jk})^p} K_p \left(\frac{x - x_{i_j}}{\alpha_k d_{jk}} \right) \quad (i = 1, 2). \tag{4.20}$$

Dabei ist α_k ein konstanter Faktor. So wird d_{jk} in Gebieten mit niedriger Dichte sehr groß, so daß weit entfernte Punkte ein größeres Gewicht erhalten, während bei einer Vielzahl von Punkten in der nächsten Umgebung d_{jk} sehr klein wird mit dem umgekehrten Effekt. Dieser sogenannte *Variable Kern* adaptiert sich an die lokale Datendichte in der Stichprobe und führt so eine sehr grobe Bias-Varianz-Abwägung durch .

Zur Bestimmung der optimalen k und α_k schlagen die Autoren eine Goodness-of-fit-Statistik vor:

$$\hat{S} := \sum_{j=1}^{n_i} \left(\hat{\omega}_{[j]} - \frac{j}{n_i} \right)^2 .$$

Dabei gilt:

$$\hat{\omega}_j = \mathrm{e}^{-n_i f_i(x_{i_j}) V(d_{j,1})} \quad \text{und} \quad \hat{\omega}_{[1]} \leq \ldots \leq \hat{\omega}_{[n_i]}.$$

$V(r) := \pi^{\frac{p}{2}} \frac{r^p}{\Gamma(\frac{p}{2}+1)}$ bezeichnet hierbei das Volumen der p-dimensionalen Sphäre mit Radius r. $h_i(x_{i_j})$ ist so proportional zu $f_i(x_{i_j})^{-\frac{1}{p}}$ ([184]).

Der Ansatz ist L_1-konsistent ([49]). Gegenüber Nearest Neighbours liefert er eine echte Dichteschätzung, die darüber hinaus alle Eigenschaften des Kernes erbt, also je nach Kern-Wahl auch Glattheit und analytische Eigenschaften, während Nearest Neighbours unstetige Schätzungen liefert. Für $p = 2$ erhalten die Autoren in zwei Datensätzen ($n = 400$) eine bessere Approximation der Dichten hinsichtlich verschiedener Fehlerkriterien ([29]).

ABRAMSON [1] schlägt eine modifizierte Schätzung mit $h_i(x_{i_j}) \propto f_i(x_{i_j})^{-\frac{1}{2}}$ vor.

Adaptive Kerne

Eine weitere Verallgemeinerung, die die Variablen Kerne mit enthält, sind die *Adaptiven Kerne*. Während bei Kernen mit globalen Bandweiten diese höchstens für die Klasse oder die einzelnen Variablen spezifisch waren, sind solche h denkbar, die sogar von den einzelnen Realisierungen der unbekannten Dichte f an den Stellen $x, x_1, \dots, x_n$ abhängen ([196], Kap. 6.6), so daß man als *adaptive Bandweite* eine Funktion erhält:

$$h := h(x, x_j, \mathcal{S}) \approx h(x, x_j, f_i). \tag{4.21}$$

Analytisch kann man die optimale adaptive Bandweite $h(x, x_j, f_i)$ als eine glatte und langsam variierende Funktion annehmen, so daß sich für $n < \infty$ nach Scott ([196], Kap. 6.6; siehe auch [184]) zwei verschiedene Ansätze einer adaptiven Schätzung ergeben:

- $h_x := h(x, x, f_i)$: Die Bandweite ist nur eine Funktion des zu klassifizierenden Punktes, die mit unterschiedlichen x variiert, aber für festes x an allen Punkten aus der Stichprobe gleich ist.
- $h_j := h(x_j, x_j, f_i)$: Die Bandweite ist spezifisch für jeden Punkt der Stichprobe, wird aber auf jeden zu klassifizierenden Punkt x angewendet.

Bei den Variablen Kernen wird der zweite Ansatz verwendet: Die Bandweite h_j wird je nach Datendichte in $\mathcal{S}$ um x_j herum gewählt. Die Nearest-Neighbours-Methode (Kapitel 4.3) setzt h_x gleich der Entfernung des k-ten nächsten Punktes aus $\mathcal{S}$ zum zu klassifizierenden Punkt x und kann so als Spezialfall eines Adaptiven Kernes gesehen werden. Im ersten Ansatz ist $\{h_x\}$ im stetigen Fall ∞-dimensional, so daß es im vornherein schwer bestimmbar ist, während $\{h_j\}$ immer n_i Dimensionen aufweist. Hier soll nur der zweite Ansatz erläutert werden.

Grundidee ist eine zweistufige Prozedur ([201], Kap. 5.3). Zunächst wird eine grobe *Pilotschätzung* $\tilde{f}_i$ der Dichte gefunden, für die $\tilde{f}_i(x_{i_j}) > 0\,(j = 1, \dots, n_i)$ gilt. Darüber werden die *lokalen Bandweiten* $(\lambda_1, \dots, \lambda_{n_i})$ als

$$\lambda_j := \left(\frac{\tilde{f}_i(x_{i_j})}{g_i}\right)^{-\alpha} \quad (j = 1, \dots, n_i) \tag{4.22}$$

bestimmt, wobei g_i das geometrische Mittel der $\tilde{f}_i(x_j)$,

$$\log g_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \log \tilde{f}_i(x_{i_j}) \quad (j = 1, \dots, n_i), \tag{4.23}$$

und $\alpha \in [0, 1]$ ein sogenannter *Sensitivitätsparameter* ist. Als endgültige, *adaptive Kernschätzung* erhält man dann

$$\hat{f}_i^{(aK)}(x) := \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{(h_i \lambda_j)^p} K_p \left(\frac{x - x_{i_j}}{h_i \lambda_j}\right) \quad (i = 1, 2), \tag{4.24}$$

mit h_i als klassenspezifischer Bandweite und λ_j als Faktor für die Bandweite am Punkt $x_{i_j} \in \mathcal{S} \cap \Pi_i$.

Die adaptive Schätzung ist wenig empfindlich gegenüber der Art der Pilotschätzung, für die oft Kernschätzungen mit globaler Bandweite (insbesondere mit schnell zu berechnenden Kernen) oder Nearest Neighbours benutzt werden ([201], Kap. 5.3).

Für α empfehlen SILVERMAN ([201], Kap. 5.3) u.a. die Wahl von $\frac{1}{2}$, obwohl auch hier eine CV-Bestimmung möglich ist. Für $\alpha = 1$ erhält man wieder globale Bandweiten. Mit der adaptiven Kernschätzung kann im Falle einer Normalverteilung in numerischer Approximation eine Verringerung des MISE um die Hälfte gegenüber globalen Bandweiten erzielt werden ([196], Kap. 6.6). Die asymptotische Ineffizienz Adaptiver Kerne ([196]) spricht daher nicht gegen deren Anwendung bei kleineren Stichprobenumfängen n_i.

Für eine Übersicht und Referenzen siehe die Arbeiten von SCOTT [196] und SAIN & SCOTT [184]. Die letztgenannten Autoren schlagen außerdem eine abgewandelte Form der Berechnung der Kernschätzung vor, indem Werte von $K(.)$ auf einem Gitter berechnet und diese dann, gewichtet nach der relativen Häufigkeit der x_j in der Umgebung der Gitterpunkte, gemittelt werden. Durch diesen sogenannten *Binned*[10] *Kernel Estimator* erfolgt so ebenfalls eine adaptive Bandweitenbestimmung.

4.3 Nearest Neighbours

4.3.1 Das Verfahren

Während klassische Kernschätzungsmethoden (Kapitel 4.2) für die Klassifizierung eines Punktes um diesen herum eine Umgebung definieren und alle Punkte aus dieser Umgebung für eine Entscheidung betrachten, gehen *Nearest-Neighbours*-Verfahren (k-NN) den umgekehrten Weg: Gemäß einer definierten Zahl $k \in \{1, \ldots, n\}$ werden um den zu klassifizierenden Punkt x die k nächsten Punkte bezüglich eines Abstandes $D(x, x_j)$ aus der Stichprobe $\mathcal{S}$ zur lokalen Dichteschätzung herangezogen.[11] Nearest-Neighbours-Verfahren können als Spezialfall Adaptiver Kerne (Kapitel 4.2.3) gesehen werden. Sie sind einer der ältesten nichtparametrischen Ansätze zur Dichteschätzung ([64]), sehr bekannt und verbreitet.

Es sei $N_{x,k}$ die Nachbarschaft um x, die genau die zu x nächsten k Punkte beider Klassen aus $\mathcal{S}$ enthält.[12] Eine Dichteschätzung erhält man dann durch

$$\hat{f}_i^{(k-NN)}(x) := \frac{k_i}{n_i} \frac{1}{V_{x,k}} \quad (i = 1, 2) \tag{4.25}$$

([105], Kap. 2.4). Dabei sollen k_i die Anzahl der Punkte in $N_{x,k}$ aus $\mathcal{S} \cap \Pi_i$ und $V_{x,k} = \int_{N_{x,k}} dP$ $(P = \pi_1 P_1 + \pi_2 P_2)$ das Volumen von $N_{x,k}$ bezeichnen. Die in $N_{x,k} \cap \mathcal{S} \cap \Pi_i$ liegenden Punkte tragen also zu gleichen Teilen zur Dichteschätzung bei, unabhängig von ihrem Abstand zu x. Allerdings ist $\hat{f}_i^{(k-NN)}$ keine wirkliche Dichteschätzung, da das Integral über Ω unendlich wird ([105]). Da man für die Klassifikationsregel jedoch nur am Dichtequotienten interessiert ist, entsteht daraus kein Problem. Die geschätzten Dichten werden in die BAYESsche Regel (2.10) eingesetzt (Plug-in), so daß die Diskriminanzregel folgendes Aussehen besitzt:

$$\hat{\delta}_{k-NN}(x) = \begin{cases} 1: & \frac{\pi_1 \frac{k_1}{n_1}}{\pi_2 \frac{k_2}{n_2}} \geq \frac{c_{21}}{c_{12}} \\ 2: & \text{sonst} \end{cases} . \tag{4.26}$$

Für $\pi_1 c_{12} = \pi_2 c_{21}$ und $n_1 = n_2$ ordnet die Regel das Objekt x derjenigen Klasse zu, aus der die Mehrzahl der Punkte $(\arg\max_i k_i)$ in der Umgebung $N_{x,k}$ stammt.

[10] zu deutsch: Behältnis

[11] In diesem Teilkapitel bezeichnet k nicht die Anzahl der Klassen, sondern aus historischen Gründen die Anzahl der betrachteten Punkte um x.

[12] Bei gleichen Abständen wird das Objekt mit dem niedrigeren Index gewählt ([48]).

Ähnlich den Kernschätzungen gibt es auch bei Nearest Neighbours zwei Wahlmöglichkeiten, die das Verhalten der Diskriminanzregel beeinflussen, indem sie die Nachbarschaft definieren: die Anzahl der betrachteten Nachbarn k und der Abstand D.

Die Dichteschätzung bezieht lokal die k nächsten Nachbarn zum zu klassifizierenden Punkt mit ein. Diese Nähe wird durch ein *Abstandsmaß* D definiert. Gebräuchlich für stetige Daten ist der EUKLIDische Abstand

$$D_2(x, y) = \sqrt{\sum_{j=1}^{p} (x_{(j)} - y_{(j)})^2} \tag{4.27}$$

als Spezialfall der MINKOWSKI-Metrik (2.35) bei $\lambda = 2$. Damit haben die Umgebungen immer die Form p-dimensionaler Kugeln. Die sofort ersichtliche Skalenvarianz bewirkt, daß beispielsweise Messungen eines Merkmals in Metern bzw. Millimetern zu verschiedenen Nachbarschaften mit anderen Punkten und so zu anderen Regeln führen können. Die Messung von Merkmalen in größeren Einheiten bewirkt eine Verringerung ihres Einflusses.

Verallgemeinernd kann der Abstand auch in Abhängigkeit von der MINKOWSKI-Metrik (2.35) und einer Gewichtsmatrix $W \in \mathcal{M}_{p\times p}$ definiert werden ([73]):

$$D_{\lambda,W}(x, y) = \left[\sum_{j=1}^{p} \left(\left(W(x-y)\right)_{(j)}\right)^{\lambda}\right]^{\frac{1}{\lambda}}. \tag{4.28}$$

Der EUKLIDische Abstand ist mit D_{2,I_p} darin enthalten. Eine Skalierung der einzelnen Variablen ist durch $W = \text{diag}(w_1, \ldots, w_p)$ möglich, so daß bei $\lambda = 2$ aus den Kugeln Ellipsoide werden. FRIEDMAN [73] listet als mögliche w_j^{-1} auf:

- die *Standardabweichung*: $w_j = \left[\frac{1}{n}\sum_{l=1}^{n}\left(x_{l(j)} - \bar{x}_j\right)^2\right]^{-\frac{1}{2}}$
- den *Extreme Range*: $w_j = \left[x_{[n](j)} - x_{[1](j)}\right]^{-1}$
- den *Interquartil Range*: $w_j = \left[x_{[\frac{3}{4}n](j)} - x_{[\frac{1}{4}n](j)}\right]^{-1}$

Damit wird die Klassifikationsentscheidung unabhängig von den einzelnen Meßskalen, wenngleich Skalierungseffekte durch die Wahl von W auftreten können, denn nun liegen alle Variablen in ähnlicher Größenordnung vor. RIPLEY ([181], Kap. 6.2) trägt weitere Metriken zusammen.

Die Definition eines Abstandes für diskrete Daten ist analog. Für binäre Merkmale läßt sich z.B. der Abstand als Anzahl der Nichtübereinstimmungen zwischen x und $x_{i_j} \in \mathcal{S}$, ${d_{ij}}^2 := \|x - x_{i_j}\|^2$, definieren ([3]). Anders als bei stetigen Daten kann jedoch die Wahrscheinlichkeit größer als Null sein, daß für einen gegebenen Abstand d die Anzahl der in $N_{x,k}$ enthaltenen Punkte größer als k ist, für $d-1$ jedoch kleiner als k. HAND ([105], Kap. 5.2.3) schlägt für diesen Fall vor, eine zufällige Auswahl aus den Punkten, die genau den Abstand d zu x haben, so zu treffen, daß wieder genau k Punkte die Nachbarschaft $N_{x,k}$ bilden. Für weitere Abstandsmaße siehe Kapitel 2.4.2. Für gemischte Daten sind Produkte von Abstandsmaßen, nach diskreten und stetigen Merkmalen getrennt, anwendbar; siehe z.B. [224].

Die *Wahl von* k hat dieselbe Bedeutung wie die Wahl der Bandweite h bei Kerndichteschätzungen (Kapitel 4.2): Große k bewirken eine stärkere Glättung als kleine und, damit verbunden, eine Verringerung der Varianz bei möglicher Erhöhung des Bias. k wirkt so als Regularisierungsparameter. LOFTSGAARDEN & QUESENBERRY [153] schlagen $\sqrt{n_i}$ als klassenspezifische Wahl für k vor. ENAS & CHOI [59] schlagen für zwei Klassen aufgrund von Simulationsstudien $n^{\frac{3}{8}}$ für große und $n^{\frac{2}{8}}$ für kleine Unterschiede in den Klassenkovarianzen vor. Sicherer erscheint der Vergleich mehrerer Werte für k (etwa $k = 1, \ldots, 10$) über die Schätzung des Risikos via Crossvalidation oder Bootstrap (siehe Kapitel 2.2.3). Im übrigen scheint die Wahl von k laut HAND ([107], Kap. 7.3) relativ unkritisch.

4.3.2 Eigenschaften

Das Nearest-Neighbours-Verfahren setzt als nichtparametrisches Verfahren keine allgemein parametrische Form der Dichten voraus, sondern als grundsätzliche und einzige Annahme die Glattheit der Dichten, d.h. daß sich der Wert von $f_i(x)$ in der Umgebung von x nur wenig ändert. Es ist darum flexibel und kann auf unterschiedliche Verteilungen angewandt werden. Bei Verwendung der EUKLIDischen Metrik zeigt das Verfahren die besten Ergebnisse, falls die betrachteten Variablen annähernd gleich bedeutsam für die Klassifizierung sind ([73]). Falls die Variablen jedoch zunehmend unterschiedlichen Einfluß besitzen, wird das Risiko schnell schlechter.

Wie bei der Wahl des Abstandsmaßes gesehen, ist das Nearest-Neighbours-Verfahren nicht skaleninvariant. Es liefert außerdem diskontinuierliche Dichteschätzungen. Auch ist es, wie die Kernschätzungen, empfindlich gegenüber dem „Fluch der Dimensionalität“ (siehe Kapitel 2.5.1): Schon für moderate p müssen riesige Stichprobenumfänge vorhanden sein, um sogar die kleinste mögliche Nachbarschaft $N_{x,1}$ „lokal“ und damit den Bias unter Kontrolle zu halten ([73]). In hohen Dimensionen sind die Abstände zwischen sämtlichen Objekten der Stichprobe groß und annähernd gleich, was einen Verlust an Information über die Verteilungen bedeutet ([72]). Mit zunehmendem p vergrößert sich entweder der Bias der lokalen Dichteschätzung, falls k konstant bleibt und somit die Umgebung größer wird; oder es vergrößert sich deren Varianz, falls man k verringert, um so die Umgebung klein zu halten, und damit weniger Objekte für die Dichteschätzung zur Verfügung hat. Für große p relativ zu n wird selbst bei $k = 1$, d.h. bei maximaler Varianz, der Bias groß sein ([75]).

Sämtliche Punkte der Stichprobe müssen für Klassifizierungen verfügbar gehalten werden; und für große Umfänge kann die Berechnung der Regel langsam werden ([105], Kap. 2.4). Der Aufwand für eine Klassifizierung liegt weniger beim Zeitpunkt der Bestimmung der Regel als bei ihrer Anwendung — dies im Gegensatz etwa zum CART-Verfahren (Kapitel 4.4), bei dem das Aufstellen der Regel sehr rechenintensiv, dafür ihre Anwendung extrem schnell ist ([75]). Jedem Punkt innerhalb der Nachbarschaft wird dasselbe Gewicht gegeben, unabhängig von seiner Entfernung zum zu klassifizierenden Punkt, so daß Ausreißer oder sehr weit entfernte Punkte in der Stichprobe — im Gegensatz zu klassischen Kerndichteschätzungen — verfälschenden Einfluß auf die Dichteschätzung nehmen können. Außerdem liegt das zu klassifizierende Objekt zwar im Zentrum seiner Umgebung, kann aber bei Betrachtung der konvexen Hülle für sich und seine Nachbarschaft an deren Rand liegen. Daraus ergibt sich bei geringer Varianz ein oft großer Bias.

Trotz dieser Nachteile wird das Nearest-Neighbours-Verfahren in verschiedenen Vergleichsstudien oft unter den besten gefunden. FRIEDMAN [73] begründet das mit der anderen Art des Klassifikationsfehlers gegenüber dem Fehler bei der Dichtenapproximation (siehe Kapitel 2.3) — starkes Glätten muß nicht zwangsläufig das Risiko verschlechtern, so daß k oft größer als im Regressionskontext gewählt wird — und mit der Erfahrung, daß multivariate Daten oft auf einer niederdimensionalen Mannigfaltigkeit liegen und so den „Fluch“ umgehen. Die Nearest-Neighbours-Schätzung ist besser zu interpretieren als Approximation des Dichte*quotienten* anstelle einer reinen Dichteschätzung, wie aus (4.26) ersichtlich ist: Es sind weniger die absoluten Größen der Dichten interessant als ihr Verhältnis zueinander. Analytisch hat die Nearest-Neighbours-Methode außerdem die interessante Eigenschaft, asymptotisch ($n \to \infty$) höchstens zweimal so schlecht bezüglich des Risikos zu sein wie die optimale BAYESsche Regel[13] ([105], Kap. 2.4) und für $k_n \stackrel{n}{\to} \infty$ mit $\frac{k_n}{n} \stackrel{n}{\to} 0$ das BAYESsche Risiko sogar zu erreichen ([48]). Dies begründet unter anderem die weite Verbreitung der Nearest-Neighbours-Verfahren. Allerdings hängt die Fehlerrate in praktischen Problemstellungen oft nicht von n ab ([161], Kap. 9.7).

4.3.3 Erweiterungen*

HAND ([105], Kap. 2.4) führt Beiträge verschiedener Autoren über modifizierte Nearest-Neighbours-Verfahren zusammen. Das *Condensed-NN*-Verfahren versucht, in geeigneter Weise Objekte aus der Stichprobe wegzulassen, um so die Anzahl der zu berechnenden Distanzen zu verringern und das Verfahren zu beschleunigen. Die Modifikationen *Reduced-NN* und *Edited-NN* versuchen, Schwächen dieses Verfahrens zu beseitigen. Da diese Ansätze vor allem in Hinblick auf die bis in die 1970er Jahre knappen Rechenkapazitäten entwickelt wurden, sollen sie hier nicht weiter behandelt werden.[14]

HILL [98] entwickelt einen adaptiv gewichteten Nearest-Neighbours-Schätzer für multivariat binäre Daten, der den Approximationsfehler von $\hat{f}_i$ im Sinne des MSE minimiert. HELLMAN [121] betrachtet ein Nearest-Neighbours-Verfahren mit einer Reject-Option[15].

Flexible Metric Nearest Neighbours

FRIEDMAN [73] schlägt mit *Flexible Metric Nearest Neighbours* (FMNN) eine Adaptierung der Gewichtsmatrix W in (4.28) an den zu klassifizierenden Punkt x vor ($W = W(x)$). Hierbei benutzt er Methoden der rekursiven Partitionierung (siehe Kapitel 4.4), um die Umgebung des zu klassifizierenden Punktes zu definieren. Grundidee ist die Annahme, daß Variablen an verschiedenen Punkten x in unterschiedlichem Maße zur Klassifizierung beitragen. Er definiert daher eine *lokale Relevanz* der Variablen:[16]

Unter der Bedingung, daß die j-te Komponente von x den Wert y_j annimmt ($x = (x_{(1)}, \ldots, x_{(p)})'$), liefert der Quadratische Vorhersagefehler

$$I_j^2(y_j) = \Big[\mathrm{E}f - \mathrm{E}(f|x_{(j)} = y_j)\Big]^2 \quad (j = 1, \ldots, p) \tag{4.29}$$

den absoluten Beitrag der j-ten Variable zum Vorhersagefehler im Punkt x. Für einen beliebigen Punkt $y = (y_1, \ldots, y_p)'$ erhält man so den relativen Einfluß r_j der j-ten Variable (*Relevanz*) als

$$r_j^2(y) = \frac{I_j^2(y_j)}{\sum\limits_{l=1}^{p} I_l^2(y_l)} \quad (j = 1, \ldots, p). \tag{4.30}$$

[13] MCLACHLAN ([161], Kap. 9.7) beschreibt das heuristisch so, daß „nur die Hälfte der Information der Stichprobe benutzt wird".

[14] Zu *Data Editing* als Mittel der Verkleinerung sehr großer Stichproben siehe auch [181], Kap. 6.2.

[15] *Reject-Option* bedeutet die Möglichkeit des Zurückweisens einer Entscheidung über ein Objekt; solche Objekte liegen dann im *Indifferenzbereich* der Diskriminanzregel (siehe Kapitel 2.6).

[16] Für eine ausführliche Herleitung siehe [73].

Um dieses Maß noch stärker lokal um y zu definieren, betrachtet man bei der Bildung des Erwartungswertes in (4.29) nur eine Umgebung $\mathcal{U}(y)$ um y, also $\mathrm{E}\Big|_{\mathcal{U}(y)}$.

In der Diskriminanzanalyse ist man an den A-posteriori-Verteilungen $\tau_i(x)$ $(i = 1, 2)$ (siehe Kapitel 2.1) als den zu modellierenden Funktionen f interessiert. Definiert man $z_i = 1 \iff x \in \Pi_i$ (und 0 sonst), ist (4.29) mit der Umgebung $\mathcal{U}(y)$ äquivalent zu

$$I_j^2(y_j) = \left[\mathrm{E}\Big|_{\mathcal{U}(y)} z_i - \mathrm{E}\Big|_{\mathcal{U}(y)} (z_i|x_{(j)} = y_j)\right]^2 \quad (j = 1, \ldots, p). \tag{4.31}$$

Eine Schätzung von $\mathrm{E}\Big|_{\mathcal{U}(y)} z_i = \pi_i\Big|_{\mathcal{U}(y)}$ ist durch

$$\hat{\mathrm{E}}\Big|_{\mathcal{U}(y)} z_i = \frac{\sum\limits_{l=1}^{n} z_{li}\, I_{x_l \in \mathcal{U}(y)}}{\sum\limits_{l=1}^{n} I_{x_l \in \mathcal{U}(y)}} \quad (i = 1, 2) \tag{4.32}$$

gegeben. Da vor allem bei stetigen Daten die Anzahl der $x \in \mathcal{S}$ mit $x_{(j)} = y_j$ gering oder gleich Null ist, muß der zweite Term in (4.31) glättend durch eine Umgebung, z.B. Nearest Neighbours, geschätzt werden. Dazu sei L als die Anzahl der lokal betrachteten Punkte gegeben und Δ_j als der Abstand entlang der j-ten Variable, durch den genau L Objekte aus der Stichprobe eingeschlossen werden. Lokal in der Umgebung $\mathcal{U}(y)$ ergibt sich so:

$$\hat{\mathrm{E}}\Big|_{\mathcal{U}(y)} (z_i|x_{(j)} = y_j) = \frac{\sum\limits_{l=1}^{n} z_{li}\, I_{|x_{l(j)} - y_j| \le \Delta_j} I_{x_l \in \mathcal{U}(y)}}{\sum\limits_{l=1}^{n} I_{|x_{l(j)} - y_j| \le \Delta_j} I_{x_l \in \mathcal{U}(y)}} \quad (i = 1, 2;\ j = 1, \ldots, p). \tag{4.33}$$

Kombinierend für beide Klassen, erhält man (4.31) als

$$I_j^2(y_j) = \sum_{i=1}^{2} \left[\mathrm{E}\Big|_{\mathcal{U}(y)} z_i - \mathrm{E}\Big|_{\mathcal{U}(y)} (z_i|x_{(j)} = y_j)\right]^2 \quad (j = 1, \ldots, p). \tag{4.34}$$

Friedman [73] empfiehlt allerdings eine Transformation der Daten, um $\pi_1 = \pi_2 = \frac{1}{2}$ (lokal in $\mathcal{U}(y)$ oder global) zu erreichen, so daß sich (4.34) zu

$$\tilde{I}_j^2(y_j) = \sum_{i=1}^{2} \left[\frac{1}{2} - \mathrm{E}\Big|_{\mathcal{U}(y)} (z_i|x_{(j)} = y_j)\right]^2 \quad (j = 1, \ldots, p) \tag{4.35}$$

reduziert. Dieses *Reinheits*-Maß ist über $\tilde{I}_j^2(y_j) = \frac{1}{2} - G_j(y_j)$ mit dem sogenannten Gini-Index verbunden (siehe Kapitel 4.4).

Mittels der so definierten lokalen Relevanz r_j^2 bietet Friedman [73] nun zwei Wege zur Definition der Nearest-Neighbours-Umgebung an. Für eine Kaskade von Anzahlen $n > M_1 > M_2 > \ldots > M' = k$ $(M_l = \alpha M_{l-1},\ \alpha \in (0, 1))$ werden am zu klassifizierenden Punkt y Regionen $\Omega = \mathcal{U}_0(y) \supset \mathcal{U}_1(y) \supset \mathcal{U}_2(y) \supset \ldots \supset \mathcal{U}'(y)$ definiert. Die sogenannte *Machete* tut dies iterativ, indem sie die Variable mit der größten lokalen Relevanz bestimmt,

$$j^*(y) = \arg\max_{1 \le j \le p} \hat{r}_j^2(y) \quad (\text{lokal in } \mathcal{U}_{l-1}(y)), \tag{4.36}$$

und den Abstand $\Delta_{j^*} := d(M_l)$ entlang dieser Variable so wählt, daß gerade M_l Punkte in der Umgebung $\mathcal{U}_l$ enthalten sind:

$$\mathcal{U}_l(y) = \Big\{x \in \mathcal{S} \cap \mathcal{U}_{l-1}(y) : \, |x_{l(j^*)} - y_{j^*}| \leq d(M_l)\Big\}. \tag{4.37}$$

Auf diese Weise wird entlang der Achsen der Variablen, die jeweils als lokal am relevantesten erachtet werden, die Umgebung um y symmetrisch beschränkt. Diese rekursive Partitionierung von Ω in betrachtete Nachbarschaft $\mathcal{U}_l(y)$ und unbeachtetes Komplement endet mit $M' = k$, so daß in der Umgebung von y genau k Objekte aus der Stichprobe enthalten sind. Die Umgebung $N_{y,k}$ hat so die Form eines achsenorientierten, p-variaten, eventuell halboffenen Quaders. Dies ist äquivalent zur einem klassischen Nearest-Neighbours-Verfahren mit $\lambda = \infty$ und $W(y) = \mathrm{diag}(\frac{1}{l_1(y)}, \ldots, \frac{1}{l_p(y)})$ in (4.28), wobei $l_j(y)$ die Länge des Intervalls der j-ten Variable $(2\Delta_j)$ in $N_{y,k}$ bezeichnet ([73]). Die Machete ist so eine *adaptive Erweiterung* des klassischen Nearest-Neighbours-Verfahrens. Für die beiden zusätzlichen Parameter schlägt FRIEDMAN [73] vor, für L Werte, die nicht in der Nähe von 1 und n liegen, und für α eher Werte, die näher zu 1 liegen, zu wählen.

Da der Machete über die Auswahl jeweils einer relevantesten Variable eine Variablenselektion inhärent ist, können ohne Probleme zusätzliche, abgeleitete Variablen[17] dem Verfahren angeboten werden (nur limitiert durch den zusätzlichen Rechenaufwand), die sogar vom zu klassifizierenden Punkt y abhängen können. FRIEDMAN [73] nennt hier drei Möglichkeiten:

- *Skalierter* EUKLID*ischer Abstand*: Mit

$$d_{\mathrm{E}} = g(x|y) = \sum_{j=1}^{p} \Big[s_j \, (x_{(j)} - y_j)\Big]^2$$

  als einziger Variable erhält man das gewöhliche Nearest-Neighbours-Verfahren. Als eine zusätzliche Variable angeboten, wird die Machete robuster in Fällen, in denen mehrere Variablen eine ähnliche Relevanz besitzen.

- *Projection Pursuit*: Durch Linearkombinationen von Variablen,

$$d_{\mathrm{PP}} = g(x|y) = \sum_{j=1}^{p} a_j(y) \, x_{(j)},$$

  etwa wenn multivariate Daten in geringerdimensionalen Hyperebenen liegen, kann man *eine* hochrelevante abgeleitete Variable erhalten. (Zum Projection Pursuit siehe [125], [81] und Kapitel 5.3.)

- *LDA*: Durch die Anwendung der LDA (Kapitel 3.2) können ebenfalls geeignete Linearkombinationen gewonnen werden.

Die *Scythe*[18] als andere Form der Definition der Umgebung $N_{y,k}$ bestimmt den Abstand in (4.37) nicht entlang der Achse der einen relevantesten Variable, sondern als Maximum der Distanzen entlang aller Variablenachsen gewichtet gemäß der lokalen Relevanz:

$$\mathcal{U}_l(y) = \{x \in \mathcal{S} \cap \mathcal{U}_{l-1}(y) : \, d_\infty(x_l, y) \leq d(M_l)\} \tag{4.38}$$

[17] d.h. Transformationen und Kombinationen der ursprünglichen Merkmale, siehe Kapitel 2.4.2

[18] zu deutsch: die Sense

mit

$$d_\infty(x_l, y) = \max_{1\le j\le p} |\omega_j(y)\,(x_{l(j)} - y_j)|. \tag{4.39}$$

Dabei gilt: $\omega_j(y) = r_j{}^\beta(y)\,(\beta \ge 0,\, j = 1,\dots,p)$. Mit $\beta = \infty$ erhält man wieder die Machete, während $\beta = 0$ das klassische Nearest-Neighbours-Verfahren mit $\lambda = \infty$ liefert. FRIEDMAN [73] empfiehlt die Wahl eher größerer β, vor allem beim Einschluß abgeleiteter Variablen.

FRIEDMAN [73] vergleicht die Machete (z.T. mit d_{E}, d_{PP}) und die Scythe für $\alpha = 0,5$ und $L = 20$ mit dem klassischen Nearest-Neighbours- und dem CART-Verfahren (Kapitel 4.4). In Simulationen, die jedes der Verfahren mindestens einmal bevorzugen, erhält er die Machete (mit d_{PP} und LDA-abgeleiteten Variablen) und die Scythe immer als optimales oder dem optimalen nachfolgendes Verfahren, so daß sie als sehr robust anzusehen sind. Verschiedene reale Datensätze liefern vergleichbare Ergebnisse. Er empfiehlt beide Verfahren in Situationen, in denen sich CART oder das klassische Nearest Neighbours schon als recht gute Klassifizierer erweisen, um diese zu verbessern. In Situationen, in denen die klassischen Verfahren versagen und die Einflüsse der Variablen eher globaler Natur sind oder Interaktionen zumindest geringer Ordnung aufweisen, rät er zu Kernschätzungen (Kapitel 4.2), der Linearen Diskriminanzanalyse (Kapitel 3.2) oder der Flexiblen Diskriminanzanalyse (Kapitel 5.2.2).

Ebenfalls FRIEDMAN [75] schlägt eine weitere Kombination von Ideen des Nearest-Neighbours-Verfahrens mit denen von CART vor, die im Kapitel 4.4.3 kurz vorgestellt wird.

Discriminant Adaptive Nearest Neighbours

HASTIE & TIBSHIRANI [118] orientieren sich an der LDA (Kapitel 3.2), um das Abstandsmaß an den zu klassifizierenden Punkt x zu adaptieren. Grundidee dieser *Discriminant Adaptive Nearest Neighbours* (DANN) genannten Methode ist, lokal in einer Umgebung um x eine lineare Trennfläche zwischen den beiden Klassen zu bestimmen und dann die Umgebung in Richtungen, die orthogonal zur Trennfläche sind, zu „stauchen“, und die, die parallel verlaufen, zu verlängern. Dadurch werden die lokalen klassenspezifischen Centroiden einander angenähert.

Die Autoren schlagen zu diesem Zweck folgende iterative Schätzung von W in (4.28) vor ($\lambda = 2$): Beginnend mit einer nichtnegativen Gewichtsmatrix W_0 (z.B. $W_0 = I_p$), wird um x über (4.28) eine Umgebung $U_L(x)$ definiert, die genau L Punkte aus $\mathcal{S}$ enthält. Nun werden in gewichteter Form die gepoolte Kovarianz[19] Σ_W beider Klassen und die Streuungssumme[20] Σ_B zwischen den Klassen in $U_L(x)$ berechnet. Diese definieren über

$$W := \Sigma_W^{-\frac{1}{2}} \left[\Sigma_W^{-\frac{1}{2}} \Sigma_B \Sigma_W^{-\frac{1}{2}} + \epsilon I_p\right] \Sigma_W^{-\frac{1}{2}} \tag{4.40}$$

eine neue Metrik ([118]). Die Iteration kann nun mit W anstelle von W_0 fortgesetzt werden. Abschließend wird W als Gewichtsmatrix in (4.28) benutzt; der weitere Verlauf entspricht dem klassischen Verfahren.

Im Detail werden Σ_B und Σ_W im Punkt x wie folgt berechnet:

[19] engl.: *within sum of squares*
[20] engl.: *between sum of squares*

$$\Sigma_B := \sum_{i=1}^{2} \frac{\sum_{x_j \in \mathcal{S} \cap \Pi_i} \omega_j}{\sum_{j=1}^{n} \omega_j} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \tag{4.41}$$

$$\Sigma_W := \sum_{i=1}^{2} \sum_{x_j \in \mathcal{S} \cap \Pi_i} \frac{\omega_j}{\sum_{j=1}^{n} \omega_j} (x_j - \bar{x}_i)(x_j - \bar{x}_i)'. \tag{4.42}$$

Dabei sind $\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$ der Mittelwert aller Beobachtungen in $\mathcal{S}$, die $\bar{x}_i$ die Mittelwerte über $\mathcal{S} \cap \Pi_i$ und die ω_j die Gewichte der einzelnen Beobachtungen in $\mathcal{S}$. Der Abstand von x zu $x_j \in \mathcal{S}$ ergibt sich über (4.28) als $d_j = \|W^{\frac{1}{2}}(x_j - x)\|$; $h := \max_{x_j \in U_L(x)} d_j$ als Bandweite ist so gewählt, daß genau L Beobachtungen in die Umgebung eingeschlossen sind. Die lokalen Gewichte berechnen sich dann als

$$\omega_j := \left[1 - \left(\frac{d_j}{h}\right)^3\right]^3 I_{\{|d_j| \leq h\}} \quad (j = 1, \ldots, n). \tag{4.43}$$

Alle Beobachtungen außerhalb von $U_L(x)$ erhalten also das Gewicht Null. Die trikubische Funktion kann durch andere relle symmetrische Funktion ersetzt werden.

Die Autoren empfehlen als „Tuning"-Parameter $L := \max(\frac{n}{5}, 50)$ und $\epsilon = 1$; außerdem genüge ein einziger Iterationsschritt. In Vergleichen mit der LDA und einem klassischen Nearest-Neighbours-Verfahren ($k = 5$) für verschiedene simulierte und reale Datensätze erhalten sie DANN meist als annähernd so gut wie das optimale Verfahren im Sinne des Risikos, in einigen Daten sogar als Optimum mit großem Abstand. Es gibt also Situationen, in denen DANN eine verbesserte Diskriminanzregel liefern kann.

4.4 CART

CART steht für *Classification and Regression Trees* und geht auf ein Buch von Breiman et al. [21] aus dem Jahre 1984 zurück. Es ist mittlerweile ein verbreitetes Klassifikationsverfahren.

4.4.1 Das Verfahren

Der Merkmalsraum wird durch Errichtung eines binären Entscheidungsbaumes (*Rekursive Partitionierung*) in Gebiete unterteilt, die jeweils einer Klasse zugeordnet werden, so daß die entsprechenden Vereinigungen dieser Gebiete zu der Zerlegung $\{\Omega_1, \Omega_2\}$ von Ω führen. Dazu werden, beginnend mit dem gesamten Merkmalsraum, Mengen in disjunkte Untermengen zerlegt. Jede der betrachteten Mengen stellt einen sogenannten *Knoten* dar; der Merkmalsraum heißt *Wurzelknoten* und wird mit t_0 bezeichnet. Knoten werden in Unterknoten durch sogenannte *Splits* zerlegt.

Definition 4.1 (Split) *Ein Split s ist eine binäre Frage, die alle Beobachtungen eines Knotens t in zwei disjunkte Untermengen (Unterknoten) t_w und t_f aufteilt:*

$$\forall x \in t: \quad s(x) = \begin{cases} \textit{wahr} & \Rightarrow \quad x \in t_w \\ \textit{falsch} & \Rightarrow \quad x \in t_f \end{cases}, \tag{4.44}$$

wobei $t_w \cap t_f = \emptyset$ und $t_w \cup t_f = t$.

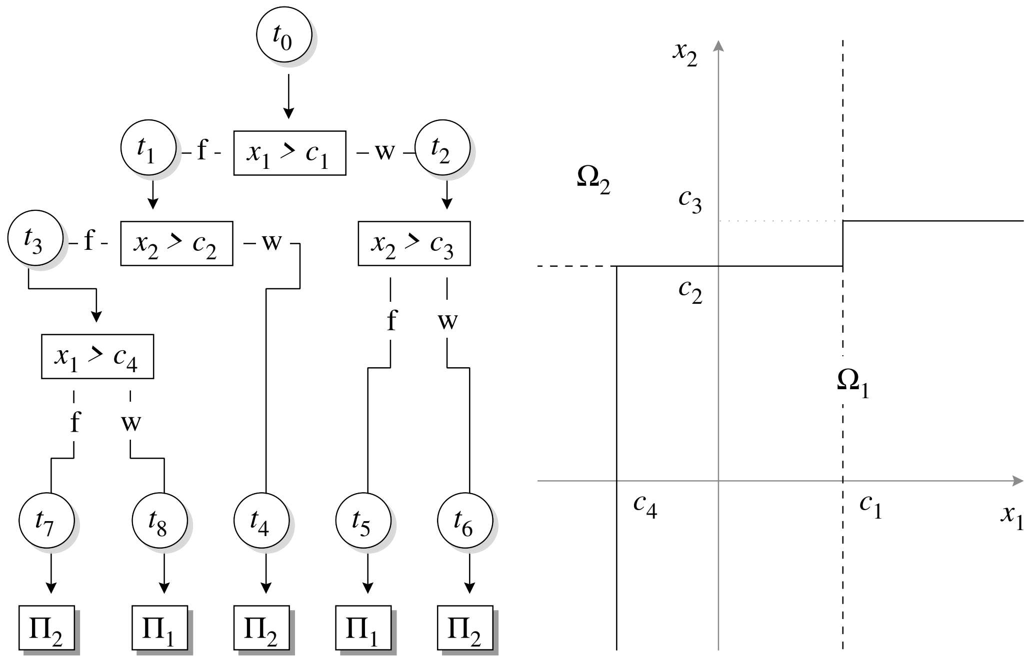

Abbildung 4.1: Beispiel für einen Entscheidungsbaum mit dem Standard-CART-Verfahren bei zweidimensionalen stetigen Daten und zwei Klassen: links der Entscheidungsbaum T mit den Knoten $T = \{t_0, \ldots, t_8\}$ ($\tilde{T} = \{t_4, \ldots, t_8\}$, $\bigcup_{t \in \tilde{T}} t = \Omega$), rechts die entsprechende Aufteilung des Merkmalgrundraumes Ω. Die Regel bewirkt folgende Zerlegung: $\Omega_1 = t_5 \cup t_8$, $\Omega_2 = t_4 \cup t_6 \cup t_7$. Die Trennfläche verläuft stückweise parallel zu den Koordinatenachsen.

Bildlich gesprochen, erzeugt ein Split auf diese Weise, von einem Knoten ausgehend, zwei Äste zu den beiden Unterknoten (siehe Abbildung 4.1). Im Standard-CART sind die Splits von folgender Form:

- *Numerische Variablen*: $\{\text{Ist } x_{(j)} \leq c \text{ ?}; c \in (-\infty, \infty), j \in \{1, \ldots, p\}\}$

  Solche Fragen bewirken Trennflächen parallel zu den Koordinatenachsen, also Teilmengen in Form (halboffener) mehrdimensionaler Quader.

- *Kategoriale Variablen*: $\{\text{Ist } x_{(j)} \in M \text{ ?}; M \subseteq \{a_1, \ldots, a_{n(j)}\}, j \in \{1, \ldots, p\}\}$

  Die Ausprägung von x im j-ten Merkmal gehört einer spezifizierten Teilmenge aller möglichen $a_{n(j)}$ Ausprägungen an.

Es sei T die Menge aller Knoten eines Baumes und $\tilde{T}$ die Menge aller *terminalen Knoten* (oder *Endknoten*), d.h. die durch keine Splits aufgeteilt wurden. Die Wahrscheinlichkeit, daß eine Beobachtung in einen Knoten t fällt, ist mit $p(t) = \int_t dP$ gegeben ($P := \pi_1 P_1 + \pi_2 P_2$); die Wahrscheinlichkeit für eine zusätzliche Zugehörigkeit zur Population Π_i mit $p(i,t) = \pi_i \int_t dP_i$.

Die Anteile der einzelnen Klassen an einem Knoten t sind definiert durch

$$p(i|t) = \frac{p(i,t)}{p(t)} \quad (i = 1, 2). \tag{4.45}$$

Offensichtlich addieren sich die Anteile zu Eins. Resubstitutionsschätzungen sind durch

$$\begin{array}{rcl} \hat{p}(i,t) & = & \pi_i \frac{\#\{x : x \in \Pi_i \cap t\}}{n} \quad (i=1,2) \\ \hat{p}(t) & = & \sum_{i=1}^{k} \hat{p}(i,t) \\ \hat{p}(i|t) & = & \frac{\hat{p}(i,t)}{\hat{p}(t)} \quad (i=1,2) \end{array} \tag{4.46}$$

gegeben ([21]). CART zerlegt so den Merkmalsraum Ω in eine endliche Anzahl von disjunkten Untermengen $\{t_c\}_c = \tilde{T}$ und modelliert dort, in der Region $t \subseteq \Omega$, lokal und konstant über die Schätzung $\hat{p}(i|t)$ direkt die A-posteriori-Dichte $\tau_i(x)$ (2.11). Eingesetzt in die BAYESsche Regel (2.10), liefert das (über das BAYESsche Theorem (2.12)) die Klassifikationsregel:

$$\hat{\delta}_{CART}(x)\Big|_{x \in t} = \begin{cases} 1: & \frac{\hat{p}(1|t)}{\hat{p}(2|t)} \geq \frac{c_{21}}{c_{12}} \\ 2: & \text{sonst} \end{cases} . \tag{4.47}$$

Die Schätzungen in der Region t erfolgen ausschließlich mit Beobachtungen aus $\mathcal{S} \cap t$. Die Diskriminanzregel ordnet jeden terminalen Knoten über (4.47) einer Population zu ($t \subseteq \Omega_i$); neu zu klassifizierende Beobachtungen fallen durch den Entscheidungsbaum in einen der Endknoten und werden der entsprechenden Klasse zugeschlagen (siehe Abbildung 4.1 zur Illustration).

Wie erfolgen nun die Splits, die zu den einzelnen Knoten führen? Der Algorithmus unterteilt sich in zwei Abschnitte: 1. die Errichtung eines Baumes mit dem Ziel möglichst „reiner“ Knoten im Sinne der Klassenzugehörigkeit und 2. das Zurückschneiden einzelner Äste, um ein Over-fitting der Regel an die Stichprobe zu vermeiden.

Aufteilung (Growing)

Über die Anteile $p(i|t)$ ist es möglich, ein *Impurity*-Maß[21] i zu definieren:

Definition 4.2 (Impurity-Maß) *Ein Maß für die Unreinheit eines Knotens t im Sinne der Klassenzugehörigkeiten seiner Elemente ist durch*

$$i(t) := \varphi\left(p(1|t), \ldots, p(k|t)\right) \tag{4.48}$$

gegeben, wobei φ eine nichtnegative Funktion ist, die auf allen k-Tupeln $(p_1, \ldots, p_k)$ mit $p_i \geq 0 \quad (i = 1, \ldots, k)$ und $\sum_{i=1}^{k} p_i = 1$ definiert ist, und folgende Eigenschaften besitzt:

(i) φ wird nur im Punkt $(\frac{1}{k}, \ldots, \frac{1}{k})$ maximal

(ii) φ wird nur in den Punkten $(1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$ minimal

(iii) φ ist symmetrisch in $p_1, \ldots, p_k$.

Die Unreinheit eines Baumes T ist dann gegeben durch

$$I(T) := \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)p(t). \tag{4.49}$$

[21] zu deutsch: Maß für die Unreinheit

Eine sehr kleine Unreinheit in einem Knoten bedeutet, daß dort die Schätzung der A-posteriori-Verteilung der einen Klasse sehr viel größer ist als die der anderen und so eine sehr sichere Zuordnung des Knotens zur ersteren Klasse erfolgen kann. Zwei Impurity-Maße werden häufig benutzt:

- *Entropie*: $i(t) := -\sum_{i=1}^{k} p(i|t) \log p(i|t)$
- GINI-*Index der Diversität*: $i(t) := -\sum_{i \neq j} p(i|t)p(j|t)$.

BREIMAN et al. [21] ziehen den GINI-Index vor.

Bei der Aufteilung eines Knotens t in zwei Unterknoten t_w, t_f durch den Split s kann sich die Unreinheit des Baumes T verringern:

$$\Delta i(s,t) \;:=\; i(t) - \frac{|t_w|}{|t|} i(t_w) - \frac{|t_f|}{|t|} i(t_f) \;\geq 0 \tag{4.50}$$

$$\Delta I(s,t) \;:=\; \Delta i(s,t)\, p(t). \tag{4.51}$$

Der Aufteilungsalgorithmus untersucht nun alle terminalen Knoten ($t \in \tilde{T}$), beginnend mit $T = \tilde{T} = \{t_0\}$, hinsichtlich ihres optimalen Splits, d.h. der Maximierung von Δi. Anschließend wird genau derjenige Endknoten in zwei neue terminale Knoten aufgeteilt, durch den die größte Verminderung der Unreinheit des Baumes ($\arg\max_{t \in \tilde{T}, s} \Delta I$) erzielt wird ([21]). Knoten werden nicht mehr für eine Aufteilung betrachtet, wenn die mögliche maximale Verringerung eine Schranke $\beta \geq 0$ unterschreitet: $\max_s \Delta I(s,t) \leq \beta$. Kann kein terminaler Knoten mehr aufgeteilt werden, bricht der Algorithmus ab. Der so erhaltene Baum heißt T_{max}.

Zurückschneiden (Pruning)

Die durch obige Aufteilung gewonnene Diskriminanzregel ist zu stark an die Stichprobe angepaßt (Over-fitting) und so von der zufälligen Auswahl $\mathcal{S}$ abhängig (erhöhte Varianz): Die Resubstitutionsschätzung $\hat{p}(i|t)$ im Impurity-Maß $I(T)$ bewirkt einen Bias in der Risikoschätzung,

$$\hat{\mathcal{R}}(T) := \sum_{t \in \tilde{T}} \min_i \left(\sum_j c_{ij} p(j|t) \right) p(t), \tag{4.52}$$

analog der Apparent Error Rate (AER, siehe Kapitel 2.2.3), durch den das BAYESsche Risiko der Regel unterschätzt wird. So fällt $\hat{\mathcal{R}}(T)$ im allgemeinen mit wachsendem $|\tilde{T}|$ und ist sogar Null, falls alle Knoten aus Objekten jeweils nur einer Klasse bestehen. Dies wird nicht durch die Schranke β verursacht, sondern ist dem Algorithmus der Aufteilung eigen ([21]).

BREIMAN et al. [21] schlagen deswegen ein „Zurückschneiden“ (engl. *Pruning*) des Entscheidungsbaumes T_{max} vor, d.h. die Rücknahme von Aufteilungen bis hin zum Zurückschneiden ganzer Unterbäume. Um die Auswahl an zu untersuchenden Bäumen zu beschränken, führen sie ein *Kosten-Komplexitäts*-Kriterium ein:

$$C_\alpha(T) = \hat{\mathcal{R}}(T) + \alpha|\tilde{T}|. \tag{4.53}$$

Dabei sind $\hat{\mathcal{R}}(T)$ die Schätzung (4.52) der AER und $\alpha \geq 0$ ein Komplexitätsparameter, der über die Anzahl der Endknoten eine zunehmende Komplexität des Baumes bestraft. Das ist ein Beispiel für die Verwendung von Straftermen als Regularisierung, durch die eine Abwägung zwischen dem Bias der Modellannahmen und einer Varianzreduktion moderiert wird (siehe Kapitel 2.3).

Für jedes α existiert ein optimaler Baum $T(\alpha)$ im Sinne von

$$T(\alpha) = \arg \min_{T \preceq T_{max}} C_\alpha(T) \tag{4.54}$$

([21]). Ist $\alpha \approx 0$, wird der resultierende Baum $T(\alpha)$ eher groß sein; für große α wird $T(\alpha) = \{t_0\}$ gelten. Da es nur endlich viele Unterbäume gibt, wird es für wachsendes α eine Zeit lang jeweils einen optimalen Baum hinsichtlich (4.53) geben. Breiman et al. [21] bestimmen diese Kaskade $T_1, T_2, \ldots$ von Bäumen über das Prinzip des *Weakest link*[22] beim Wegschneiden. Zunächst werden in T_{max} all jene Äste weggeschnitten (d.h. die Knoten werden aus dem Baum entfernt und entsprechende Aufspaltungen zurückgenommen), die die AER (4.52) konstant lassen ($\alpha = 0$):

$$T_1 := T(0) := \arg \min_{T \preceq T_{max}} \left\{ \mathcal{R}(T) = \mathcal{R}(T_{max}) \right\}.$$

Dazu werden je zwei terminale Knoten $t_w, t_f \in \tilde{T}_{max}$, die durch Teilung eines Knotens t entstanden sind ($t_w \cup t_f = t$), untersucht; falls $\mathcal{R}(t) = \mathcal{R}(t_w) + \mathcal{R}(t_f)$ gilt, werden beide durch Aufhebung der Aufspaltung weggeschnitten (d.h. aus T entfernt), und t wird ein neuer terminaler Knoten.

Es bezeichne t nun einen nichtterminalen Knoten von T_1 ($t \in T_1 \setminus \tilde{T}_1$) und T_t den Unterbaum mit t als Wurzelknoten. Es gilt: $\mathcal{R}(t) > \mathcal{R}(T_t)$ ([21], Kap. 3.3). Mit wachsendem α werden sich allerdings die Kosten-Komplexitäten beider angleichen:

$$\mathcal{R}(T_t) + \alpha|\tilde{T}_t| = c_\alpha(T_t) < C_\alpha(\{t\}) = \mathcal{R}(t) + \alpha.$$

Bei Gleichheit („=“ statt „<“) ist nun $\{t\}$ vorzuziehen, da er als einzelner Knoten kleiner als der Unterbaum T_t ist. Breiman et al. [21] definieren so eine Funktion auf T_1,

$$g_1(t) := \begin{cases} \frac{\mathcal{R}(t) - \mathcal{R}(T_t)}{|\tilde{T}_t| - 1} : & t \notin \tilde{T}_1 \\ \infty \quad : & t \in \tilde{T}_1 \end{cases},$$

die dieses α für jeden Knoten liefert, und bestimmen so den Knoten $t_1 \in T_1$, der als erster mit wachsendem α die gleiche Kosten-Komplexität wie sein Unterbaum aufweist (Weakest link):

$$t_1 := \arg \min_{t \in T_1} g_1(t).$$

Mit $\alpha_2 := g_1(t_1)$ erhält man $T_2 := T(\alpha_2) := T_1 - T_{t_1}$, d.h. T_1 wird im Unterbaum T_{t_1} auf t_1 zurückgeschnitten. Analog werden so g_2 auf T_2, t_2, T_3 und α_3 etc. definiert, bis man den Wurzelknoten als Baum erhält: $T_1, T_2, \ldots, \{t_0\}$. Dieses rekursive Pruning ist rechentechnisch sehr schnell und nimmt nur einen Bruchteil der Zeit für die Aufstellung der Regel in Anspruch ([21]).

Aus den erhaltenen Bäumen wird abschließend derjenige ausgewählt, der das Bayessche Risiko (2.4) minimiert ([21], Kap. 3.4):

$$T_0 := \arg \min_{r \in \{1,2,\ldots\}} \mathcal{R}(T_r). \tag{4.55}$$

[22] zu deutsch: gemäß der schwächsten Verbindung, schwächstes Glied der Kette

T_0 wird dann als Entscheidungsbaum in der Diskriminanzregel (4.47) benutzt. Für die Schätzung des Risikos schlagen BREIMAN et al. [21] eine Kreuzvalidierung mit 10 oder 25 Gruppen vor. Aber natürlich sind auch Bootstrap oder ein Testdatensatz anwendbar (siehe Kapitel 2.2.3).

VENABLES & RIPLEY ([218], Kap. 14.2) schlagen eine kreuzvalidierte Bestimmung von α vor. Ziel ist die Minimierung des Quotienten aus Risiko $\mathcal{R}$ und $\alpha \geq 0$. Dazu unterteilen sie die Stichprobe in 10 Gruppen, von denen jeweils 9 zum Aufstellen der Regel (mit mehreren α beim Pruning) und die 10. als Testdatensatz zur Bestimmung des Quotienten dient. Durch die Mittelung mehrfacher Wiederholungen erhalten sie eine Schätzung der Kurve $\frac{\mathcal{R}}{\alpha}$ und wählen so den Komplexitätsparameter als $\arg\min_{\alpha \geq 0} \frac{\mathcal{R}}{\alpha}$.

4.4.2 Eigenschaften

CART ist ein sehr einfach anzuwendendes Verfahren: Einmal bestimmt, müssen bei neu zu klassifizierenden Objekten nur noch wenige simple Entscheidungen getroffen werden, um sie einer Klasse zuzuordnen. Es ist daher sehr intuitiv, schnell und effizient in der Anwendung, da beinahe der gesamte Aufwand für eine Klassifizierung in der Phase der Aufstellung der Regel geleistet wurde ([75]). Darüber hinaus benötigt das Verfahren zur Anwendung nur sehr wenig Speicherplatz. Die einfache Form bedingt eine gute Interpretierbarkeit. Durch geeignete Fragen für die Splits ist CART auf sämtliche Datentypen ohne vorherige Transformation anwendbar. Es ist invariant gegenüber allen monotonen Transformationen geordneter Variablen (also auch skaleninvariant). Durch die Betrachtung jeweils nur einer Variable für eine Aufteilung entsprechend dem größten Trennvermögen findet eine automatische Variablenauswahl und Komplexitätsreduktion statt, so daß der „Fluch der Dimensionalität“ (siehe Kapitel 2.5.1) hier keine Bedeutung hat. Außerdem ist CART so extrem robust gegen Ausreißer und falschklassifizierte Objekte in der Stichprobe; daneben ist es unter milden Bedingungen BAYES-Risiko-konsistent ([21]). Die Risikoschätzung des gewählten optimalen Baumes ist durch die gleichzeitige Benutzung in der Modellwahl negativ verfälscht analog zur Apparent Error Rate (2.23) und unterschätzt so das Risiko des Gesamtverfahrens. Eine zweistufige Kreuzvalidierung wäre möglich, aber aufgrund des Aufwandes nicht unbedingt sinnvoll.

Die Variabilität der Regel ist sehr hoch: Verschiedene Stichproben können zu unterschiedlichen ersten Splits führen, so daß völlig verschiedene Bäume entstehen. Auf der anderen Seite ist das Modell sehr restriktiv: Die Trennfläche zwischen den Populationen kann nur stückweise entlang der Koordinatenachsen verlaufen. Bei komplizierteren BAYESschen Trennflächen erhöht sich entweder der Bias oder — durch die Approximation mittels vieler kleiner achsenparalleler Stücke — die Varianz der Regel, so daß sich das Risiko verschlechtert und CART in Nachteil zu anderen Klassifikationsverfahren gerät. Es kann so nur in ausgewählten Situationen das optimale Verfahren sein ([73]).

Es besteht weiterhin die Gefahr von Fehlinterpretationen der Baumstruktur, wenn etwa eine Variable durch eine andere maskiert wird[23] und so im Entscheidungsbaum nicht auftaucht oder wenn durch instabile Baumstrukturen eine Gleichsetzung von wichtigster Variable und erstem Split nicht möglich ist. Durch die Sortier- und Suchprinzipien bei der Bestimmung optimaler Splits kann das Aufstellen der Regel relativ langwierig sein ([154]).

[23] d.h. beide Variablen führen zu annähernd gleichen Risiken der Regel, jedoch ist eine der beiden immer geringfügig besser und wird deswegen für den Split benutzt

Eine besondere Eigenschaft von CART ist, daß es durch die rekursive Unterteilung und lokale Betrachtung jeweils eines Knotens bereits adaptiv ist und so bedingte Informationen gut verwerten kann.[24] Mit jeder Teilung eines Knotens verringert sich aber die Anzahl der Beobachtungen in den Unterknoten, so daß die Stichprobe mit zunehmenden Baumebenen für sinnvolle Schätzungen in den Endknoten sehr groß werden muß, um die Varianz der Regel unter Kontrolle zu halten.

4.4.3 Erweiterungen*

Um das Manko von nur achsenparallelen Trennflächen aufzuheben, betrachten Breiman et al. ([21], Kap. 5.2) bei numerischen Daten *Linearkombinationen* der Variablen, so daß eine Split-Frage die Form

$$\sum_m a_m x_{(m)} \leq c \,? \quad (c \in (-\infty, \infty))$$

besitzt. Dadurch sind beliebige Hyperebenen im Raum als Trennflächen möglich. Das wird aber durch einige schwerwiegende Nachteile erkauft: Es existiert kein natürlicher Algorithmus mehr für die Maximierung von Δi (d.h. möglicherweise nur ein Auffinden von Suboptima), Verlust der Invarianz unter monotonen Transformationen einzelner Variablen, verminderte Interpretierbarkeit und, vielleicht am schwerwiegensten, keine automatische Variablenauswahl mehr. Dafür entdeckt diese Modifikation beliebige lineare Strukturen.

Für binäre Variablen betrachten die gleichen Autoren ([21], Kap. 5.2) Boole*sche Kombinationen* der Form

$$\bigcap_m x_{(m)} \quad \text{und} \quad \bigcup_m x_{(m)},$$

wie sie häufig bei medizinischen Diagnosen oder in den Sozialwissenschaften vorkommen, und bestimmen Δi schrittweise. Bei fehlenden Daten benutzen sie sogenannte *Surrogate Splits* ([21], Kap. 5.3).

FACT

Um der möglichen Langsamkeit bei der Aufstellung der Regel abzuhelfen, haben Loh & Vanichsetakul [154] durch Kombination von CART und LDA (Kapitel 3.2) den *FACT*[25]-Algorithmus entwickelt. Für jeden Knoten berechnen sie die Kovarianzmatrix Σ_t, ermitteln deren Hauptkomponenten, um Fast-Singularitäten bei wenigen Beobachtungen in t zu vermeiden. Sie benutzen dann alle Hauptkomponenten, deren Eigenwerte nicht zu klein im Vergleich zum größten Eigenwert sind (oft ≥ 5 %), um darauf eine lineare Diskriminanzregel zu errichten, die den Knoten in zwei Unterknoten teilt. Je nach Eigenschaften der Daten werden dazu die obigen Hauptkomponenten, deren zentrierte Pendants oder sogar ihre Repräsentationen in Polarkoordinaten benutzt. Kategoriale Daten werden in binäre umgewandelt, von denen dann das kanonische Variat (siehe Kapitel 2.4.2) benutzt wird. Der Algorithmus stützt sich allein auf die Apparent Error Rate (2.23), um ein Stoppkriterium abzuleiten: Die AER verändert sich nicht mehr, oder eine einzige Klasse verfügt über eine gewisse Anzahl von Beobachtungen im Knoten. Es findet kein Pruning statt.

[24] Breiman et al. [21] beschreiben das am Beispiel der Schiffserkennung durch Radarbilder, indem bestimmte Merkmale erst relevant werden, falls das Schiff z.B. einen Turm besitzt.

[25] *Fast Algorithm for Classification Trees*

Die Autoren ([154]) erhielten in vergleichender Anwendung auf verschiedenartige Daten, daß CART oft etwas besser war, dafür FACT aber sehr viel schneller.

Breiman & Friedman [28] kritisieren in ihrem Kommentar allerdings die Umständlichkeit des Verfahrens, den Verlust von Interpretierbarkeit und Invarianz, und stellen vor allem die Fraglichkeit der Benutzung von Linearkombinationen[26] und der Bevorzugung der Schnelligkeit beim *Aufstellen* der Regel gegenüber Genauigkeit und Interpretierbarkeit in der (industriellen) Praxis heraus. Die relativ willkürliche Benutzung der Daten bzw. ihrer Transformationen sowie das alleinige Stützen auf die AER zur Risikoschätzung machen skeptisch gegenüber dem FACT-Ansatz und stehen im Widerspruch zum Geist dieser Arbeit.

DART

Friedman [75] verallgemeinert CART, indem er Ideen der Nearest-Neighbours-Methode (siehe Kapitel 4.3) benutzt: Anstatt durch einen binären Baum eine Zerlegung, d.h. disjunkte terminale Knoten, zu erzeugen, schlägt er überlappende Regionen vor. Für eine anschließende Klassifizierung soll dann diejenige Region benutzt werden, in welcher das zuzuordnende Objekt am meisten zentriert liegt. Er hofft so, Nachteile de rekursiven Partitionierung, nämlich möglichen großen Bias durch Randlage des Objekts in einem Knoten und erhöhte Varianz aufgrund der Datenfragmentation einer disjunkten Zerlegung, zu verringern.

Die Überlappung wird erreicht, indem die binäre Frage verändert wird. An die Stelle von $\{x_{(j)} \leq c\}$ im Falle stetiger Daten in (4.44) treten zwei Schwellenwerte $c_1 \leq c_2$. Eine bisherige terminale Region R wird dadurch in zwei Unterregionen R_l, R_r folgendermaßen aufgeteilt:

$$R \overset{Split}{\longrightarrow} (R_l, R_r) := \left\{ \begin{array}{lcl} x_{(j)} \leq c_2 & \Rightarrow & x \in R_l \\ x_{(j)} > c_1 & \Rightarrow & x \in R_r \end{array} \right. . \tag{4.56}$$

Objekte x, für deren j-te Komponente $x_{(j)} \in (c_1, c_2]$ gilt, sind so in *beiden* Regionen enthalten. Friedmans Algorithmus *DART* ([75]) erlaubt nun nicht nur achsenorientierte Trennflächen, sondern wählt die c_1, c_2 als δ- bzw. $(1-\delta)$-Quantil der Beobachtungen in R ($x \in R \cap \mathcal{S}$) entlang einer linearen Splitrichtung $\alpha' x$, wobei für kleinere Werte von $\delta \in (0, \frac{1}{2}]$ sich die Regionen stärker überlappen. Terminale Regionen werden nicht mehr aufgeteilt, falls sie weniger als K Objekte der Stichprobe enthalten. Diese beiden Parameter δ und K sind die Steuerungsparameter für das Verfahren; kleine Werte von δ bzw. $\frac{K}{n}$ bewirken extrem viele Schnittmengen (auch mit Regionen anderer Baumzweige), während $\delta = \frac{1}{2}$ wieder eine Partitionierung liefert (mit dem Median als Splitkriterium). Für die Bestimmung der Parameter hat Friedman [75] folgende Vorschläge, die er in die rechentechnische Umsetzung von DART implementiert hat: K soll über Cross-validation bestimmt und δ so klein gewählt werden, wie es die Rechenmöglichkeiten zulassen. Für die Bestimmung der Splitrichtung α schlägt er die Maximierung eines Kriteriums vor, das die Entfernung der Punkte mit der größten systematischen Abweichung von der lokalen Approximation der modellierten Funktion (hier: die A-posteriori-Verteilung) aus der überlappenden Region bewirkt.

Bei der Anwendung der Regel kann sich ein Zeitproblem, besonders bei vielen Schnittmengen, ergeben, da eine umfassende Suche nach der Region, in der das Objekt am meisten zentriert liegt, durchgeführt werden muß. Das läßt sich jedoch

[26] Linearkombinationen seien intuitiv einleuchtend, hielten aber in der Praxis ihr Versprechen nicht. ([28])

umgehen, indem der erhaltene Baum in einen binären Entscheidungsbaum überführt wird, der so die Schnelligkeit von CART in der Anwendung erreicht ([75]). Die relativ aufwendige Bestimmung eines DART-Baumes läßt sich vermindern, indem für ein zu klassifizierendes Objekt x nur der Baum mit den Regionen, die x enthalten, bestimmt wird (*HYESS*-Verfahren, siehe auch das LOESS-Verfahren in [42]). Das verlagert jedoch bei der Vorhersage den Aufwand aus der Trainings- in die Anwendungsphase in Analogie zur Nearest-Neighbours-Methode (Kapitel 4.3). Durch die komplexere Struktur verliert DART gegenüber CART an Interpretierbarkeit.

Weitere Ansätze

Für CART existieren auch Prädiktive Ansätze (siehe Kapitel 2.2.1). Denison et al. [47] führen eine Wahrscheinlichkeitsverteilung über dem Raum aller möglichen Bäume ein und ermitteln iterativ eine Anzahl „guter“ Bäume im Sinne des Risikos über eine stochastische Suche. Ihr Ansatz ist aber nicht komplett Bayes-gemäß, da sie einen solchen zur Zeit für undurchführbar halten. Siehe auch [38]. Bildlich kann man sich diese Ansätze als eine Mittelung über einige gute Bäume vorstellen. Siehe auch [41].

Ripley ([181], Kap. 7.5) führt weitere Split-Kriterien, wie *Incremental learning* oder *Soft splits*, auf; siehe auch [21], Kap. 4. Außerdem trägt er alternative Kriterien für das Pruning zusammen ([181], Kap. 7.2). So teilen Gelfand et al. [85] die Stichprobe in zwei Hälften auf und benutzen diese alternierend für die Aufteilung und das Pruning: Mit der ersten Hälfte erhält man eine Aufteilung, die über Fehlerschätzungen mittels der zweiten Hälfte „zurückgeschnitten“ wird. Der erhaltene Baum wird nun mit der zweiten Hälfte aufgeteilt und über die erste verkleinert. Das Verfahren bricht ab, wenn keine Veränderung des Baumes mehr erzielt wird. Gelfand et al. [85] begründen das Verfahren damit, daß beim Standard-CART durch die Benutzung der AER (2.23) der optimale Baum nicht unbedingt in $\{T(\alpha)\}$ enthalten sein muß. Quinlan [174] verwendet in seinem C4.5-Verfahren eine andere Schätzung für $\mathcal{R}(t)$, indem er sie durch das $87,5$-Quantil der Binomialverteilung $(n_t, \frac{j_t}{n_t})$ approximiert, wobei n_t die Anzahl der Beobachtungen aus $\mathcal{S} \cap t$ darstellt und j_t die davon falschklassifizierten. H. Zhang [226] wendet Klassifikationsbäume auf rein binäre Daten an.

Eine Erweiterung der Nearest-Neighbours-Methode durch Friedman [73], die Ideen des rekursiven Partitionierens aufgreift, ist in Kapitel 4.3 beschrieben. Helmbold & Schapire [122] geben einen alternativen Pruning-Algorithmus an.

4.5 Neuronale Netze

Die frühesten Vorläufer der *Neuralen* oder *Neuronalen Netze* (Neural Networks) kamen bereits in den 1940er Jahren auf und hatten ihren Ursprung in der Modellierung von Hirnfunktionen. Es gibt mittlerweile eine unüberschaubare Vielfalt an Modellen, die lose unter dem populären Namen zusammengefaßt werden. Die Neuronalen Netze sollen deswegen in dieser Arbeit nur kurz eingeführt werden, insbesondere die sogenannten *Feed-Forward-Netze*. Eine gute Einführung geben z.B. Bishop [17], Ripley [180], Michie et al. [162] und Rojas [182].

4.5.1 Feed-Forward-Netze

Grundidee der Feed-Forward-Netze ist die Annahme von Knoten in geordneten Schichten, die auf die Knoten nachfolgender Schichten wirken (aber nie umgekehrt).[27] Die oberste Schicht besteht aus den Merkmalsvariablen (bzw. ihren Transformierten) als den einzelnen Knoten und wird als *Input* bezeichnet. Die unterste Schicht heißt *Output* und kodiert mit ihren Knoten die resultierenden Antwortgrößen. Zwischen diesen äußeren Schichten liegen oft eine oder mehrere verdeckte Schichten von Knoten, die nicht beobachtet werden können (sogenannte *Hidden Layers*).

Ein gebräuchliches Feed-Forward-Netz ist das *One-Hidden-Layer* und soll hier beschrieben werden. Zunächst wirken die Input-Knoten $x_{(1)}, \ldots, x_{(p)}$ als Linearkombination auf jeden Knoten, etwa z_l, der mittleren, verdeckten Schicht über eine univariate Funktion φ_l (*Link*) ein:

$$z_l = \varphi_l \left(\alpha_l + \sum_{j=1}^{p} w_{jl} x_{(j)} \right)$$

Die $w_{jl} \geq 0$ gewichten den Einfluß der einzelnen Merkmale $x_{(j)}$ auf den Zwischenschicht-Knoten z_l: $\sum_{j=1}^{p} w_{jl} = 1$. α_l ist eine Konstante (*Intercept*). Sind einige der Gewichte gleich Null, so kann man mit der Schreibweise $j \to l$ nur die positiven ansprechen.

Analog wirken die Zwischenknoten $\{z_l\}$ auf den Output $y_1, \ldots, y_k$. Sind zusätzlich noch Wirkungen direkt vom Input zum Output unter Umgehung der Zwischenschicht erlaubt (sogenannte *Skips*), so läßt sich ein One-Hidden-Layer-Modell wie folgt schreiben:

Definition 4.3 (One-Hidden-Layer-Netz) *Die allgemeine Form des Feed-Forward-Netzes mit einer Zwischen-schicht lau-tet:*

$$y_i = \varphi_i \left(\alpha_i + \sum_{j \to i} w_{ji} x_{(j)} + \sum_{l \to i} w_{li} \varphi_l \left(\alpha_l + \sum_{j \to l} w_{jl} x_{(j)} \right) \right). \tag{4.57}$$

Unter Umständen — etwa zur Garantie der Identifizierbarkeit bei linearen oder Indikator-Funktionen — kann zusätzlich gefordert werden, daß sich die Gewichte w_{li}, w_{ji} und w_{jl} dabei jeweils zur Einheit addieren. Die Abbildung 4.2 gibt eine graphische Illustration eines Netzwerkes.

Für die Wahl der Funktionen $\varphi_.$ sind üblich:

- *Lineare Funktionen*: $\varphi(x) = ax + b$
- *Indikator-Funktionen*[28]: $\varphi(x) = I_{x > const.}$
- *Logistische Funktionen*[29]: z.B. $\varphi(x) = \frac{e^x}{1+e^x}$
- *Tangens hyperbolicus*: $\varphi(x) = \tanh x = \frac{e^x - 1}{e^x + 1}$

Kommentar 4.1 *Die logistische Funktion kann als Glättung der Indikatorfunktion $I_{x>0}$ gesehen werden und modelliert z.B. Wachstumskurven für Populationen in der Ökologie (siehe [241], Kap. 11.4.). Allerdings ist ihr Einsatz eher aufgrund ihres sigmoidalen Verlaufs motiviert.*

---

[27] Netzwerke mit Rückkopplungen werden als *symmetrische rekurrente Netzwerke*, *Attraktor-Netzwerke* oder HOPFIELD-*Netze* bezeichnet. ([180])

[28] auch: *Schwellwert*-Funktion

[29] auch: *„sanfter" Schwellenwert*, *Aktivierungsfunktion*

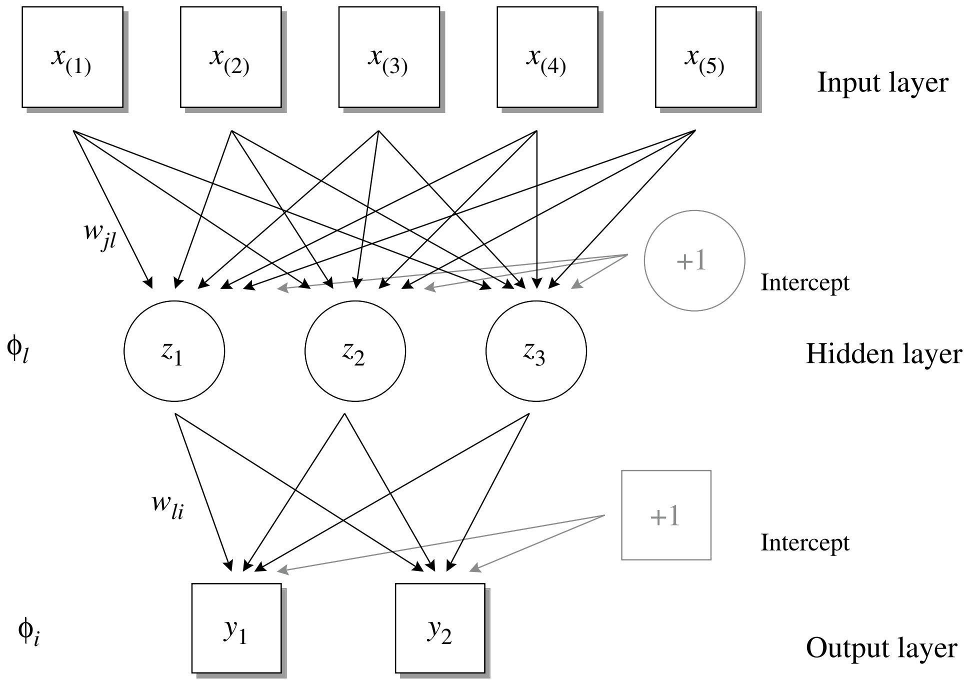

Abbildung 4.2: Beispiel für ein Feed-Forward Neuronales Netzwerk (One-Hidden-Layer) mit $k = 2$, $p = 5$ und einer Zwischenschicht ohne Skips. Die Knoten sind in Schichten angeordnet und wirken über eine Linearkombination als Argument einer Funktion auf die Knoten nachfolgender Schichten, jedoch nie umgekehrt. Nur die äußeren Schichten — Input und Output — sind beobachtbar.

Kommentar 4.2 (Perceptron) *Die Urform eines Neuronalen Netzes — das Perceptron — hat keine Zwischenschichten und nur einen Output-Knoten der Form* $\varphi(x) = sgn(ax + b)$ *([181], Kap. 3.6). Die Regionen werden durch stückweise lineare Hyperebenen getrennt; ihnen sind die Werte* $+1$ *bzw.* -1 *zugeordnet ([217], Kap. 0.1).*

Oft wird für die Knoten einer Schicht nur eine Funktion zugelassen, z.B. die logistische für die Zwischenschicht und eine lineare für den Output. Führt man im Input und der Zwischenschicht neue Knoten $x_{(0)} \equiv 1$, $z_0 \equiv 1$ mit den Gewichten $w_{0l} := \alpha_l$, $w_{0i} := \alpha_i$ ein, ergibt sich (4.57) in kompakterer Form als

$$y_i = \varphi_o \left(\sum_{j \to i} w_{ji} x_{(j)} + \sum_{l \to i} w_{li} \varphi_h \left(\sum_{j \to l} w_{jl} x_{(j)} \right) \right). \tag{4.58}$$

Das Neuronale Netz versucht so, einen angenommenen funktionellen, nichtlinearen Zusammenhang zwischen In- und Output möglichst gut zu approximieren. Cybenko [45] zeigt, daß zwei verdeckte Schichten für jede stetige Funktion genügen. Außerdem ist bekannt, daß Neuronale Netze mit linearen Output-Knoten und ursprünglich einer Zwischenschicht jede stetige Funktion auf Kompakta durch Hinzunahme zusätzlicher Zwischenschichten gleichmäßig approximieren können ([181]). Ripley ([181], Kap. 5.2) gibt als heuristischen Grund für das Funktionieren von Neuronalen Netzen die Dimensionsverringerung an, die bei mehreren Projektionen durch die Linearkombinationen der Variablen erreicht wird. Verschiedene Kombinationen repräsentieren verschiedene Hyperebenen in Ω, die unterschiedliche Variablenzusammenhänge beleuchten. Das Neuronale Netz erhält so Zugriff auf *mehrere* lineare Zusammenhänge zwischen einzelnen Variablen.[30]

4.5.2 Netzwerk-Schätzung: Back-propagation

Die Anpassung eines Netzes erfolgt iterativ, indem für Objekte mit bekanntem Output dieser mit dem Output des Netzes verglichen und die Gewichte entsprechend angepaßt werden. φ bezeichne im folgenden den k-dimensionalen Output des Netzes, $y(x)$ den bekannten Output der Objekte der Trainingsstichprobe und w den Vektor aller Gewichte. Gebräuchliche Kriterien ([180]) sind der *Quadratische Abstand*

$$E_{LS}(w) = \sum_{j=1}^{n} \|y(x_j) - \varphi(x_j; w)\|^2 \tag{4.59}$$

und der *Log-Lineare Ansatz*

$$E_{LL}(w) = \sum_{j=1}^{n} \sum_{i=1}^{k} \left[y_i(x_j) \log \frac{y_i(x_j)}{\varphi_i(x_j)} + (1 - y_i(x_j)) \log \frac{(1 - y_i(x_j))}{(1 - \varphi_i(x_j))} \right]. \tag{4.60}$$

Klassisches Mittel zur Bestimmung der Gewichte ist der *Back-propagation*-Algorithmus[31]. Als Abstiegsverfahren[32] benutzt er zur Minimierung des Kriteriums E die Ableitungen und setzt damit die Differenzierbarkeit der Funktionen $\varphi_.$ voraus, so daß er auf Indikatorfunktionen nicht anwendbar ist (was aber mit der logistischen Funktion zu umgehen ist). Der Iterationsschritt hat die Form ([180]):

$$w_{jl}^{neu} \quad \leftarrow \quad w_{jl}^{alt} \underbrace{-\eta \frac{\partial E}{\partial w_{jl}}}_{=:\Delta w_{ij}} . \tag{4.61}$$

Das Gewicht w_{ij} wird so um die konstante Schrittweite η in Richtung des steilsten Abstiegs korrigiert. Über alle Trainingsobjekte wird das Fehlerkriterium E bestimmt, die Gewichte werden adjustiert, und eine neue Iteration folgt. Das kann durch *Exponentielles Glätten*,

$$\Delta w_{ij} = -(1 - \alpha)\eta \frac{\partial E}{\partial w_{jl}} + \alpha(\Delta w_{ij})_{\text{alt}}, \tag{4.62}$$

oder durch eine Justierung nach jedem Objekt (*On-line-Algorithmus*) verallgemeinert werden ([180]).

Problem ist, wie bei allen Abstiegsverfahren, das häufige Auffinden nur lokaler Minima. Durch einen Multistart-Algorithmus mit verschiedenen Ausgangsgewichten läßt sich diese Gefahr verringern. Ripley [180] gibt als weiteres Problem eine mögliche Nichtidentifizierbarkeit der Parameter bei vollständig verbundenen Netzwerken an, da hier mehrere Mengen optimaler Parameter existieren werden.

[30] Das Finden von im gewissen Sinne „interessanten" Hyperebenen ist auch Ziel des *Projection Pursuit*, siehe z.B. [125], [81].

[31] auch: *Generalized delta rule*

[32] siehe z.B. [195], Kap. 6

4.5.3 Netzwerkkomplexität*

Wie sollte nun ein Netzwerk dimensioniert werden? Neben dem Umfang des Inputs, d.h. der Anzahl der Merkmale, sieht RIPLEY ([181], Kap. 5.6) drei „Hebel", um die Komplexität eines Netzwerkes zu kontrollieren:

- Anzahl der Hidden Layers,
- Anzahl der Links,
- Modifiziertes Fehlerkriterium.

Durch Hereinnahme genügend vieler Schichten läßt sich jede Funktion approximieren. Wegen der damit verbundenen Erhöhung der Parameteranzahl, insbesondere bei vollständig verbundenen Netzwerken, stößt dieser Ansatz durch den limitierten Stichprobenumfang schnell an seine Grenzen. Die Anzahlen der Hidden Layers und der Knoten sowie die Beträge der Gewichte spiegeln erneut die Abwägung von Bias und Varianz wider, um das Fehlerkriterium (z.B. das Risiko) unter Kontrolle zu halten. Dabei gibt es analog den Modellwahlverfahren (siehe Kapitel 2.5.1) schrittweise Verfahren, die sequentiell Netzwerke konstruieren bzw. „ausdünnen", und solche, die ein Kriterium optimieren wollen.

Links im Netzwerk sind dadurch bestimmt, daß die Gewichte einer Wirkung von Knoten zu Knoten größer als Null sind. Durch *Pruning* werden einige dieser Gewichte auf Null gesetzt, indem über schrittweise Selektion oder Cross-validation ihr Beitrag zum Fehlerkriterium getestet wird. Beispiele für solche Ansätze sind *Optimal Brain Damage* ([151]) und *Optimal Brain Surgeon* ([112]), deren Namen an die ursprünglich biologische Motivation Neuronaler Netze anknüpfen. REED [177] gibt einen teilweisen Überblick über Pruning-Algorithmen. Über modifizierte Fehlerkriterien ist ebenfalls eine Verringerung der Anzahl von Links erreichbar (siehe unten). Umgekehrt existieren auch Konstruktionsalgorithmen, die analog dem Forward-selection-Ansatz der Modellwahl eine Sequenz von Netzwerken durch Hinzunahme weiterer Knoten in die Hidden Layers oder insgesamt weiterer verdeckter Schichten generieren. Beispiele dafür sind *Sequential Network Construction* ([165]) und der *Kaskaden-Korrelations*-Algorithmus ([62]). Siehe auch [181], Kap. 5.6, für weitere Ansätze und Referenzen. Analog der Modellwahl besteht die Gefahr des Auffindens nur stark suboptimaler Netze.

Im zu minimierenden Fehlerkriterium können zusätzliche Strafterme auftauchen:

$$\tilde{E} = E + \lambda C, \tag{4.63}$$

wobei C der Strafterm ist und λ die Stärke der Strafe kontrolliert. Einige Beispiele sind in Tabelle 4.3 aufgeführt. Durch diese Form einer Regularisierung versucht man, ein Over-fitting des Netzwerkes an die Stichprobe, d.h. eine zu große Varianz, zu vermeiden. Der *Weight-decay*-Ansatz bestraft vor allem große Gewichte und wird so eher ausgeglichene Schätzungen der ω_i hervorbringen. Er ist der Ridge-Regression bei der konventionellen Kurvenapproximation analog ([17], Kap. 9.2). Die Regularisierung über *Weight elimination* wird dagegen eher wenige große Gewichte bevorzugen, während unbedeutende Gewichte im Sinne eines Beitrages zum Risiko eliminiert werden. Strafterme können als zusätzlichen Effekt ein besseres numerisches Verhalten des Fehlerkriteriums bewirken ([218], Kap. 11.4).

RIPLEY ([181], Kap. 5.6) empfiehlt für die allermeisten Fälle Regularisierungen zur Netzwerkschätzung, da dies meist zu besseren Approximationen bei festem Stichprobenumfang führt und außerdem oft auch rechentechnisch vorteilhafter ist.

Name	*Strafterm C*	*Bemerkung*
Weight decay	$\frac{1}{2}\sum\limits_i \omega_i^2$	
Weight elimination	$\sum\limits_i \frac{\omega_i^2}{\hat{\omega}^2+\omega_i^2}$	$\hat{\omega}$ ist Skalenparameter
TIKHONOV-Regularisierer	$\frac{1}{2}\sum\limits_{r=0}^{R}\int\limits_a^b h_r(x)\left(\frac{d^r y}{dx^r}\right)^2 dx$	$h_r \geq 0 \quad (r=0,\ldots,R-1)$, $h_R > 0$
Zweite Ableitung	$\int \sum_i \frac{\partial^2 y}{\partial x_i^2}\, dx$	

Tabelle 4.3: Regularisierungen bei Neuronalen Netzen. Es sind einige Beispiele für Strafterme in zu minimierenden Fehlerkriterien aufgeführt, um eine zu große Varianz des Netzwerkes zu verhindern ([181] Kap. 5.6, [17] Kap. 9.2 und 9.5.4, [180], [218] Kap. 11.4).

4.5.4 Netzwerke in der Diskriminanzanalyse

Für eine einfache Diskriminanzregel besteht die Output-Schicht aus einem einzigen Knoten, der beispielsweise den Quotienten der A-posteriori-Dichten modelliert,

$$y = \varphi(x) := \frac{\tau_1(x)}{\tau_2(x)}, \tag{4.64}$$

eine logistische Funktion davon,

$$y = \varphi(x) := \log\left(\frac{\tau_1(x)}{\tau_2(x)}\right), \tag{4.65}$$

oder auch die Differenz analog zu HALL & WAND [104]:

$$y = \varphi(x) := \tau_1(x) - \tau_2(x). \tag{4.66}$$

Für eine Diskriminanzregel werden Schätzungen für die obigen Modellierungen in die BAYESsche Regel (2.10) eingesetzt; die Zuweisung eines Objekt zur einen oder zur anderen Klasse erfolgt dann über eine Indikatorfunktion als Output-Funktion, also z.B. für (4.64) als:

$$\hat{\delta}_{NNet}(x) = \begin{cases} 1: & y \geq \frac{1}{2}\frac{c_{21}}{c_{12}} \\ 2: & \text{sonst} \end{cases} . \tag{4.67}$$

Neuronale Netze modellieren so den Quotienten oder die Differenz der A-posteriori-Verteilungen (2.11) direkt, ohne den Umweg über Klassendichten.

Für die Schätzung der Netzwerkparameter wird keine direkte Risikoschätzung benutzt, sondern andere Optimalitätskriterien, gewöhnlich der Quadratische Abstand E_{LS} ([180]). Um ein Over-fitting des Netzwerkes an die Stichprobe zu vermeiden, schlägt RIPLEY [180] den Abbruch der Iteration vor, sobald ein hoher Anteil der Objekte aus $\mathcal{S}$ korrekt klassifiziert wird, d.h. die Apparent Error Rate (2.23) einen gewissen Wert unterschreitet. Bei Testdaten schlägt er den Abbruch vor, sobald sich das Risiko wieder erhöht; er weist aber darauf hin, daß es seiner Erfahrung nach dadurch oft zu früh zum Stopp kommt. Durch wiederholte Versuche mit verschiedenen Startgewichten und den Vergleich der Risikoschätzungen via Cross-validation oder Testdatensatz sollte das Auffinden annähernd optimaler Netzwerke möglich sein.

Eine anderer Ansatz modelliert die A-posteriori-Dichten getrennt als Knoten y_1, y_2. Um Werte im Intervall $[0, 1]$ zu garantieren, wird als Dichteschätzung oft

$$\hat{\tau}_i(x) = \frac{\exp(\hat{y}_i)}{\exp(\hat{y}_1) + \exp(\hat{y}_2)} \tag{4.68}$$

benutzt. Dieser Ansatz ist als *Softmax* ([30]) bekannt. Die Regel ordnet dann ein Objekt der Klasse mit der größten, um die Kosten korrigierten, Wahrscheinlichkeit y^* zu:

$$\delta_{NNet}(x) = \arg \max_{i \in \{1,2\}} y_i^*. \tag{4.69}$$

Eigenschaften

Neuronale Netze können als Verallgemeinerung der multiplen logistischen Regression gesehen werden ([180]). Die allgemeine Form (4.57) eröffnet eine riesige Vielfalt an möglichen Modellierungen. Dadurch sind Neuronale Netze sehr flexibel und können gut „nichtparametrische" Trennflächen approximieren. Durch Hinzunahme weiterer Schichten können theoretisch beliebige Funktionen approximiert werden. Neue Schichten oder mehr Knoten pro Schicht bedeuten aber immer auch mehr zu schätzende Parameter. Die Konstruktion eines Netzwerkes — die Anzahl der Schichten, Knoten und Links — stellt so eine Abwägung zwischen Bias und Varianz dar (siehe Kapitel 2.3).[33]

Das liefert die Kehrseite: Für Neuronale Netzwerke müssen eine Vielzahl von Entscheidungen getroffen werden, etwa die Wahl der Anzahl der Schichten, der Anzahl der Knoten in den einzelnen Schichten, die Art der Funktionen $\varphi_.$ etc.[34] Dieses weite Feld läßt eine automatisierte Bestimmung der Diskriminanzregel nur begrenzt zu. Neuronale Netzwerke besitzen außerdem eine hohe Variabilität ([23]), d.h. je nach Stichprobe können die geschätzten Netzwerke eine unterschiedliche Gestalt (in Form der Gewichte) annehmen. Während kleine Netzwerke noch interpretierbar sein können, sind größere dafür viel zu komplex, so daß sie als „Black box" wirken.

Neuronale Netzwerke sind relativ langsam in der Bestimmung der Regel. Durch Bildung mehrerer Linearkombinationen von Ausgangsvariablen findet eine Dimensionsreduktion statt. Weitere Aussagen allgemeiner Art sind aufgrund der Vielfalt kaum zu machen. Es gibt Beispiele, daß Neuronale Netze die beste Klassifikationsregel liefern, wobei sie vorher in ihrer Struktur „per Hand" auf das Problem angepaßt wurden.

[33] Zu starke Varianz aufgrund zu komplex gewählter Strukturen wird in der Praxis manchmal so beschrieben, daß das Netz sich die Beobachtungen „merkt": Es macht keine Fehler bei Anwendung auf die Trainingsstichprobe (die Apparent Error Rate (2.23) ist klein), neue Beobachtungen werden aber zu erheblichen Teilen falsch klassifiziert (das Risiko (2.4) ist groß).

[34] Vapnik kommentiert das folgendermaßen: „The designers of neural networks compensate the mathematical shortcomings with the high art of engineering." ([217], S. 157)

Kapitel 5

Neuere Ideen

Dieses Kapitel stellt weitere nichtparametrische Verfahren in kurzer Form vor, die in den 1990er Jahren für die Diskriminanzanalyse entwickelt oder auf sie angewandt worden sind. Allen ist der massive Einsatz von Rechnern gemein, der Teil der Verfahren ist und sie praktisch erst ermöglicht.

5.1 Resampling-Verfahren zur Varianzreduktion

Aufbauend auf den Erkenntnissen von FRIEDMAN [74] (siehe Kapitel 2.3) u.a.[1] wurden in jüngster Zeit Verfahren konstruiert, die sich um eine vornehmliche Reduktion der Varianz einer Klassifikationsregel bemühen. Zwei dieser *P & C*-Verfahren[2] sollen im folgenden vorgestellt werden.

5.1.1 Bagging

BREIMAN [23] schlägt vor, aus der Trainingsstichprobe $\mathcal{S}$ eine Stichprobe $\mathcal{S}^{(B)}$ vom Umfang $n = |\mathcal{S}|$ mit Zurücklegen zu ziehen, wobei alle Beobachtungen in $\mathcal{S}$ dieselbe Wahrscheinlichkeit $\frac{1}{n}$ besitzen. Aufgrund von $\mathcal{S}^{(B)}$ wird nun eine Diskriminanzregel $\delta(\mathcal{S}^{(B)})$ über eine sogenannte *Basisregel*, z.B. CART, bestimmt. Dieser Vorgang wird L-mal wiederholt; abschließend wird ein Objekt x der Klasse zugeordnet, für die sich die Mehrzahl der bestimmten Regeln „ausspricht“ (*Voting*):

$$\delta_A(x) = \arg \max_{i \in \{1,2\}} \sum_{l=1}^{L} I_{\delta(\mathcal{S}_l^{(B)};\, x)=i} \,. \tag{5.1}$$

δ_A ist so eine zusammengesetzte Regel, in der jede Einzelregel $\delta(\mathcal{S}_l^{(B)})$ dasselbe Gewicht beim Voting besitzt. Der Erhalt der Stichproben über Ziehen mit Zurücklegen aus der Ausgangsstichprobe (*Bootstrap*) ist analog zur Fehlerschätzung in Kapitel 2.2.3 und war namensgebend: **b***ootstrap* **agg***regation*.

Grundidee des Bagging ist eine Varianzreduktion: BREIMAN [26] zerlegt den Fehler einer Diskriminanzregel in BAYESschen Anteil (nicht verringerbar), Bias und Varianz. Ein Verfahren wie CART (Kapitel 4.4) ist restriktiv in seiner Struktur und besitzt so oft eine hohe Varianz (Instabilität) beim Versuch der Approximation komplizierter Trennflächen, während die LDA oder Nearest Neighbours (Kapitel 3.2 und 4.3) sehr stabil sind (geringe Varianz) bei oftmals großem Bias ([23]). Durch die Aggregation beim Bagging wird nun die Varianz reduziert, während der Bias

[1] siehe z.B. [129], [26], [53]

[2] Perturb & Combine ([26]), zu deutsch: stören und kombinieren

relativ unverändert bleibt, so daß sich eine Verringerung des Risikos ergibt ([26]). Damit macht Bagging nur Sinn bei Verfahren mit hoher Varianz und niedrigem Bias als Basisregel; BREIMAN [26] zeigt dieses eindrucksvoll an einigen synthetischen und realen Datensätzen. Die LDA zum Beispiel wird *nicht* verbessert. Bei CART bewirkt die Aggregation immer noch stückweise achsenparallele Trennflächen, die aber ohne Erhöhung der Varianz die ideale Trennfläche besser approximieren. Allerdings geht das auf Kosten der Interpretierbarkeit.

Bagging ist einfach in seiner Struktur und damit im Verständnis seiner Funktionsweise; außerdem ist es hoch parallelisierbar. Durch den Bootstrap-Ansatz ist das wiederholte Stichprobenziehen wesentlich für das Verfahren. BREIMAN [23] erhält eine Verringerung des Risikos bei CART als Basisregel von 6 bis zu 77 %, im Schnitt von 20-30 %. Für L empfiehlt er Werte zwischen 25 und 50.

5.1.2 Arcing & Boosting

FREUND & SCHAPIRE [67] schlagen ein äußerlich ähnliches Verfahren vor. Im Gegensatz zu Bagging ist es aber serieller Natur, da die Wahrscheinlichkeiten der Beobachtungen in $\mathcal{S}$ mit jeder Stichprobe neu bestimmt werden. Außerdem unterscheiden sich die Gewichte der Einzelregeln in der zusammengesetzten Regel:

Beginnend mit einer Ausgangsverteilung der Beobachtungen, $p^{(1)}(x_j) = \frac{1}{n}\,(j = 1, \ldots, n)$, wird aus $\mathcal{S}$ eine Stichprobe $\mathcal{S}^{(1)}$ vom Umfang n mit Zurücklegen gezogen. Mit der durch $\mathcal{S}^{(1)}$ gewonnenen Regel $\delta(\mathcal{S}^{(1)})$ werden sämtliche Beobachtungen aus $\mathcal{S}$ klassifiziert; über die Anzahl der Fehlklassifikationen lassen sich dann

$$\epsilon_1 := \sum_{x_j \in \mathcal{S}} p^{(1)}(x_j) d(x_j) \tag{5.2}$$

$$\beta_1 := \frac{1 - \epsilon_1}{\epsilon_1} \tag{5.3}$$

definieren, wobei $d(x_j) = I_{\delta(\mathcal{S}^{(1)};\, x_j) \neq z_j}$ die Falschklassifizierung und z_j die Klassenzugehörigkeit von x_j kodieren. β_1 ist so ein Indikator für die Güte der Regel: Für große ϵ_1, d.h. häufiges Falschklassifizieren von Beobachtungen mit eher größerer Wahrscheinlichkeit $p^{(1)}$, geht β_1 gegen Null, während es bei wenigen und eher „unbedeutenden“ Falschklassifizierungen, die beim Voting wahrscheinlich überstimmt werden, groß wird. Die neue Verteilung auf $\mathcal{S}$ für das Stichprobenziehen ergibt sich dann als

$$p^{(2)}(x_j) := \frac{p^{(1)}(x_j)\beta_1^{d(x_j)}}{\sum\limits_{x_j \in \mathcal{S}} p^{(1)}(x_j)\beta_1^{d(x_j)}} \quad (j = 1, \ldots, n). \tag{5.4}$$

Gemäß $p^{(2)}$ wird nun eine weitere Stichprobe $\mathcal{S}^{(2)}$ aus $\mathcal{S}$ gezogen usw. Abschließend werden die Regeln gewichtet kombiniert, d.h. je nach „Schwere“ der Fehlklassifikationen haben die Einzelregeln ein unterschiedliches Stimmgewicht:

$$\delta(x) = \arg \max_{i \in \{1,2\}} \sum_{l=1}^{L} \log \beta_l \, I_{\delta(\mathcal{S}^{(l)};\, x) = i}. \tag{5.5}$$

Regeln, die nur wenig Fehlklassifizierungen aufweisen oder nur Objekte von nur geringer Wahrscheinlichkeit falsch zuordnen, erhalten so ein größeres Gewicht bei der „Abstimmung“.

Breiman [26] verändert zwei Details: Für $\epsilon_l > \frac{1}{2}$ wird β_l negativ, für $\epsilon_l = 0$ ist es nicht definiert; er schlägt in diesen Situationen einen Neustart mit einer Gleichverteilung der Beobachtungen auf $\mathcal{S}$ vor und erhält damit gute Resultate.

Freund & Schapire [68] nennen ihr Verfahren *AdaBoost* als eine spezielle Form des *Boosting*[3]; Breiman [26] benennt es *arc-fs* als besondere Form des *Arcing*[4]. Er erhält in der Anwendung auf verschiedene Daten AdaBoost als gleichwertiges, oft sogar besseres Verfahren im Vergleich zu Bagging. Das deckt sich mit Ergebnissen von Bauer & Kohavi [11], Freund & Schapire [67], Dietterich [50] und Quinlan [175]. Außerdem wurde eine Bias-Reduktion bei AdaBoost gefunden. Allerdings wurden mit dem C4.5-Verfahren ([174]) als Basisregel sowohl mit Bagging als auch mit AdaBoost in seltenen Fällen Verschlechterungen des Risikos registriert.

Das Verständnis des Verfahrens ist etwas schwieriger als bei Bagging. Da Arcing sequentiell ist, kann es nicht als eine Aggregation unabhängiger Regeln gesehen werden. Die intuitive Idee für das Funktionieren von Arcing ist die Betonung von „unsicheren Fällen“, die oftmals falsch klassifiziert werden, im Sinne einer häufigen Mitgliedschaft in $\mathcal{S}^{(l)}$, so daß Arcing als gewichtetes Bootstrap verstanden werden kann. Breiman [26] konstruiert einen alternativen Algorithmus, *arc-x4*, der eine andere Gewichtung durchführt und trotzdem das gleiche Risiko besitzt. Er zeigt so, daß die *adaptive* Gewichtung wesentlich ist. Er schlägt deswegen für CART eine einfache Neugewichtung sämtlicher Beobachtungen in $\mathcal{S}$ anstatt wiederholter Stichproben vor und erhält vergleichbare Ergebnisse. Über Simulationen zeigt er außerdem eine nur verrauschte Beziehung zwischen der Anzahl der Falschklassifizierungen eines Objekts x_j und seiner Wahrscheinlichkeit $p^{(l)}(x_j)$. Schapire et al. [189] leiten von der Apparent Error Rate (2.23) nichtasymptotische Schranken für das Risiko her und vermuten einen Zusammenhang zwischen einer von ihnen definierten Größe (*Margin*) und dem Risiko. Das wird allerdings von Breiman [25] widerlegt.

Friedman et al. [79] erklären Boosting als Additives Logistisches Regressionsmodell und geben eine technische Modifikation an. Schapire [187] gibt einen kurzen Überblick über aktuelle theoretische Arbeiten (1999). Friedman [78] und Breiman [27] kombinieren Bagging und Boosting. Freund et al. [69] entwickeln *RankBoost* als Variante eines Boosting-Verfahrens für die Anwendung bei Meta-Suchen, etwa von Seiten im World Wide Web.

Kong & Dietterich [135] betrachten die obigen Methoden als *homogene Voting*-Verfahren, weil stets dieselbe Basisregel benutzt wird. Dem stellen sie Verfahren mit *nichthomogenem Voting* entgegen, die die Ergebnisse unterschiedlicher Basisregeln, also z.B. CART und Nearest Neighbours, miteinander kombinieren und unter Umständen eine Reduktion von Bias und Varianz erlauben (siehe [135] für Referenzen solcher Verfahren).

5.2 Nichtparametrische Regression

5.2.1 Verallgemeinerungen linearer Modelle

In neueren Arbeiten wird die Diskriminanzanalyse als Regressionsproblem gesehen, wobei die abhängige Variable Y als Ausprägungen die Klassenzugehörigkeiten $\{1, 2\}$

[3] Der Begriff des *Boosting* stammt aus der Machine-Learning-Theorie (siehe z.B. [67], [187]). AdaBoost wurde ursprünglich dafür konstruiert, die Apparent Error Rate (2.23) iterativ schnellstmöglich auf Null zu bringen ([26]).

[4] **a***daptive* **r***esampling and* **c***ombining*

besitzt und damit kategorial ist. Eine allgemeine Form besitzt das Regressionsmodell in

$$Y = f(X) + \epsilon, \tag{5.6}$$

mit ϵ einer Wahrscheinlichkeitsverteilung folgend, wobei aus Identifizierbarkeitsgründen $\mathrm{E}[\epsilon] = 0$ gilt, und $X = (X_{(1)}, \ldots, X_{(p)})'$ als Vektor der Argumente. Die Klassenzugehörigkeit wird so im Mittel durch eine Funktion der Merkmale beschrieben: $\mathrm{E}Y|X = f(X)$. Die bekannteste Form von (5.6) ist das *Lineare Regressionsmodell*:

$$f(X) = a_0 + \sum_{j=1}^{p} a_j X_{(j)}. \tag{5.7}$$

Es besitzt Linearität und Additivität in den Variablen als grundlegende Eigenschaften. Um komplexere Funktionen f approximieren zu können und trotzdem zumindest eine der Eigenschaften zu erhalten, wurden verschiedene Verallgemeinerungen entwickelt.

Verallgemeinerte Lineare Modelle (GLM) lassen die abhängige Variable Y, von der eine zur Exponentialfamilie gehörende Verteilung angenommen wird, nur indirekt über eine sogenannte *Link*-Funktion G von der Linearkombination der $X_{(j)}$ abhängen:

$$G(Y) = a_0 + \sum_{j=1}^{p} a_j X_{(j)} + \epsilon \tag{5.8}$$

([91], Kap. 5.2). Die Modellierung (3.26) des Dichtequotienten in der Logistischen Diskriminanzanalyse (Kapitel 3.4) und die Neuronalen Netze (Kapitel 4.5) entspringen diesem Kontext. Die lineare Beziehung und mit ihr die Methoden der Schätzung bleiben so erhalten. Der andere Weg *Additiver Modelle* (AM) betont die Additivität und sieht die $X_{(j)}$ auf Y über eine Linearkombination univariater, glatter Funktionen g_j wirken:

$$Y = g_0 + \sum_{j=1}^{p} g_j(X_{(j)}) + \epsilon, \tag{5.9}$$

mit $\mathrm{E}\left[g_j(X_{(j)})\right] = 0\,(j = 1, \ldots, p)$ als zusätzlicher Bedingung ([190]). Beide Ansätze können über die *Verallgemeinerten Additiven Modelle* (GAM) zusammengeführt werden (Y wieder einer Verteilung der Exponentialfamilie unterliegend, [190]):

$$G(Y) = g_0 + \sum_{j=1}^{p} g_j(X_{(j)}) + \epsilon. \tag{5.10}$$

Beispiel 5.1 (Diagnostik) *In der Medizin werden selten alle Symptome für eine Krankheit beobachtet; oft ist aber ein „je mehr, desto wahrscheinlicher“ gegeben. So ist etwa eine Risikoformel für kardiovaskuläre Endpunkte aus Daten der Framingham-Herz-Studie ([228]) über die Verteilung einer Größe modelliert worden, die über ein Verallgemeinertes Additives Modell aus Geschlecht, Alter und Phänotypen dargestellt wird (ähnlich die PROCAM-Formel, siehe [229]). Auch viele grobe Regeln, etwa für die Bestimmung von Obergrenzen für Blutfett- und Blutdruckwerte ([244]), summieren „Risikofaktoren“, die eventuell vorher binärisiert wurden.*

Für das Lineare Modell existieren die bekannten expliziten Schätzungen (siehe [126]). Die Verallgemeinerungen werden oft über eine iterative Bestimmung von Maximum-Likelihood-Schätzungen (*Backfitting*[5]) gelöst. GREEN & SILVERMAN ([91]) bieten einen Überblick über die GLM, während SCHIMEK & TURLACH ([190]) und HASTIE & TIBSHIRANI ([115]) dies für die GAM tun. Letztere führen auch sogenannte *Scatterplot Smoother* als Schätzer ein und geben Erweiterungen wie z.B. *MARS*[6] von FRIEDMAN [71] an. Siehe auch [196], Kap. 8.

Die Besonderheit der Anwendung solcher Modelle auf die Diskriminanzanalyse ist die Art des Fehlerkriteriums. Der Regressionskontext mißt den Approximationsfehler oft in Größen wie dem MSE (vgl. Tabelle 4.2 auf S. 51),

$$\mathrm{E}\left[Y - \hat{f}(X)\right]^2, \tag{5.11}$$

oder verwandten Kriterien. Diese werden manchmal noch um Strafterme erweitert (*Penalization*), um durch diese Regularisierung glattere Schätzungen zu erhalten ([72]) und so die Bias-Varianz-Abwägung zu moderieren (siehe Kapitel 2.3). In der Diskriminanzanalyse wird statt dieses L_2-Abstandes ein 0-1-Verlust in einer Beobachtung betrachtet; d.h. der quantitative Verlust wird durch einen qualitativen ersetzt. Es ist so nicht irgendein Abstand zur optimalen Trennfläche oder ähnliches entscheidend, sondern daß man auf der „richtigen“ Seite liegt. Ein Ad-hoc-Maß für Fehler ist beispielsweise die Anzahl der Falschklassifizierungen:

$$\frac{1}{n}\sum_{j=1}^{n} I_{z_j \neq \hat{\delta}(x_j)}. \tag{5.12}$$

Dies läßt sich natürlich wieder um Populationsanteile und Kosten korrigieren (siehe [72], Kap. 6). Auch eine Binärisierung von y in k Variablen $y_1, \ldots, y_k$, die über $y_i = 1 \iff x \in \Pi_i$ (und 0 sonst) die Klassenzugehörigkeit kodieren, ist denkbar. Die $y_i = \frac{f_i(x)}{\sum_l f_l(x)} + \epsilon$ modellieren dann die A-posteriori-Verteilungen (2.11) in einem multivariaten Regressionskontext ([72]).

5.2.2 Modifizierte klassische Verfahren

HASTIE et al. [120] zeigen die Äquivalenz der Linearen Diskriminanzanalyse (LDA, Kapitel 3.2) mit einer multivariaten linearen Regression unter Benutzung Optimaler Scores (Kapitel 2.4.2) für die Kodierung der Klassenzugehörigkeiten. Ihr Ansatz der *Flexiblen Diskriminanzanalyse* (FDA) ersetzt nun die lineare Regression durch nichtlineare nichtparametrische Verfahren, insbesondere durch MARS ([71]) und BRUTO ([113]). Die Autoren möchten mit den dadurch erhaltenen nichtlinearen Trennflächen die Schwäche der LDA bei nichtlinearen und komplexen BAYESschen Trennflächen beheben. Bei Anwendung auf synthetische und reale Daten erhielten sie teils kleine, teils dramatische Verringerungen des Risikos.

HASTIE & TIBSHIRANI [117] verallgemeinern die LDA auf einem anderen Wege, indem sie nichtnormale Verteilungen durch eine Mischung von Normalverteilungen approximieren: *Mixture Discriminant Analysis* (MDA)[7]. Sie modellieren die Klassendichten f_i als Summe von R_i Normalverteilungen mit derselben Kovarianzmatrix Σ, aber unterschiedlichen Anteilen π_{ij} an der Summe. Die Schätzung der Parameter

[5] iterativer GAUSS-SEIDEL-Algorithmus

[6] *Multivariate Adaptive Regression Splines*

[7] Die Autoren geben die MDA als eine geglättete Version des *Learning-Vector-Quantization*-Verfahrens aus der Pattern-Recognition-Literatur an.

erfolgt über einen verallgemeinerten EM-Algorithmus ([150]) zur Maximierung einer klassenspezifischen Likelihood-Funktion. Über das BAYESsche Theorem (2.12) sind die A-posteriori-Verteilungen τ_i schätzbar; die Klassifikationsregel (2.6) ordnet dann x der Klasse mit $\arg\max_i \hat{\tau}_i(x)$ zu. Auch eine Bestimmung über Optimale Scores und weitere Verfeinerungen sind möglich (siehe [117]).

Über die Verallgemeinerten Additiven Modelle (5.10) läßt sich auch die Modellierung des Logarithmus des Klassendichtenquotienten, etwa in der Logistischen Diskriminanzanalyse (Kapitel 3.4), verallgemeinern:

$$\log\left(\frac{f_1(x)}{f_2(x)}\right) = \sum_{j=1}^{p} g_j\left(x_{(j)}\right). \tag{5.13}$$

Das modifizierte Verfahren kann damit ebenso wie die LDA oder MDA nichtlineare Trennflächen beschreiben und ist aufgrund der angewandten nichtparametrischen Verfahren, wie z.B. MARS, lokal adaptiv.

5.3 Projection Pursuit

Grundidee des *Projection Pursuit*[8] (PP) ist das Auffinden „interessanter" oder in einem gewissen Sinne optimaler Projektionen der Variablen, die dann über univariate Funktionen auf eine abhängige Variable wirken. Im Regressionskontext werden diese Funktionen additiv verbunden ([115], Kap. 4.2):

$$Y = \sum_{l=1}^{L} g_l(\alpha_l' X) + \epsilon. \tag{5.14}$$

HASTIE & TIBSHIRANI ([115], S. 85) nennen das „eine direkte Attacke auf die Dimensionalitätsfrage". In der Tat treten bei der Schätzung oder Glättung der g_l keine Dimensionalitätsprobleme auf. Für die Dichteschätzung und die Diskriminanzanalyse ist eine multiplikative Verknüpfung sinnvoll ([125], Kap. 11):

$$f_L(x) = f_0(x) \prod_{l=1}^{L} g_l(\alpha_l' x), \tag{5.15}$$

wobei f_0 eine Initialdichte, oft die Normalverteilung, darstellt. Diese Form stammt von FRIEDMAN et al. [81]. Die Dichte f_L in x wird so als Produkt verschiedener Randdichten in den Richtungen $\alpha_l' x$ approximiert, die iterativ bestimmt werden. Die Optimalität wird im Sinne eines Index erzielt, meist von der Form ([127]):

$$I(f) = \int J\left(f(z)\right) f(z)\, dz = \mathrm{E}_f\left[J(f)\right]. \tag{5.16}$$

Dieser muß maximiert bzw. minimiert werden. FRIEDMAN et al. [81] etwa benutzen die KULLBACK-LEIBLER-*Distanz* mit $J(f) = \log(f_L)$ für die Dichteschätzung. Durch die Projektionen aus Ω in eindimensionale Räume über die Bildung von Linearkombinationen der Variablen umgeht der Projection Pursuit den „Fluch der Dimensionalität" (siehe Kapitel 2.5.1). Er hat Zugriff auf mehrere lineare Zusammenhänge; sind diese nicht gegeben, erhöht dies den Bias.

Für einen allgemeinen Überblick über den Projection Pursuit siehe HUBER [125]. Speziell für den Regressionskontext siehe KLINKE & GRASSMANN ([134]).

[8] zu deutsch: Projektive Weiterverfolgung

POLZEHL [170] wendet den Ansatz (5.15) auf die Diskriminanzanalyse an. Um der anderen Natur des Fehlers in der Diskriminanzanalyse Rechnung zu tragen, ersetzt er aber (5.16) durch eine Schätzung des BAYESschen Risikos (2.4). Dabei werden die Richtungen α_l so iterativ gesucht, daß die geglättete Risikoschätzung für die Regel, die aus den Schätzungen der Klassendichten resultiert, minimiert wird (siehe [81] und [170] für Details). POLZEHL [170] sieht Platz für die Diskriminanzanalyse via Projection Pursuit vor allem in Situationen, in denen parametrische Verfahren nicht flexibel genug und die Stichprobenumfänge zu klein für vollständig nichtparametrische Methoden sind.

5.4 Support Vector Machines

Vom Gebiet des *Machine Learning* kommend, schlägt VAPNIK [217] die sogenannte *Support Vector Machine* (SVM) vor. Der p-dimensionale Merkmalsvektor aus Ω (*input vector*) wird zunächst durch nichtlineare Abbildungen in einen hochdimensionalen Raum $\mathcal{Y}$ abgebildet (*feature vector*), z.B.:

$$\Omega \ni x = \begin{pmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(p)} \end{pmatrix} \longmapsto \begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(p)} \\ x_{(1)}^2 \\ \vdots \\ x_{(p)}^2 \\ x_{(1)}x_{(2)} \\ \vdots \\ x_{(p-1)}x_{(p)} \end{pmatrix} = y \in \mathcal{Y}$$

([217], Kap. 5.6). Dabei wird die Abbildung a-priori gewählt. In $\mathcal{Y}$ sucht die SVM nun nach lokal linearen Trennflächen, die beide Klassen in einem gewissen Sinne optimal separieren. Die lineare Fläche läßt sich als

$$w'y + b = 0 \tag{5.17}$$

beschreiben. Kodiert man die Klassenzugehörigkeit über

$$z^*(x) := \begin{cases} 1 : x \in \Pi_1 \\ -1 : x \in \Pi_2 \end{cases} \tag{5.18}$$

und nimmt linear separierbare Klassen (in $\mathcal{Y} \cap \mathcal{S}$) an[9], erhält man als kompakte Beschreibung einer Trennfläche:

$$z_i^* \left[w'y_i + b \right] \geq 1 \quad (i = 1, \ldots, n). \tag{5.19}$$

Die *optimale separierende Hyperebene* ist als diejenige definiert, die den minimalen Abstand eines Punktes aus $\mathcal{Y} \cap \mathcal{S}$ zur Trennfläche maximiert ([217], Kap. 5.4). Für die Errichtung einer Hyperebene sind nun nur diejenigen Punkte wichtig, die Gleichheit erreichen im Ausdruck (5.19), d.h. die auf dem Rand des Korridors um diese Hyperebene liegen (Normalisierung der Parameter der Hyperebene). VAPNIK

[9] VAPNIK ([217], Kap. 5.5) gibt auch eine Verallgemeinerung für nicht linear separierbare Klassen an, die die Anzahl der Fehlklassifikationen minimiert.

([217], Kap. 5.5) konstruiert aus diesen *Support Vectors* in einer lokalen Umgebung in $\mathcal{Y}$ die optimale Hyperebene als Linearkombination. Die Bestimmung der Support Vectors erfolgt über die Maximierung einer quadratischen Form. Für die Diskrimination eines neuen Objekts wird dessen Merkmalsvektor ebenfalls in $\mathcal{Y}$ abgebildet, der Funktionswert der lokalen Trennfläche in (5.17) bestimmt und je nach Vorzeichen Π_1 bzw. Π_2 zugeordnet. Die SVM modelliert so implizit die A-posteriori-Verteilungen (2.11) ohne Betrachtung von Klassendichten.

Mit zunehmender Dimension von $\mathcal{Y}$ steigt die Anzahl der Support Vectors weit weniger stark. Vapnik ([217], Kap. 5.7) demonstriert das bei einem Beispiel mit $p = 256$ über den Grad eingeschlossener Polynome bei der Abbildung $\Omega \to \mathcal{Y}$: Für eingeschlossene Polynome 1. Grades ($\dim\mathcal{Y} = 256$) erhielt er durchschnittlich 282 Support Vectors, die die lineare Trennfläche bestimmten, für Polynome bis zum 7. Grad ($\dim\mathcal{Y} \approx 10^{16}$) waren es 422. Als interessante analytische Eigenschaft gibt Vapnik ([217], Kap. 5.6) an, daß die durchschnittliche Wahrscheinlichkeit einer Fehlklassifikation durch

$$\frac{\mathrm{E}[\text{Anzahl der Support Vectors}]}{|\mathcal{S}| - 1}$$

als oberer Schranke gebunden ist. Das unterfüttert theoretisch die ausschließliche Benutzung der Support Vectors zur Konstruktion der Regel.

Vapnik ([217], Kap. 5.12) erklärt für die SVM ein neues Dogma: Bisherige multidimensionale Modellierungen von funktionalen Zusammenhängen basieren auf der Annahme, daß reale Probleme durch eine geringe Anzahl „starker“ Variablen gut approximiert werden können. Für die SVM erklärt er als Glaubensgrundlage, daß eine Vielzahl „schwacher“ Variablen existiert und die Probleme durch geeignete oder schlaue[10] Linearkombinationen dieser gut angenähert werden können. Die SVM erlaubt den Einschluß vieler abgeleiteter Variablen, um so vermutete nichtlineare Zusammenhänge „faßbar“ zu machen. Je nach Art der Abbildung $\Omega \to \mathcal{Y}$ kann die Trennfläche in Ω stark nichtlineare Formen annehmen. Durch das Konzept der Support Vectors findet eine teilweise Variablenauswahl statt. Für weitere Details siehe [217].

[10] engl. *smart*

Kapitel 6

Der Fall mehrerer Klassen

Diese Arbeit beschränkt sich bei der Vorstellung der Verfahren zumeist auf zwei Klassen. Im folgenden soll kurz angerissen werden, welche Möglichkeiten es gibt, die Verfahren auf Probleme mit mehr als zwei Klassen ($k \geq 3$) anzuwenden.

6.1 Die allgemeine Bayessche Regel

Prinzipiell erlauben viele Verfahren die Erweiterung auf mehrere Klassen: Wenn die klassenspezifischen Dichten f_i oder A-posteriori-Verteilungen τ_i — ob lokal oder global — modelliert werden, können diese in die allgemeine Bayessche Regel (2.6) eingesetzt werden:

$$\delta(x) = \arg \min_{i \in \{1,\ldots,k\}} \sum_{l=1}^{k} c_{li} \pi_l f_l(x) \tag{6.1}$$

(bzw. τ_i anstelle von $\pi_i f_i(x)$). Das ist zum Beispiel für Kernschätzungen, Nearest Neighbours, CART und Neuronale Netzwerke der Fall (Kapitel 4.2 - 4.5). Andere Verfahren, etwa die Lineare, Quadratische und Logistische Diskriminanzanalyse (Kapitel 3.2 - 3.4), betrachten allein den Quotienten der Verteilungen zweier Klassen. Da bei ersteren aber die Normalverteilungsannahme zugrunde liegt, können hier ebenfalls Klassendichten geschätzt und eingesetzt werden.

Da die Logistische Diskriminanzanalyse nur den Quotienten der Klassendichten modelliert, ist sie auf obigem Wege nicht zu verallgemeinern. Seber ([198], Kap. 6.9) schlägt als „natürliche“ Erweiterung auf k Klassen die wiederholte paarweise Modellierung der Logarithmen der Dichtequotienten,

$$\log\left(\frac{f_i(x)}{f_k(x)}\right) = \beta_{0i} + \beta_i' x \quad (i = 1, \ldots, k-1), \tag{6.2}$$

vor. Daraus folgt für die A-posteriori-Dichten:

$$\tau_i(x) = \frac{\mathrm{e}^{z_i}}{\sum\limits_{j=1}^{k} \mathrm{e}^{z_j}} \quad (i = 1, \ldots, k), \tag{6.3}$$

wobei $z_i = \beta_{0i} + \beta_i' x$ für $i = 1, \ldots, k-1$ und $z_k = 0$ gilt (β_{0i} enthält die A-priori-Wahrscheinlichkeiten als additiven Anteil $\log(\pi_i/\pi_k)$). Die Parameter werden wieder über ML-Schätzungen bestimmt. Die erhaltenen A-posteriori-Dichten werden anschließend in die Bayessche Regel (2.6) eingesetzt.

Um Verfahren, die prinzipiell nur auf zwei Klassen angewendet werden können, auch bei mehreren Klassen benutzen zu können oder um unter gewissen Umständen Verbesserungen der allgemeinen k-Klassen-Regel (6.1) zu erzielen, gibt es zwei Ideen, die wieder auf Zwei-Klassen-Probleme zurückgreifen:

- *Paarweise Regeln* separieren jeweils zwei Klassen unabhängig von den anderen; ihre Ergebnisse werden dann geeignet kombiniert,
- *Ansätze über Superklassen* fassen die k Klassen wiederholt zu unterschiedlichen Superklassen-Paaren zusammen und leiten dafür eine Regel ab; deren Ergebnisse werden ebenfalls kombiniert.

Beiden Ansätzen ist so ein *Voting* der Ergebnisse verschiedener Regeln, d.h. deren Kombination zu einer zusammengesetzten Regel, gemein. Diese Eigenschaft teilen sie mit Verfahren zur Varianzreduktion, siehe Kapitel 5.1.

6.2 Paarweise Regeln

Friedman [76] betrachet den Fall gleicher Kosten für eine Fehlklassifikation in eine beliebige andere Population:

$$c_{ij} \equiv c_{ik} =: c_i \ \forall j, k \neq i \quad \text{und} \quad c_{ii} = 0. \tag{6.4}$$

Die allgemeine Regel (6.1) reduziert sich dadurch zu

$$\delta(x) = \arg \max_{i \in \{1,\ldots,k\}} c_i \tau_i(x), \tag{6.5}$$

so daß bei symmetrischen Kosten die Klasse mit der höchsten Wahrscheinlichkeit unter der gezogenen Stichprobe gewählt wird; $\tau_i(x)$ wird dabei eventuell durch $\pi_i f_i(x)$ ersetzt. Durch einige Umformungen erhält Friedman [76]

$$\delta(x) = \arg \max_{i \in \{1,\ldots,k\}} \sum_{l=1}^{k} I_{c_i \tau_i^{(il)}(x) > c_l \tau_l^{(il)}(x)} \tag{6.6}$$

als äquivalenten Ausdruck zu (6.5), wobei $\tau_i^{(il)}(x) := \frac{\tau_i(x)}{\tau_i(x)+\tau_l(x)}$ die A-posteriori-Verteilung im paarweisen Vergleich bezeichnet. Die Regel (6.6) ordnet x also derjenigen Klasse zu, die am häufigsten von den paarweisen Regeln gewählt wurde. Die allgemeine Bayessche Regel für k Klassen kann so als Voting-Kombination der Bayesschen Regeln zur Diskrimination aller $\binom{k}{2}$ möglichen Paare von Populationen (Π_i, Π_l) beschrieben werden. Benutzt man in (6.1) und (6.6) dieselben separaten Schätzungen[1] $\hat{\tau}_i(x)$ bzw. $\hat{\pi}_i \hat{f}_i(x)$, so liefern sie die gleichen Ergebnisse. Die Idee von Friedman [76] besteht nun darin, die $\hat{\tau}_i(x)$ bzw. $\hat{f}_i(x)$ so zu bestimmen, daß sie stattdessen das Risiko für das Problem (Π_i, Π_l) minimieren. So können, heuristisch gesehen, die Information aus $\mathcal{S} \cap (\Pi_i \cup \Pi_l)$ genutzt und die Schätzungen direkt in Hinsicht auf ein Klassifikationsrisiko anstelle eines Approximationsfehlers optimiert werden, mit „lokaler“ Bias-Varianz-Abwägung.

Durch die wiederholte Betrachtung von Zwei-Klassen-Problemen müssen wesentlich mehr Trennflächen mit jeweils weniger Trainingsdaten geschätzt werden. Eine Kompensation dafür, auch für die möglicherweise wachsende Varianz, können aber

[1] d.h. nur aus $\mathcal{S} \cap \Pi_i$ und nach Gütekriterien der allgemeinen Kurvenapproximation (siehe Tabelle 4.2, S. 51)

einfachere Trennflächen mit geringem Bias in den Teilproblemen sein, so daß sich das Risiko der zusammengesetzten Regel insgesamt verringert. Beispiele erfolgreicher Anwendungen geben z.B. FRIEDMAN [76] und HASTIE & TIBSHIRANI [119]. So könnte sich im Falle dreier Normalverteilungen mit unterschiedlichen Kovarianzen der Bias bei Anwendung der Linearen Diskriminanzanalyse (Kapitel 3.2) verringern, weil Gleichheit nur jeweils zwischen Paaren von Klassen anstatt in allen drei gleichzeitig vorausgesetzt wird. Ein Problem anderer Art ist die Möglichkeit von Indifferenzbereichen, wenn mehrere Klassen das maximale Voting in (6.6) erreichen, also z.B. bei einem 3-Klassen-Fall jede Klasse im paarweisen Vergleich einmal „gewinnt" und einmal „verliert".

HASTIE & TIBSHIRANI [119] verallgemeinern die Strategie (6.6), indem sie die Schätzungen $\left\{\hat{\tau}_i^{(il)}(x)\right\}$ der paarweisen A-posteriori-Verteilungen so ermitteln, daß ihr mittlerer Abstand zu $\left\{\frac{\tau_i(x)}{\tau_i(x)+\tau_l(x)}\right\}$ gemäß einer abgewandelten KULLBACK-LEIBLER-Distanz minimal wird. Unter anderem erreichen sie dadurch eine Auflösung von Indifferenzbereichen.

6.3 Ansätze über Superklassen

Anstatt paarweise Regeln zu betrachten, stellen JAMES & HASTIE [130] eine Methode vor, die auf DIETTERICH & BAKIRI ([51], [135], [52]) zurückgeht und Ähnlichkeiten mit Bagging (siehe Kapitel 5.1.1) besitzt. Die verschiedenen Klassen werden durch Vereinigung wiederholt zu je zwei Superklassen mit den Marken $i^* \in \{0, 1\}$ zusammengefaßt. Wenn Z_{ij} die Zugehörigkeit der i-ten Klasse bei der j-ten Wiederholung kodiert, erhält man eine $k \times J$-dimensionale Matrix $Z = \left(Z^1, \ldots, Z^J\right)$, die beispielsweise folgende Form annehmen kann ($k = 4$):

	Wiederholungen					
Klasse	Z^1	Z^2	Z^3	Z^4	...	Z^J
Π_1	1	0	0	1	...	1
Π_2	1	1	1	0	...	0
Π_3	0	1	0	1	...	0
Π_4	0	0	1	1	...	1

Population Π_1 ist in diesem Beispiel bei der 1., 4. und J-ten Wiederholung in der zweiten Superklasse enthalten, bei der 2. und 3. in der ersten.

Für die Generierung von Z sind die Wege des Designs, d.h. des Festsetzens der Klassenzugehörigkeiten, und der Randomisierung, d.h. der zufälligen Zuordnung, möglich. Für jede Wiederholung wird eine sogenannte Basisregel[2] δ^j „trainiert", d.h. hinsichtlich eines minimalen BAYESschen Risikos für das binäre Problem der beiden Superklassen geschätzt. Es kommen dabei nur solche Verfahren in Betracht, die eine Schätzung der A-posteriori-Verteilung τ_{i^*} liefern, z.B. CART (Kapitel 4.4) oder auch die LDA (Kapitel 3.2), die dies mittels Klassendichte f_{i^*} und A-priori-Verteilung π_{i^*} über das BAYESsche Theorem (2.12) erreicht. Bei einem neu zu klassifizierenden Objekt x liefern die Regeln $\{\delta^1, \ldots, \delta^J\}$ nun einen Vektor von Schätzungen für die Zugehörigkeit zu den Superklassen:

$$\hat{\tau}_j(x) := P\left(i_j^* = 1|x\right) \quad (j = 1, \ldots, J). \tag{6.7}$$

[2]engl. *base classifier*

Der L_1-Abstand[3] zwischen $p = (\hat{\tau}_1, \ldots, \hat{\tau}_J)$ und dem bekannten Zeilenvektor Z_i der Zuordnungen der Klasse Π_i zu den einzelnen Superklassen,

$$d_i := \sum_{j=1}^{J} |\hat{\tau}_j(x) - Z_{ij}| \quad (= 1, \ldots, k), \tag{6.8}$$

mißt den Abstand zwischen wahrer und vorhergesagter Zugehörigkeit zu den Superklassen, dient also als Verlustfunktion. Das Objekt x wird dann der Klasse zugeordnet, die die größte Übereinstimmung im Sinne von d_i besitzt:

$$\delta_{ECOC}(x) := \arg \min_{i \in \{1,\ldots,k\}} d_i \,. \tag{6.9}$$

In Anlehnung an die Notation von Dietterich & Bakiri [51] nennen James & Hastie [130] diesen Ansatz *ECOC*[4]*-PICT*[5]. Sie sehen ihn als ein gewichtetes Voting der J Regeln an, das in Analogie zu Bootstrap-Schätzungen die Varianz der zusammengesetzten Regel und dadurch eventuell ihr Risiko verringert. Kong & Dietterich [135] zeigen auch eine mögliche Bias-Verringerung. Sie sehen den Erfolg dieses Ansatzes vor allem in Fällen, in denen der Bias der einzelnen Voting-Regeln „unkorreliert" ist, also nicht stets an derselben Stelle im Raum auftritt (z.B. bei Baumalgorithmen, Kapitel 4.4), und die Basisregeln ihn in der zusammengesetzten Regel gegenseitig verringern können.

Schapire [186] und Schapire & Singer [188] kombinieren die ECOC-Methode mit dem Boosting-Ansatz (Kapitel 5.1.2), indem für das Problem der beiden Superklassen die Regel per Boosting bestimmt wird. Dies dient auch dazu, den Fall *mehrfacher* Zugehörigkeiten zu verschiedenen Klassen (*multi-label*) zu behandeln.

[3] siehe die Definition des Minkowski-Abstandes (2.35) auf S. 21
[4] *Error coding output correction*
[5] *plug-in classification technique*

Kapitel 7

Eine Demonstration an Datensätzen

Das Ziel dieses Kapitels ist eine Demonstration ausgewählter theoretischer Aussagen der vorausgegangenen Kapitel an praktischen Beispielen. Dazu werden einige der vorgestellten Verfahren und Schätzmethoden auf verschiedene generierte oder reale Datensätze angewandt. Der Einfachheit halber werden die Kosten einer Fehlallokation konstant auf $c_{ij} \equiv c := 1$ gesetzt. Sämtliche Analysen wurden in der statistischen Programmiersprache S-PLUS gerechnet (siehe z.B. [218]).

7.1 Herkunft der Daten & Prozeduren

Der Quellcode für die Generierung der nachfolgend aufgeführten Datensätze in S-PLUS ist im Kapitel A.1 dokumentiert.

Generierte Datensätze nach Friedman

Die Datensätze Beispiel 1 bis Beispiel 5 sind künstlich generierte Datensätze. Als solche besitzen sie sehr genau definierte Eigenschaften. Sie wurden für diese Arbeit in Analogie zu den Angaben von Friedman [73] erzeugt, um mal das eine, mal das andere Verfahren zu bevorzugen. Es wurden jeweils zwei Klassen erzeugt. Im folgenden nun ihre Beschreibung:

- Beispiel 1: Es wurden zwei Klassen mit je $n_i = 100$ Beobachtungen in $p = 10$ Merkmalen erzeugt. Die Stichprobe wurde aus folgenden Verteilungen gezogen:

$$\Pi_1 : \mathrm{N}(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, I_p) \qquad \Pi_2 : \mathrm{N}(\begin{pmatrix} \frac{\sqrt{1}}{2} \\ \vdots \\ \frac{\sqrt{p}}{2} \end{pmatrix}, \mathrm{diag}\,(\frac{1}{\sqrt{1}}, \ldots, \frac{1}{\sqrt{p}})).$$

- Beispiel 2: Es wurden zwei Klassen mit je $n_i = 100$ Beobachtungen in $p = 10$ Merkmalen erzeugt. Die Stichprobe wurde aus folgenden Verteilungen gezogen:

$$\Pi_1 : \mathrm{N}(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, I_p) \qquad \Pi_2 : \mathrm{N}(\begin{pmatrix} \frac{\sqrt{p-1+1}}{2} \\ \vdots \\ \frac{\sqrt{p-p+1}}{2} \end{pmatrix}, \mathrm{diag}\,(\frac{1}{\sqrt{1}}, \ldots, \frac{1}{\sqrt{p}})).$$

- BEISPIEL 3: Zunächst wurden $n = 200$ Beobachtungen in $p = 10$ Merkmalen aus $\mathrm{N}((0,\ldots,0)', I_p)$ erzeugt. Anschließend erfolgte die Klassenzuordnung über:

$$x \in \Pi_1 \iff \sum_{j=1}^{p} \frac{x_{(j)}^2}{j} \leq 2.5\,, \quad \text{sonst}\, x \in \Pi_2.$$

- BEISPIEL 4: Zunächst wurden $n = 500$ Beobachtungen in $p = 10$ Merkmalen aus $\mathrm{N}((0,\ldots,0)', I_p)$ erzeugt. Anschließend erfolgte die Klassenzuordnung über:

$$x \in \Pi_1 \iff \sum_{j=1}^{p} x_{(j)}^2 \leq 9.8\,, \quad \text{sonst}\, x \in \Pi_2.$$

- BEISPIEL 5: Zunächst wurden $n = 200$ Beobachtungen in $p = 10$ Merkmalen aus $\mathrm{N}((0,\ldots,0)', I_p)$ erzeugt. Anschließend erfolgte die Klassenzuordnung über:

$$x \in \Pi_1 \iff \sum_{j=1}^{p} x_{(j)} \leq 0\,, \quad \text{sonst}\, x \in \Pi_2.$$

Für alle fünf Datensätze wurden außerdem Testdatensätze vom jeweils halben Umfang erzeugt, um mit diesen eine Fehlerschätzung zu ermöglichen.

Datensätze für die graphische Demonstration

Die Datensätze BEISPIEL 6 bis BEISPIEL 11 sind ebenfallls künstlich generiert. Sie dienen jedoch der graphischen Illustration und besitzen deswegen nur zwei Dimensionen ($p = 2$). Es wurden immer zwei Klassen mit je $n_i = 100$ Beobachtungen erzeugt. Die Stichproben wurden aus folgenden Normalverteilungen (für Π_1 bzw. Π_2) gezogen:

- BEISPIEL 6: $\mathrm{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
- BEISPIEL 7: $\mathrm{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right)$
- BEISPIEL 8: $\mathrm{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}\right)$
- BEISPIEL 9: $\mathrm{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 5 \end{pmatrix}\right) \quad \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}\right)$
- BEISPIEL 10: $\mathrm{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 5 \end{pmatrix}\right) \quad \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right)$
- BEISPIEL 11: $\mathrm{N}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}\right) \quad \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}\right)$.

Reale Datensätze

Als realer Datensatz wird LIVER DISORDERS („BUPA") von RICHARD S. FORSYTH (BUPA Medical Research Ltd.) verwendet. Er enthält Messungen an 345 Personen in 6 stetigen Merkmalen bei zwei Klassen. Dieser und weitere Datensätze können über eine der folgenden Internet-Adressen bezogen werden:

- `ftp.ics.uci.edu/pub/machine-learning-databases/`
- `lib.stat.cmu.edu/datasets/`

Dort ist auch eine detailierte Beschreibung über Herkunft, bisherige Benutzung und Analyse der Daten sowie eine inhaltliche Erläuterung der Messungen erhältlich.

Verfahren & Prozeduren

Folgende Verfahren wurden benutzt: Lineare Diskriminanzanalyse (LDA, Kapitel 3.2), Quadratische Diskriminanzanalyse (QDA, Kapitel 3.3), Nearest Neighbours (k-NN, Kapitel 4.3), CART (Kapitel 4.4) und Neuronales Netz (NNet, Kapitel 4.5).

Bei LDA und QDA wurden die Standardverfahren verwendet. Der CART-Algorithmus wurde gemäß der Funktion `tree` in S-PLUS ohne Pruning benutzt, d.h. die resultierende Regel ist zu stark an die Stichprobe angepaßt. Dies wurde aber in Kauf genommen, da lediglich einige theoretische Sachverhalte demonstriert werden sollen. Als Neuronales Netz wurde ein One-Hidden-Layer mit zwei Knoten in der Zwischenschicht verwendet. Die Anzahl der Knoten der Zwischenschicht wurde auf 2 gesetzt, um einer zu starken Varianz vorzubeugen.[1] Beim Nearest-Neighbours-Verfahren wurde der gewöhnliche EUKLIDische Abstand[2] L_2 benutzt.

Die angewendeten Diskriminanzregeln sind entweder Teil der S-PLUS-Auslieferung (`lda`, `qda`, `tree`), bzw. der Bibliothek `nnet` von VENABLES & RIPLEY [218] (`nnet`), oder wurden selbst programmiert (`kNN`, in starker Anlehnung an eine Prozedur von POLZEHL). Sämtliche Funktionen zur Generierung der Daten, zur Fehlerschätzung und graphischen Darstellung sind selbst programmiert, dabei teilweise in Anlehnung an [218].

7.2 Ergebnisse

Der Quellcode für die Generierung der nachfolgend aufgeführten Tabellen und Graphiken in S-PLUS ist im Kapitel A.2 dokumentiert.

Kommentar 7.1 (Vorbemerkung) *Bei der Begutachtung der praktischen Demonstrationen wurde der viel zu geringe Umfang der Testdatensätze kritisiert: Für* $\pi_1 = \pi_2 = \frac{1}{2}$, $c_{ij} \equiv c$ *und* $n_1 = n_2 = 100$ *ist das Risiko der Diskriminanzregel binomialverteilt mit Standardabweichung* $\sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 0,05$. *Somit ist innerhalb eines Bereiches von* $\pm 0,15$ *aufgrund der* 3σ*-Regel keine Signifikanz für die getroffenen Aussagen in Bezug auf die Testdatensätze gegeben. Dies gilt für die Tabellen 7.1, 7.2 und 7.4. Um den Faktor* 100 *oder* 1.000 *erhöhte Umfänge des Testdatensatzes wären sinnvoll.*

[1] Betrachtet man ein vollständig verbundenes Netz mit Skips, ergeben sich bei p Einflußgrößen und a Knoten in der Zwischenschicht $ap + p + a$ zu schätzende Parameter. Bei $p = 10$ und $a = 2$ sind das 32. Bei einem Stichprobenumfang von $n = 200$ erscheint dies realistisch, um eine zu starke Varianz zu verhindern. Die mögliche Kreuzvalidierung zur Wahl von a wurde nicht angewendet, da nur theoretische Aspekte demonstriert werden sollen.

[2] siehe die Definition auf S. 60

	Datensatz	*AER*	*2-CV*	*10-CV*	*25-CV*	*LOO*	*Test*
LDA	Beispiel 1	1,5	3,0	2,0	2,0	2,0	3,0
	Beispiel 2	0,0	0,5	0,5	0,5	0,5	3,0
	Beispiel 3	46,0	52,5	59,0	61,5	59,0	44,0
	Beispiel 4	39,2	46,6	42,0	41,6	41,6	35,6
	Beispiel 5	4,0	8,0	5,0	6,0	5,5	2,0
	Liver disorders	29,6	32,2	33,0	31,6	30,0	
QDA	Beispiel 1	0,0	0,5	0,5	0,5	0,5	1,0
	Beispiel 2	0,0	0,5	1,0	1,0	1,0	0,0
	Beispiel 3	11,0	27,0	23,5	23,5	22,5	26,0
	Beispiel 4	9,0	19,2	13,4	13,2	14,4	13,2
	Beispiel 5	5,5	16,5	10,0	11,0	10,5	8,0
	Liver disorders	36,5	44,8	40,0	40,0	41,6	
CART	Beispiel 1	2,5	14,5	6,5	8,5	6,0	9,0
	Beispiel 2	4,0	12,0	9,5	9,0	9,0	10,0
	Beispiel 3	11,5	32,0	27,0	25,5	27,0	27,0
	Beispiel 4	8,0	33,4	27,4	29,2	25,0	28,8
	Beispiel 5	7,5	34,5	31,0	34,5	29,0	36,0
	Liver disorders	13,0	40,3	33,3	33,0	34,2	
NNet	Beispiel 1	0,0	5,0	4,0	3,0	4,0	2,0
	Beispiel 2	0,0	1,0	2,0	0,5	1,5	3,0
	Beispiel 3	31,0	47,5	38,0	46,5	43,0	32,0
	Beispiel 4	26,0	45,2	41,2	38,4	40,0	33,6
	Beispiel 5	0,0	4,0	4,5	3,5	3,0	2,0
	Liver disorders	57,6	38,3	37,7	37,7	34,5	

Tabelle 7.1: Unterschiedliche Risikoschätzungen. Schätzungen der Fehlklassifikationswahrscheinlichkeiten verschiedener Verfahren bei Anwendung auf einige generierte und reale Datensätze. Als Schätzmethoden wurden die Resubstitutionsmethode (AER), Cross-validation (CV) mit 2, 10, 25 und n Gruppen (Leave-one-out, LOO) sowie ein Testdatensatz — so vorhanden — benutzt (siehe Kapitel 2.2.3; Angaben in %).

7.2.1 Unterschiedliche Risikoschätzungen

Die Tabelle 7.1 zeigt die Ergebnisse der Risikoschätzungen durch verschiedene Methoden (siehe Kapitel 2.2.3) für die Datensätze Beispiel 1–5 und unterschiedliche Diskriminanzregeln. Es fand keine Variablenauswahl statt.

Als erstes Resultat sticht die teils wenig, teils dramatisch geringere Fehlerschätzung durch die Resubstitutionsmethode (AER (2.23)) hervor. Dies ist eine Folge des Bias dieser Schätzung, durch den sie das Risiko beständig unterschätzt. Sie ist, wie theoretisch in Kapitel 2.2.3 ausgeführt wurde, nicht für eine verläßliche Fehlerschätzung geeignet.

Die Schätzungen über Cross-validation (CV, Kapitel 2.2.3) mit 10 bzw. 25 Gruppen liefern einander ähnliche Ergebnisse. Leave-one-out (LOO) liegt bei diesen Beispielen in derselben Größenordnung. 10-CV erhält diese Ergebnisse aber mit einem Bruchteil des Rechenaufwandes von 25-CV bzw. LOO.

Interessanterweise zeigt die Fehlerschätzung über einen Testdatensatz größere Abweichungen gegenüber Cross-validation sowohl nach oben als auch nach unten.

	Datensatz	*Anzahl aller Variablen*	*Risiko*	*Umfang der besten Teilmenge*	*Risiko*
LDA	Beispiel 1	10	3,0	6	0,0
	Beispiel 2	10	3,0	6	0,0
	Beispiel 3	10	44,0	6	39,0
	Beispiel 4	10	35,6	4	33,6
	Beispiel 5	10	2,0	10	2,0
QDA	Beispiel 1	10	1,0	7	0,0
	Beispiel 2	10	0,0	7	0,0
	Beispiel 3	10	26,0	4	14,0
	Beispiel 4	10	13,2	9	13,2
	Beispiel 5	10	8,0	10	8,0

Tabelle 7.2: Effekt einer Variablenauswahl. Für zwei Verfahren (LDA und QDA) ist der Anzahl der ursprünglichen Variablen und ihres geschätzten Risikos bei Anwendung auf einige generierte Datensätze der Umfang der Variablenteilmenge gegenübergestellt, die die geringste Risikoschätzung erzielte. Die Schätzungen des Risikos wurden über die Testdatensätze ermittelt (Angaben in %).

Dies ist ein Zeichen für die größere Variabilität der Schätzung. Angesichts des relativ kleinen Stichprobenumfangs ($n = 200$ beim Trainingsdatensatz, $n = 100$ beim Testdatensatz) hinsichtlich der Dimensionalität des Problems ($p = 10$), ist das jedoch nicht weiter verwunderlich. Die Ähnlichkeit der CV-Schätzungen, die ja alle auf derselben Stichprobe beruhen, verdeckt eher das Problem der Varianz der Fehlerschätzung. Die Auswahl des geeigneten Verfahrens ist so selbst ein Schätzproblem; die Entscheidung über das endgültige Verfahren sollte daher in einer Gruppe von Verfahren mit ähnlich niedrigen Risiken nach weiteren Kriterien erfolgen (siehe Kapitel 2.6). — *Bitte siehe Kommentar 7.1.* —

Die Schätzung über Cross-validation mit zwei Gruppen (2-CV), d.h. die Halbierung der Stichprobe, führt im Mittel zu größeren Fehlerschätzungen. Das ist in dem verminderten Umfang der Stichprobe, die zur Bestimmung der Regel zur Verfügung steht, begründet. Dieser Ansatz ist daher ungeeignet, da er zuviel Information aus der Stichprobe verschenkt.

7.2.2 Effekt einer Variablenauswahl

Die in Kapitel 2.5.1 erwähnte Möglichkeit einer Risikoverminderung durch Variablenauswahl soll an zwei Verfahren (LDA und QDA) demonstriert werden. Beide wurden auf die fünf generierten Beispieldatensätze nach Friedman angewandt. Der Risikoschätzung bei Einschluß sämtlicher Variablen wird die der Teilmenge mit dem geringsten Risiko gegenübergestellt (Tabelle 7.2).

Die ersten vier Beispiele zeigen, daß eine geringere Fehlerrate oder dieselbe mit weniger Variablen, eventuell verbunden mit geringeren Kosten der Erhebung, erreichbar ist. Besonders deutlich ist dieser Effekt bei der QDA und dem Beispiel 3. Das Beispiel 5 zeigt dagegen den Fall, daß sämtliche Variablen für eine Klassifikation wichtig sind und jedes Weglassen einer Variablen das Risiko vergrößert.

Weiterhin ist zu bemerken, daß für verschiedene Verfahren unterschiedliche Variablen relevant sein können. So benutzt die LDA im Beispiel 4 nur 4 Variablen (und erzielt aufgrund eines großen Trennflächenfehlers nur ein unbefriedigendes Er-

	$k =$							
Datensatz	*1*	*2*	*3*	*5*	*7*	*10*	*15*	*20*
Beispiel 1	1,0	1,0	1,5	1,5	1,5	1,5	1,5	1,0
Beispiel 2	1,0	0,5	1,5	1,0	1,5	0,5	1,5	1,5
Beispiel 3	35,0	46,0	37,5	38,0	39,0	42,5	43,0	50,0
Beispiel 4	33,8	39,8	37,0	37,6	38,4	41,6	41,8	42,6
Beispiel 5	24,5	25,5	20,5	16,5	15,0	16,5	14,5	14,0
Liver disorders	37,7	44,4	37,1	33,0	29,6	31,9	31,3	31,6

Tabelle 7.3: Bias-Varianz-Abwägung. Geschätzte Fehlklassifikationswahrscheinlichkeiten des Nearest-Neighbours-Verfahrens bei Anwendung auf einige generierte und reale Datensätze mit verschiedenen Werten für k. Die Schätzung erfolgte über Cross-validation mit 10 Gruppen (Angaben in %).

gebnis), während die QDA auf 9 Variablen zurückgreift. — *Bitte siehe Kommentar 7.1.* —

7.2.3 Effekt der Bias-Varianz-Abwägung

Um den Effekt der Bias-Varianz-Abwägung (siehe Kapitel 2.3) zu demonstrieren, wurde das Nearest-Neighbour-Verfahren gewählt, da hier nur ein einziger Parameter, k, diese Abwägung moderiert. Die über Cross-validation mit 10 Gruppen geschätzten Fehlklassifikationswahrscheinlichkeiten für verschiedene k sind in der Tabelle 7.3 aufgeführt.

Die ersten beiden Beispiele erscheinen sehr gut separabel, so daß die Glättung aufgrund einer sehr geringen Varianz praktisch keinen Effekt hat. Bei Beispiel 3 und 4 ist eine eher geringe Glättung mit $k = 1, \ldots, 5$ von Nutzen, da hier anscheinend der Bias das größere Problem darstellt und mehr zum Risiko beiträgt. Bei Beispiel 5 und Liver disorders verbessert eine starke Glättung bis hoch zu $k = 20$ das Risiko; die Reduktion der Varianz, die in der Zufälligkeit der Stichprobe begründet ist, verringert auch das Risiko, da die mögliche Erhöhung des Bias gering ausfällt.

Datensatz	*LDA*	*QDA*	*CART*	*One-Hidden-Layer*
Beispiel 1	0,0	0,0	10,0	0,0
Beispiel 2	0,0	0,0	7,0	0,0
Beispiel 3	39,0	14,0	27,0	19,0
Beispiel 4	33,6	13,2	25,6	29,6
Beispiel 5	2,0	8,0	36,0	2,0

Tabelle 7.4: Wechselnde optimale Verfahren. Risikoschätzungen für vier Verfahren bei Anwendung auf einige generierte Datensätze. Bei LDA, QDA und NNet wurde eine Variablenauswahl durchgeführt, bei CART alle Variablen angeboten. Die Schätzungen des Risikos wurden über die Testdatensätze ermittelt (Angaben in %).

7.2.4 Wechselnde optimale Verfahren

Um die Aussage zu illustrieren, daß kein Verfahren optimal für alle Situationen ist, wurden vier Verfahren in fünf Datensätzen miteinander verglichen. Für die LDA, die QDA und das One-Hidden-Layer wurde jeweils die beste Variablenteilmenge bestimmt; bei CART wurden sämtliche Variablen eingeschlossen. Die Risikoschätzungen mittels der Testdatensätze sind in der Tabelle 7.4 aufgeführt.

Die ersten beiden Beispiele scheinen sehr gut separabel zu sein, so daß bis auf CART aufgrund seiner achsenparallelen Limitierung alle Verfahren keine großen Probleme besitzen. BEISPIEL 3 und 4 besitzen jeweils stark unterschiedliche Kovarianzmatrizen, so daß die theoretisch optimale QDA auch praktisch den anderen Verfahren eindeutig überlegen ist. BEISPIEL 5 besitzt aufgrund seiner Konstruktion eine simple lineare Trennfläche. Das One-Hidden-Layer tut sich nicht schwer, diese zu finden. Die LDA erreicht ebenfalls die optimale Trennfläche, da die Kovarianzmatrizen in beiden Klassen annähernd gleich sein werden, so daß sie gegenüber der QDA im Vorteil ist. — *Bitte siehe Kommentar 7.1.* —

Diese Beispiele besitzen eine generierte, bekannte Struktur. Bei realen Daten ist das meist nicht der Fall. Um ein möglichst gutes Verfahren zu finden, sollten daher in ungewissen Situationen mehrere Verfahren angewandt werden (siehe Kapitel 2.5.2).

7.2.5 Formen der Trennflächen

Abschließend soll eine Vorstellung davon gegeben werden, wie die Trennflächen der einzelnen Verfahren aussehen können. Dazu wurden die vier benutzten Verfahren auf die zweidimensionalen Datensätze BEISPIEL 5 - 11 angewandt. Die Abbildungen im folgenden zeigen die Objekte der beiden Klassen („+“ und „o“ für Objekte aus Population Π_1 bzw. Π_2) sowie die Linie, die die Zuordnungsgebiete der jeweiligen Regel, Ω_1 und Ω_2, trennt (Trennfläche). Das Risiko der Regeln wurde über 10-Gruppen-Cross-validation geschätzt. Die bei der Darstellung auftretende Rasterung der Trennlinien ist technischer Natur; sie sollten im Auge des Betrachters als Approximation glatter Linien erscheinen. Nicht achsenparallele Teilstücke bei CART sind ebenfalls eine Folge der Rasterung bei der Darstellung.

Die Abbildung 7.1 zeigt die Trennflächen für zwei normalverteilte Populationen mit gleicher Kovarianz. Zwischen LDA, QDA und Neuronalem Netz sind kaum Unterschiede auszumachen; alle drei bemühen sich um die Approximation der theoretisch besten linearen Trennfläche (vgl. Kapitel 3.2). Lediglich CART hat aufgrund seiner achsenparallelen Beschränkung der Trennfläche arge Schwierigkeiten, die theoretisch optimale „schiefe“ Gerade $y = -x$ nachzuvollziehen.

Die Abbildungen 7.2 bis 7.5 zeigen normalverteilte Populationen mit unterschiedlichen Kovarianzen in den Klassen mit Mittelwerten in $(-1, -1)'$ und $(1, 1)'$. Hier zeigen sich die unterschiedlichen „Antworten“ der Verfahren auf die Daten. Der Bias der LDA zur theoretisch optimalen quadratischen Trennfläche kommt kaum zum Tragen im Risiko, vermutlich aufgrund der relativ ähnlichen klassenspezifischen Kovarianzmatrizen, die sich auch nur geringfügig von der Einheitsmatrix unterscheiden. QDA und Neuronales Netz weisen in wesentlichen Abschnitten der Trennfläche, d.h. im Zentrum der Beobachtungen, Übereinstimmungen auf. Bemerkenswert an Abbildung 7.4 ist das bessere Abschneiden des Neuronalen Netzes, das die theoretisch optimale QDA auf den zweiten Platz verweist. Ursache hierfür ist der begrenzte Umfang der Stichprobe; im Mittel über viele Stichproben und Beobachtungen wird die QDA besser sein. Es zeigt aber die Zufälligkeit der Auswahl des optimalen Verfahrens.

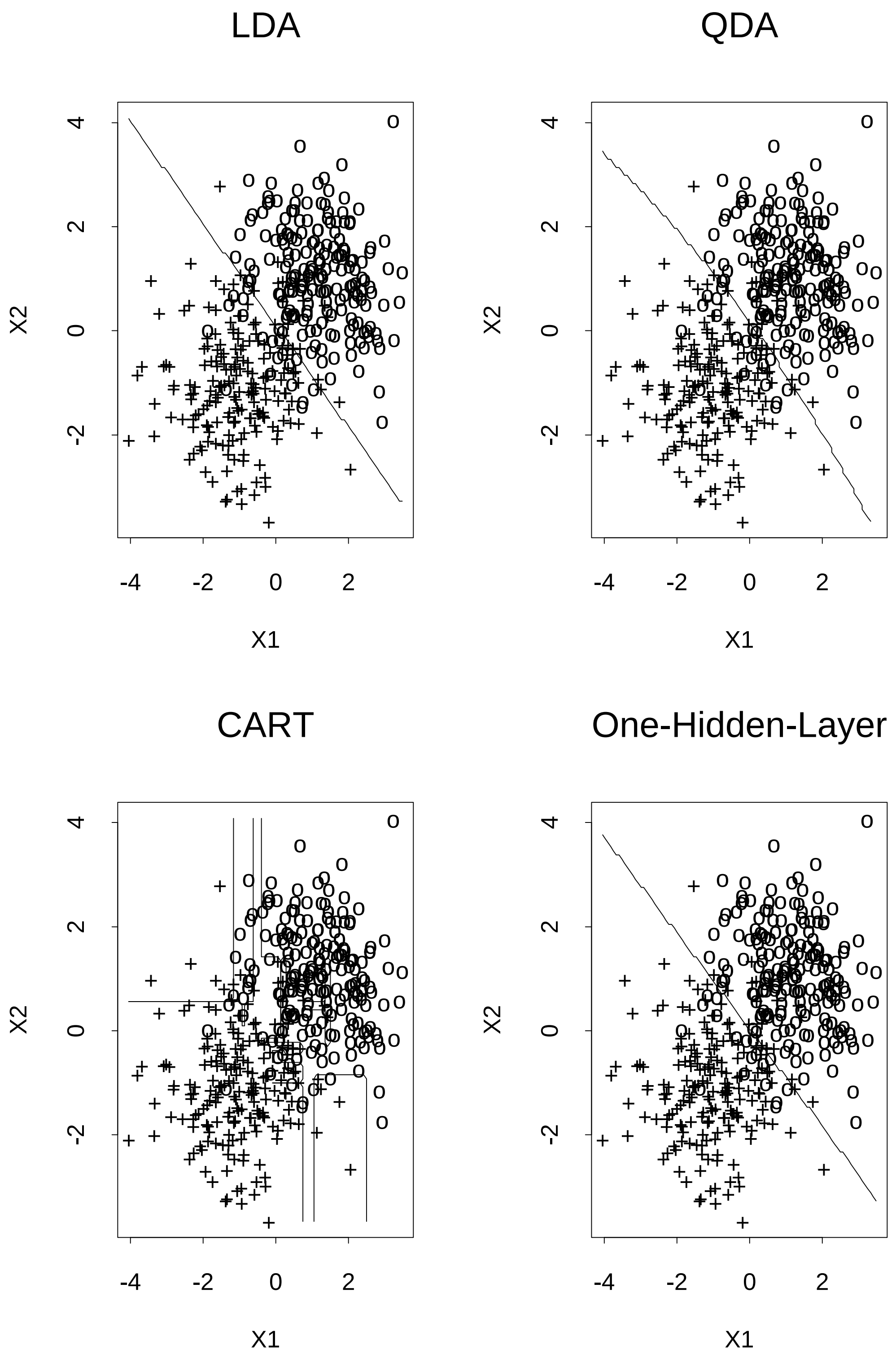

Abbildung 7.1: Die Trennflächen von vier Verfahren für den generierten Datensatz Beispiel 7. Objekte aus der Population Π_1 sind durch „+" dargestellt, solche aus Π_2 durch „o". Die über 10-CV ermittelten Fehlklassifikationswahrscheinlichkeiten betragen: LDA 9,8 %, QDA 9,5 %, CART 15,2 %, One-Hidden-Layer 10,0 %. („Brüche" in der Trennlinie sind technischer Natur bei der Darstellung.)

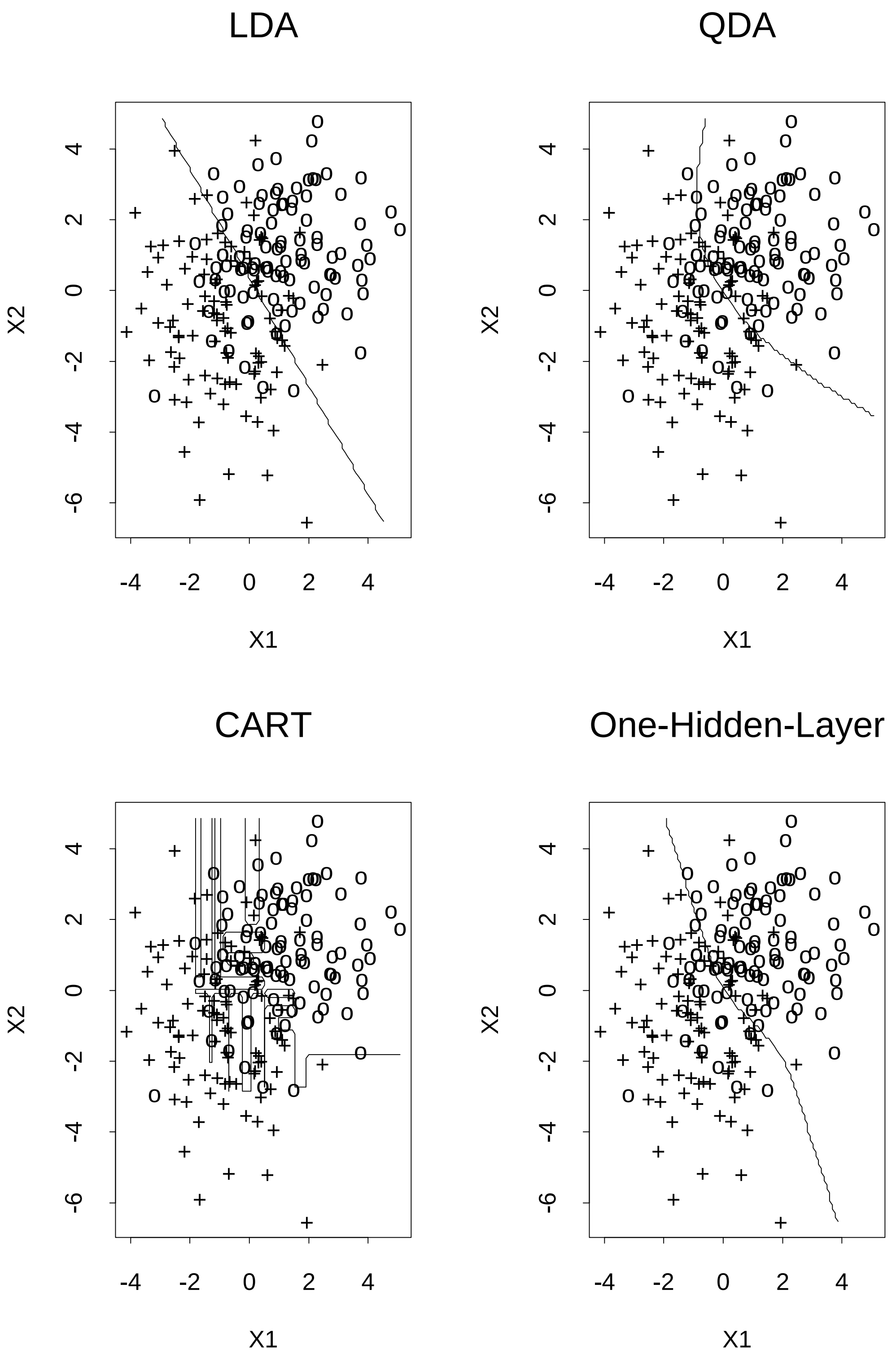

Abbildung 7.2: Die Trennflächen von vier Verfahren für den generierten Datensatz BEISPIEL 10. Objekte aus der Population Π_1 sind durch „+" dargestellt, solche aus Π_2 durch „o". Die über 10-CV ermittelten Fehlklassifikationswahrscheinlichkeiten betragen: LDA 22,0 %, QDA 22,0 %, CART 29,5 %, One-Hidden-Layer 24,0 %. („Brüche" in der Trennlinie sind technischer Natur bei der Darstellung.)

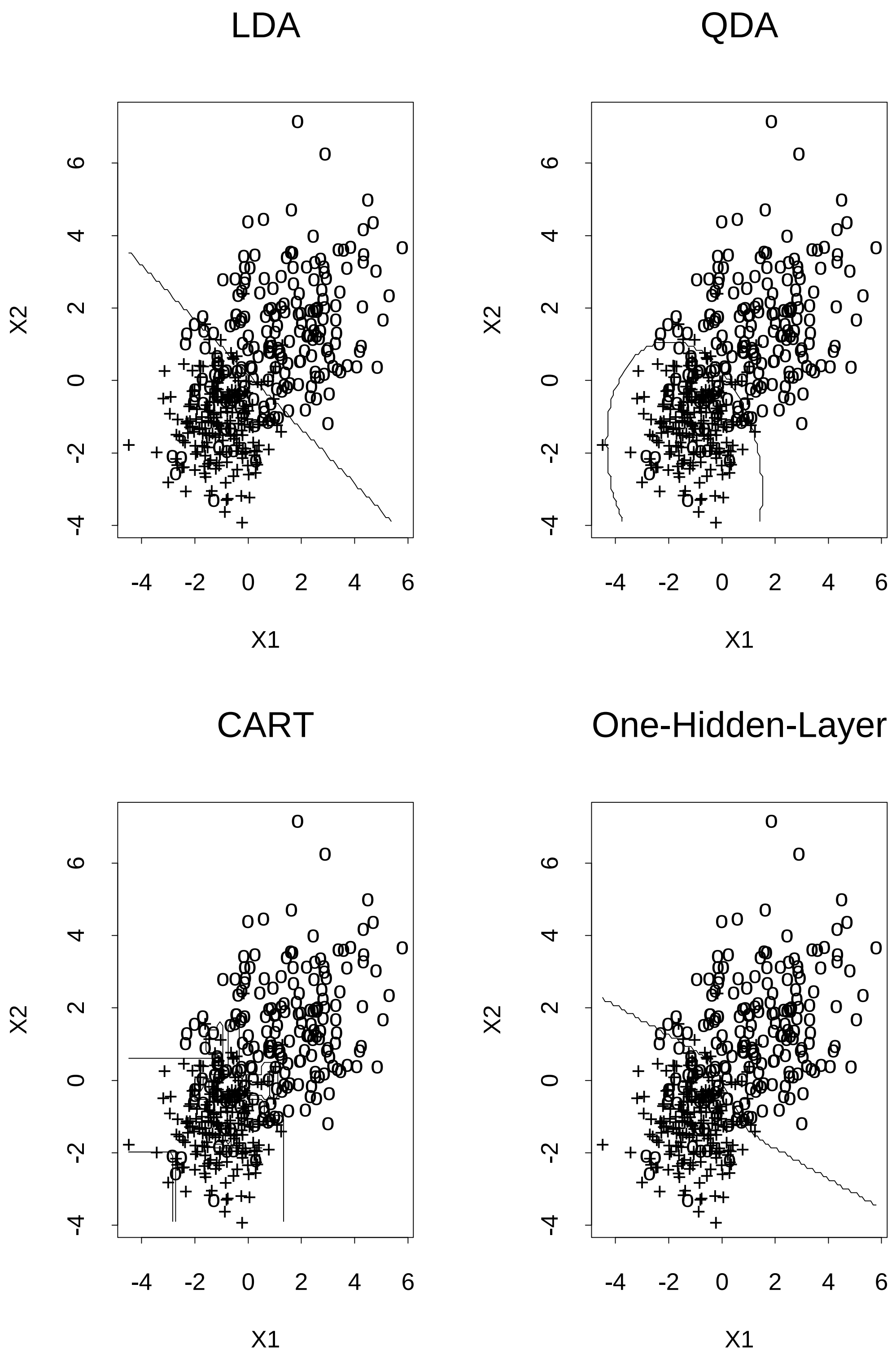

Abbildung 7.3: Die Trennflächen von vier Verfahren für den generierten Datensatz Beispiel 8. Objekte aus der Population Π_1 sind durch „+" dargestellt, solche aus Π_2 durch „o". Die über 10-CV ermittelten Fehlklassifikationswahrscheinlichkeiten betragen: LDA 16,0 %, QDA 15,8 %, CART 20,5 %, One-Hidden-Layer 15,0 %. („Brüche" in der Trennlinie sind technischer Natur bei der Darstellung.)

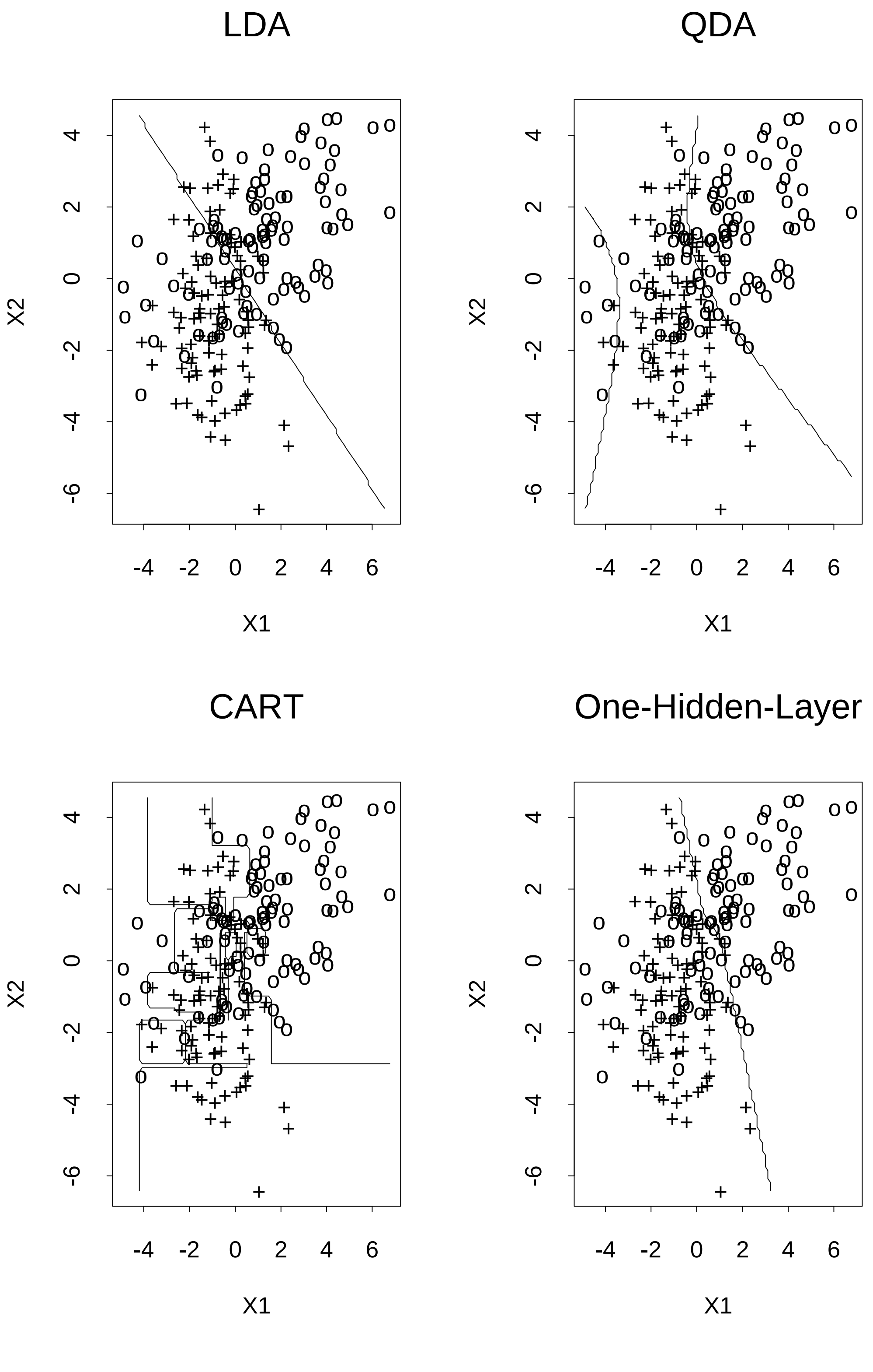

Abbildung 7.4: Die Trennflächen von vier Verfahren für den generierten Datensatz BEISPIEL 9. Objekte aus der Population Π_1 sind durch „+“ dargestellt, solche aus Π_2 durch „o“. Die über 10-CV ermittelten Fehlklassifikationswahrscheinlichkeiten betragen: LDA 27,0 %, QDA 27,0 %, CART 27,0 %, One-Hidden-Layer 22,0 %. („Brüche“ in der Trennlinie sind technischer Natur bei der Darstellung.)

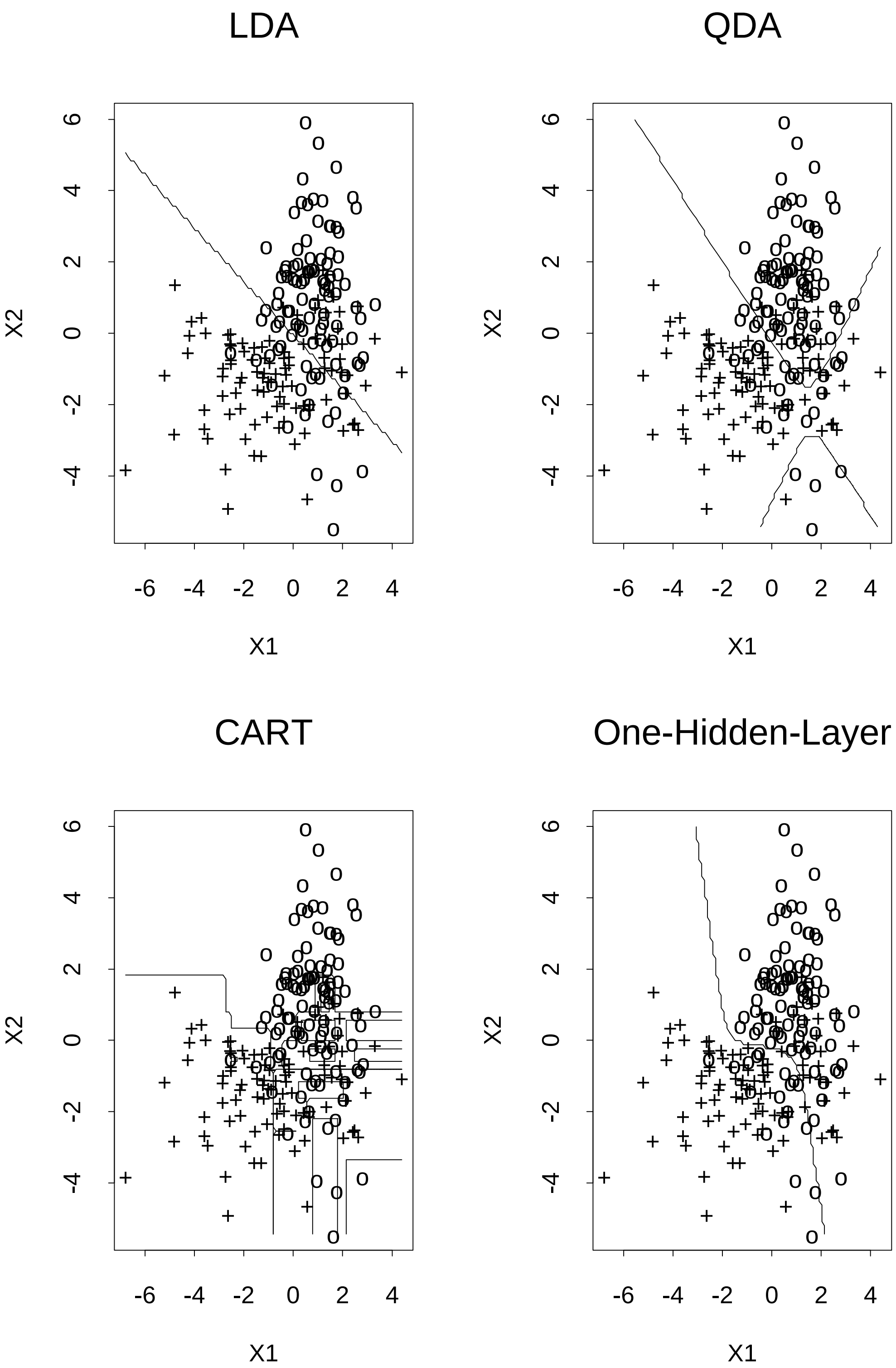

Abbildung 7.5: Die Trennflächen von vier Verfahren für den generierten Datensatz Beispiel 11. Objekte aus der Population Π_1 sind durch „+" dargestellt, solche aus Π_2 durch „o". Die über 10-CV ermittelten Fehlklassifikationswahrscheinlichkeiten betragen: LDA 24,0 %, QDA 24,5 %, CART 31,0 %, One-Hidden-Layer 25,0 %. („Brüche" in der Trennlinie sind technischer Natur bei der Darstellung.)

Kapitel 8

Praktischer Ausblick

Wie soll man nun bei praktischen Anwendungen entscheiden? Welche ist die optimale Diskriminanzregel für das gestellte Problem? Zunächst kann man feststellen:

- Bei sehr großen Stichprobenumfängen (auch relativ zu p) und gesicherter Normalverteilung in den Klassen wird man — je nach Kovarianzstruktur — mit einer Linearen oder Quadratischen Diskriminanzanlyse (Kapitel 3.2 und 3.3) auf der sicheren Seite liegen.
- Gleiches gilt für nichtnormale Verteilungen, wie sie in Kapitel 3.5 aufgeführt sind.
- Bei sehr wenigen diskreten Merkmalen ($p \leq 5$) und, relativ dazu, sehr großen Stichprobenumfängen wird das Multinomialmodell (Kapitel 4.1.1) kaum zu verbessern sein.

Obige Situationen sind aber extreme Spezialfälle: Die Verteilungen sind bekannt; und die Größe der Stichprobe bewirkt, daß theoretische Optimalitätskriterien auch praktisch gelten. Sind die Verteilungen nicht bekannt oder ist n klein relativ zu p, so daß die Varianz der geschätzten Regel wächst, finden theoretische Optimalitätskriterien ihre Grenze.

Zusammenfassend daher einige praktische Empfehlungen:

- In den meisten Anwendungen sollte eine *Bias-Varianz-Abwägung* (Kapitel 2.3) für die aufzustellende Diskriminanzregel stattfinden. Regularisierende Annahmen verringern die Varianz bei möglicher Erhöhung des Bias (Verfälschung). Solche Regularisierungen sind entweder schon durch den Verfahrensansatz bestimmt[1] oder werden durch adaptive Verfahren der Stichprobe angepaßt[2]. Da die Diskriminanzanalyse-Verfahren verschiedene Grundannahmen über die Daten machen[3] mit unterschiedlichen Folgen für Bias und Varianz, liegt ihrer Benutzung ebenfalls schon eine grobe Bias-Varianz-Abwägung zugrunde.[4] Deswegen sollten nach Möglichkeit *verschiedene Verfahren parallel* angewendet werden.

[1] z.B. bestimmte Kovarinazstrukturen in Linearer und Quadratischer Diskriminanzanalyse (Kapitel 3.2 und 3.3), die Wahl von h_i oder k in Kernschätzungen bzw. dem Nearest-Neighbour-Verfahren (Kapitel 4.2 und 4.3) oder die Wahl der Schichten in Neuronalen Netzen (Kapitel 4.5)

[2] z.B. die Regularisierte Diskriminanzanalyse (Kapitel 3.3.3), Adaptive Kerne (Kapitel 4.2), Pruning bei CART (Kapitel 4.4) oder Strafterme bei der Netzwerkschätzung (Kapitel 4.5)

[3] z.B. globale parametrische Form oder lokale Glattheit der Klassendichten

[4] z.B. bei der Anwendung der Linearen Diskriminanzanalyse (Kapitel 3.2) bei nichtnormalen Klassenverteilungen

- Eine *Variablenauswahl* kann die Varianz der Diskriminanzregel verringern; der Einschluß einer zu geringen Anzahl an Variablen kann jedoch den Bias erhöhen (Kapitel 2.5.1). Die Variablenauswahl in Hinblick auf ein konkretes Verfahren ist so ebenfalls ein Mittel der Bias-Varianz-Abwägung. Teststatistiken über den Einschluß von Variablen — hergeleitet unter bestimmten Verteilungsannahmen, die nicht erfüllt sein müssen — und sequentielles Vorgehen mit einer Abbruchbedingung bergen stets die Gefahr suboptimaler Teilmengen in sich, d.h. ein erhöhtes Risiko aufgrund einer schlechten Abwägung. Es sollten daher — wenn möglich — sämtliche Teilmengen für ein Verfahren evaluiert werden (Kapitel 2.5.1).

- Geeignete *Transformationen* von Variablen können die Voraussetzungen oder Optimalitätskriterien von Verfahren erfüllen helfen und so deren Risiko senken. Der Einschluß abgeleiteter Variablen aus vielen Klassen von Transformationen birgt damit das Potential einer Verbesserung der Diskriminanzregeln (Kapitel 2.4.2).

- Das primäre Interesse bei einer Diskriminanzregel besteht in ihrem minimalen BAYESschen Risiko (2.4) (siehe Kapitel 2.1). Darum sollte eine möglichst genaue Schätzung dieses Risikos als *Vergleichskriterium* zwischen den verschiedenen Verfahren mit den unterschiedlichen Teilmengen dienen. Durch Crossvalidation-, Bootstrap- oder Testdaten-Schätzungen (Kapitel 2.2.3) wird das geleistet. Andere Kriterien, wie z.B. Fehlermaße aus der Kurvenapproximation (vgl. Tabelle 4.2, S. 51), sind nicht unbedingt an das BAYESsche Risiko gekoppelt und sollten daher vermieden werden.

- In den meist vorkommenden, nicht durch besondere Eigenschaften ausgezeichneten Fällen wird eine *umfassende Suche* über viele Verfahrensklassen, viele Klassen an Transformationen und alle Variablenteilmengen mit einer Risikoschätzung als Vergleichskriterium (Prinzip der Adaptiven Diskriminanzanalyse, Kapitel 2.5.2) ein sehr gutes Verfahren liefern (wenn auch nicht immer das theoretisch optimale, siehe Abbildung 7.5), da die Bias-Varianz-Abwägung über ein sehr großes Spektrum an Ansätzen erfolgt.

- Die Risikoschätzungen sind durch den begrenzten Stichprobenumfang ebenfalls fehlerbehaftet, so daß das ermittelte optimale Verfahren (in gewissen Grenzen) zufällig ist. Darum sollte das letztlich benutzte Verfahren nicht allein nach einer minimalen Fehlerschätzung ausgesucht werden. Unter den Verfahren mit ähnlich niedrigem Risiko sollte stattdessen diejenige Regel gewählt werden, die *weiteren Anforderungen*, z.B. maximaler Geschwindigkeit in der Anwendung oder minimalen Kosten der Datengrundlage (siehe Kapitel 2.6), am ehesten entspricht.

Anhang A

Programme für S-PLUS

A.1 Allgemeine Prozeduren

Generierte Datensätze

```
# a function to generate random multivariate Gaussians
# (aus der S-PLUS-Online-Hilfe)
# ---------------------------------------------------------------------------
rmultnorm <- function(n, mu, vmat, tol = 1e-07) {
               p <- ncol(vmat)
               if(length(mu)!=p)
                       stop("mu vector is the wrong length")
               if(max(abs(vmat - t(vmat))) > tol)
                       stop("vmat not symmetric")
               vs <- svd(vmat)
               vsqrt <- t(vs$v %*% (t(vs$u) * sqrt(vs$d)))
               ans <- matrix(rnorm(n * p), nrow = n) %*% vsqrt
               ans <- sweep(ans, 2, mu, "+")
               dimnames(ans) <- list(NULL, dimnames(vmat)[[2]])
               ans
       }

# Beispiel 1
# ----------------------------------------------------
Gen.Beispiel.1 _ function (p=10, n=100) {
  Klasse1 _ rmultnorm (n, rep (0,p)       , diag (rep (1,p))     )
    Klasse1 _ cbind (Klasse1, 1)
  Klasse2 _ rmultnorm (n, c (sqrt (1:p)/2), diag (1/sqrt (1:p)) )
    Klasse2 _ cbind (Klasse2, 2)
  Klasse _ data.frame (rbind (Klasse1, Klasse2) )
  names (Klasse) _ c (paste ("X", 1:p, sep=""), "group")
  Klasse [,p+1] _ as.factor (Klasse [,p+1])
  Klasse
 }
Beispiel1 _ Gen.Beispiel.1 ()

# Beispiel 2
# ----------------------------------------------------
Gen.Beispiel.2 _ function (p=10, n=100) {
  Klasse1 _ rmultnorm (n, rep (0,p)       , diag (rep (1,p))     )
    Klasse1 _ cbind (Klasse1, 1)
  Klasse2 _ rmultnorm (n, c (sqrt (p-(1:p)+1)/2), diag (1/sqrt (1:p)) )
    Klasse2 _ cbind (Klasse2, 2)
  Klasse _ data.frame (rbind (Klasse1, Klasse2) )
  names (Klasse) _ c (paste ("X", 1:p, sep=""), "group")
  Klasse [,p+1] _ as.factor (Klasse [,p+1])
  Klasse
 }
Beispiel2 _ Gen.Beispiel.2 ()
```

```
# Beispiel 3
# ---------------------------------------------------
Gen.Beispiel.3 _ function (p=10, n=200) {
  Klasse _ rmultnorm (n, rep (0,p)       , diag (rep (1,p))     )
    Klasse _ cbind (Klasse, 2)
    kIndex _ rep (F, dim (Klasse) [1])
    for (i in 1: dim (Klasse) [1]) {
     kIndex[i] _ { (Klasse [i,1:p]) %*% (Klasse [i,1:p]/(1:p) ) <= 2.5 } }
    Klasse [kIndex,p+1] _ 1
  Klasse _ data.frame (Klasse)
  names (Klasse) _ c (paste ("X", 1:p, sep=""), "group")
  Klasse [,p+1] _ as.factor (Klasse [,p+1])
  Klasse
 }
Beispiel3 _ Gen.Beispiel.3 ()

# Beispiel 4
# ---------------------------------------------------
Gen.Beispiel.4 _ function (p=10, n=500) {
  Klasse _ rmultnorm (n, rep (0,p)       , diag (rep (1,p))     )
    Klasse _ cbind (Klasse, 2)
    kIndex _ rep (F, dim (Klasse) [1])
    for (i in 1: dim (Klasse) [1]) {
     kIndex[i] _ { (Klasse [i,1:p]) %*% (Klasse [i,1:p] ) <= 9.8 } }
    Klasse [kIndex,p+1] _ 1
  Klasse _ data.frame (Klasse)
  names (Klasse) _ c (paste ("X", 1:p, sep=""), "group")
  Klasse [,p+1] _ as.factor (Klasse [,p+1])
  Klasse
 }
Beispiel4 _ Gen.Beispiel.4 ()

# Beispiel 5
# ---------------------------------------------------
Gen.Beispiel.5 _ function (p=10, n=200) {
  Klasse _ rmultnorm (n, rep (0,p)       , diag (rep (1,p))     )
    Klasse _ cbind (Klasse, 2)
    kIndex _ rep (F, dim (Klasse) [1])
    for (i in 1: dim (Klasse) [1]) {
     kIndex[i] _ { sum (Klasse [i,1:p])  <= 0 } }
    Klasse [kIndex,p+1] _ 1
  Klasse _ data.frame (Klasse)
  names (Klasse) _ c (paste ("X", 1:p, sep=""), "group")
  Klasse [,p+1] _ as.factor (Klasse [,p+1])
  Klasse
 }
Beispiel5 _ Gen.Beispiel.5 ()

# Generierung zweidimensionaler Datensaetze fuer Graphiken
# --------------------------------------------------
# Beispiele mit gleichen Kovarianzen
# ---------------------------------------------------
Gen.GraphBsp.LDA _ function (n=100) {
  p _ 2
  Klasse1 _ rmultnorm (n, c (-1,-1), diag (rep (1,p)) )
    Klasse1 _ cbind (Klasse1, 1)
  Klasse2 _ rmultnorm (n, c ( 1, 1), diag (rep (1,p)) )
    Klasse2 _ cbind (Klasse2, 2)
  Klasse _ data.frame (rbind (Klasse1, Klasse2) )
  names (Klasse) _ c (paste ("X", 1:p, sep=""), "group")
  Klasse [,p+1] _ as.factor (Klasse [,p+1])
  Klasse
 }

# Beispiele mit verschiedenen Kovarianzen
```

```
# ---------------------------------------------------
Gen.GraphBsp.QDA _ function (n=100, Sigma1=diag (rep (1,2)),
                                    Sigma2=diag (rep (1,2))) {
  Klasse1 _ rmultnorm (n, c (-1,-1), Sigma1 )
    Klasse1 _ cbind (Klasse1, 1)
  Klasse2 _ rmultnorm (n, c ( 1, 1), Sigma2 )
    Klasse2 _ cbind (Klasse2, 2)
  Klasse _ data.frame (rbind (Klasse1, Klasse2) )
  names (Klasse) _ c (paste ("X", 1:2, sep=""), "group")
  Klasse [,3] _ as.factor (Klasse [,3])
  Klasse
 }

# Beispiel 6
# ---------------------------------------------------
GraphBsp.LDA.200 _ Gen.GraphBsp.LDA (n=200)

# Beispiel 7
# ---------------------------------------------------
GraphBsp.QDA.200    _ Gen.GraphBsp.QDA (n=200, diag (rep (1,2)),
                                               diag (rep (2,2)))

# Beispiel 8
# ---------------------------------------------------
GraphBsp.QDA.200.s  _ Gen.GraphBsp.QDA (n=200, diag (rep (1,2)),
                                               matrix (c(1,3,3,1),2,2))

# Beispiel 9
# ---------------------------------------------------
GraphBsp.QDA.200.s2 _ Gen.GraphBsp.QDA(100, matrix (c(2,0,0,5),2,2),
                                            matrix (c(5,2,2,3),2,2))

# Beispiel 10
# ---------------------------------------------------
GraphBsp.QDA.200.s3 _ Gen.GraphBsp.QDA(100, matrix (c(2,0,0,5),2,2),
                                            matrix (c(3,0,0,2),2,2))

# Beispiel 11
# ---------------------------------------------------
GraphBsp.QDA.200.s4 _ Gen.GraphBsp.QDA(100, matrix (c(5,0,0,2),2,2),
                                            matrix (c(1,0,0,4),2,2))
```

Ansprechen aller Teilmengen

```
# ---------------------------------------------------------------------
# Generierung aller Teilmengen als Liste
# ---------------------------------------------------------------------
GenSubsets _ function (Dimension) {
  Mengen _ vector ("list", 2^Dimension-1)
  Mengen [[1]] _ c (1)
  for (i in 2:(2^Dimension-1)) {
    Mengen [[i]] _ NextSubset (Mengen [[i-1]], Dimension)
   }
  Mengen
 }

# ---------------------------------------------------------------------
# Nachfolgende Teilmenge gemaess der Laenge des Vektors aufrufen;
# falls unmoeglich, dann leer setzen
# Bsp. fuer Dimension=2:
# ... -> (2,4) -> (3,4) -> ()
# ---------------------------------------------------------------------
SkipSubset _ function (Vektor, Dimension) {
        x _ c()
        if (length (Vektor)>1) {
            x _ SkipSubset (Vektor [-1], Dimension) # zunaechst weiter
                                                    # "rechts" erhoehen
```

```
            if (!(length (x)==0)) {
                x _ c (Vektor [1], x)
              } else {       # Erhoehung auf letzter Stelle ging nicht;
                             # muss aktuelle Stelle erhoehen
                if (Vektor[1] < Dimension - length (Vektor) + 1) {
                             # letzte Stellen auf (6,7,8,..) setzen
                    x _ c ((Vektor[1]+1) : (Vektor[1]+length (Vektor)))
                  } else {
                    x _ c()
                  }
             }
          } else {
            if (Vektor[1]<Dimension) {
                x _ Vektor + 1
              } else {
                x _ c()
             }
          }
        x                 # Funktionswert uebergeben
      }


# -------------------------------------------------------------------------
# Nachfolgende Teilmenge aufrufen mittels SkipSubset;
# falls unmoeglich, Laenge des Vektors um 1 erhoehen und auf (1,2,3,...)
# setzen
# Bsp. fuer Dimension=4:
# (1) -> ... -> (2,4) -> (3,4) -> (1,2,3) -> (1,2,4) -> (1,3,4) ->
# -> (2,3,4) -> (1,2,3,4) -> ()
# -------------------------------------------------------------------------
NextSubset _ function (Vektor, Dimension) {
    if (length (Vektor)==0) {
        NeuVektor _ c (1)
      } else {
        NeuVektor _ SkipSubset (Vektor, Dimension)
        if (length (NeuVektor)==0 & length (Vektor)<Dimension) {
          NeuVektor _ c (1:(length (Vektor)+1)) # Vektor auf (1,2,... )
                                                # setzen
         }
     }
    NeuVektor             # Funktionswert uebergeben
  }


# ----------------------------------------------------------------------
# Test fuer die Funktionsfaehigkeit
# ----------------------------------------------------------------------
v_0; p_4
cat ("\n")
while (length (v) > 0) { v _ NextSubset (v, p); cat ("# ", v, "\n") }
rm (v,p)
# -- liefert:  --
#  1
#  2
#  3
#  4
#  1 2
#  1 3
#  1 4
#  2 3
#  2 4
#  3 4
#  1 2 3
#  1 2 4
#  1 3 4
#  2 3 4
#  1 2 3 4
```

Risikoschätzungen

Der Quellcode lehnt sich zum Teil an VENABLES & RIPLEY [218], S. 493-494, an.

```
# Bestimmung der Apparent Error Rate (Resubstitutionsmethode)
# ---------------------------------------------------------------------------
Risk.AER _ function (Methode, Daten, ...) {
    if (missing(Methode)) {stop ("Keine Methode angegeben")}
    if (missing(Daten  )) {stop ("Keine Daten angegeben")}

    attach (Daten)
    x _ rep (NA, length (Methode))
    for (i in (1:length(Methode))) {
      x[i] _ FehlKlass (group,
                  VerwendeRegel (BestimmeRegel(Methode[i],
                                   Daten),
                                 Daten, ...))
     }
    detach ("Daten")
    x
 }

# Bestimmung des Risikos ueber Cross-validation
# ---------------------------------------------------------------------------
Risk.CV _ function (Methode, Daten, d=10, ...) {
    if (missing(Methode)) {stop ("Keine Methode angegeben")}
    if (missing(Daten  )) {stop ("Keine Daten angegeben")}

    attach (Daten)
    CVSp _ GenCV (dim (Daten) [1], d)  # Generierung der CV-Zerlegung

    x _ rep (NA, length (Methode))
    for (i in (1:length(Methode))) {
      resultat _ group
      for (j in sort(unique (CVSp))) {
        resultat[CVSp==j] _ VerwendeRegel (BestimmeRegel (Methode[i],
                                 Daten[CVSp!=j,,drop=F]),
                               Daten[CVSp==j,,drop=F], ...)
       }
      x[i] _ FehlKlass (group, resultat)
     }
    detach ("Daten")
    x
 }

# Bestimmung des Risikos ueber Leave-one-out
# ---------------------------------------------------------------------------
Risk.Loo _ function (Methode, Daten, ...) {
    if (missing(Methode)) {stop ("Keine Methode angegeben")}
    if (missing(Daten  )) {stop ("Keine Daten angegeben")}

    Risk.CV (Methode, Daten, d=dim(Daten)[1], ... )
 }

# Bestimmung des Risikos ueber einen Testdatensatz
# ---------------------------------------------------------------------------
Risk.Test _ function (Methode, TrainDaten, TestDaten, ...) {
    if (missing(Methode   )) {stop ("Keine Methode angegeben")}
    if (missing(TrainDaten)) {stop ("Keine Trainingsdaten angegeben")}
    if (missing(TestDaten )) {stop ("Keine Testdaten angegeben")}

    attach (TrainDaten)
    x _ rep (NA, length (Methode))
    for (i in (1:length(Methode))) {
      resultat _ group
      x[i] _ FehlKlass (group,
                  VerwendeRegel (BestimmeRegel(Methode[i],
                                   TrainDaten),
```

```
                             TestDaten, ...))
     }
    detach ("TrainDaten")
    x
 }

# Beste Variablenteilmenge bestimmen (Risiko ueber 10-CV)
# -------------------------------------------------------------------------
Best.Subset _ function (Methode, Daten, d=10, ...) {
    if (missing(Methode   )) {stop ("Keine Methode angegeben")}
    if (missing(Daten     )) {stop ("Keine Daten angegeben")}

    Resultat  _ rep (NA, length (Methode))
    Gruppe _ length (Daten)          # letzte Variable ("group")=
                                     #   Klassenzugehoerigkeit
    Subsets _ GenSubsets (Gruppe-1)  # alle Teilmengen der Variablen
    ZahlSubsets _ length (Subsets)
    Ergebnisse _ matrix (100, nrow=ZahlSubsets,
                             ncol=length (Methode))
    cat ("\n")
    for  (j in 1:ZahlSubsets) {
      cat (".")
      Ergebnisse [j,] _ Risk.CV (Methode,
                                 Daten[,c(Subsets[[j]], Gruppe),drop=F],
                                 d, ...)
     }
    BesteMengen _ apply (Ergebnisse, c(2), argmin)
    for  (j in 1:length(Methode)) {
      Resultat [j] _ Ergebnisse [BesteMengen[j],j]
     }
    list (Variablen=Subsets [BesteMengen], Risiko=Resultat)
 }

# Beste Variablenteilmenge bestimmen (Risiko ueber Test-Datensatz)
# -------------------------------------------------------------------------
Best.Subset.Test _ function (Methode, Daten, TestDaten, d=10, ...) {
    if (missing(Methode   )) {stop ("Keine Methode angegeben")}
    if (missing(Daten     )) {stop ("Keine Daten angegeben")}
    if (missing(TestDaten)) {stop ("Keine TestDaten angegeben")}

    Resultat  _ rep (NA, length (Methode))
    Gruppe _ length (Daten)          # letzte Variable ("group")=
                                     #   Klassenzugehoerigkeit
    Subsets _ GenSubsets (Gruppe-1)  # alle Teilmengen der Variablen
    ZahlSubsets _ length (Subsets)
    Ergebnisse _ matrix (100.00, nrow=ZahlSubsets,
                                 ncol=length (Methode))
    cat ("\n")
    for  (j in 1:ZahlSubsets) {
      cat (".")
      Ergebnisse [j,] _ Risk.Test (Methode,
                                   Daten[,c(Subsets[[j]],
                                            Gruppe),drop=F],
                                   TestDaten[,c(Subsets[[j]],
                                                Gruppe),drop=F],
                                   ...
                                  )
     }
    BesteMengen _ apply (Ergebnisse, c(2), argmin)
    for  (j in 1:length(Methode)) {
      Resultat [j] _ Ergebnisse [BesteMengen[j],j]
     }
    list (Variablen=Subsets [BesteMengen], Risiko=Resultat)
 }

# Berechnung des Prozentsatzes an Fehlklassifikationen (Fehlerrate)
# -------------------------------------------------------------------
```

```
# x=wahre, y=vorhergesagte Zugehoerigkeit
FehlKlass <- function (x, y){
   if (missing(x)|missing(y)) {stop ("Fehlender Vergleichsvektor")}
   tab <- table (x,y)  # Kontingenztafel der Klassenzuordnungen
   diag (tab) <- 0     # korrekte Allokationen werden nicht gezaehlt
   round (100*sum(tab)/length(x),2)    # Ausgabe FKW
 }

# Ausgabe der Fehlklassifikationen aus den einzelnen Populationen
#   sowie den Prozentsatz an Fehlklassifikationen (Fehlerrate)
# ------------------------------------------------------------------
# x=wahre, y=vorhergesagte Zugehoerigkeit
FehlTab <- function (x, y, fkw="Fehlerrate"){
   if (missing(x)|missing(y)) {stop ("Fehlender Vergleichsvektor")}
   cat ("\n")          # Leerzeile
   tab <- table (x,y)  # Kontingenztafel der Klassenzuordnungen
   cat ("\n")          # Leerzeile
   print (tab)         # Ausgabe der Kontingenztafel
   diag (tab) <- 0     # korrekte Allokationen werden nicht gezaehlt
   cat (fkw, " = ", round (100*sum(tab)/length(x),2), "%\n")
                       # Ausgabe FKW
   invisible ()
  }

# Generierung eines Indexfeldes fuer CV bei n Beob.n und d Gruppen
# ------------------------------------------------------------------
GenCV <- function (n, d=10, ss=500){
   if (missing(n)) {stop ("Fehlender Stichprobenumfang")}
   set.seed (ss)            # Angabe des Zufallsmusters (1-1000) (S. 166)
   sample (d,n,replace=T)   # zufaellige Aufteilung der n Beobachtungen
                            # auf d Gruppen (d.h. annaehernd d-CV,
                            # es wird auch Gruppen mit n/d-1 und n/d+2
                            # Beobachtungen geben
 }
```

Ansprechen der Regeln

Der Quellcode der Nearest-Neighbour-Prozedur lehnt sich an eine Prozedur von Polzehl an.

```
# Bestimmung der Regel
# (Uebergabe der gewuenschten Methode als String)
# ------------------------------------------------------------------------
BestimmeRegel _ function (Methode, Daten, ...) {
    if (missing(Methode)) {stop ("Keine Methode angegeben")}
    if (missing(Daten  )) {stop ("Keine Daten angegeben")}
    Regel _ NULL
      if (Methode=="lda"  ) {Regel _ lda (group ~ . , Daten)}   else {
      if (Methode=="qda"  ) {Regel _ qda (group ~ . , Daten)}   else {
      if (Methode=="cart" ) {Regel _ tree(group ~ . , data=Daten,
                                  mindev=0, minsize=2)} else {
      if (Methode=="knn"  ) {Regel _ Daten                    } else {
      if (Methode=="nnet" ) {Regel _ nnet(group ~ . , data=Daten,
                                    size=2, decay=1e-2,
                                    maxit=500)
                                                   } else {
         stop (paste ("Konnte Methode nicht identifizieren: ",Methode))
           }}}}}
    Regel
 }
```

```
# Verwendung der Regel
# (Uebergabe der gewuenschten Regel, Erkennung derselben am
#  class-Attribut)
# ----------------------------------------------------------------------
VerwendeRegel _ function (Regel, Daten,
                          type="class", k=1,
                          ...) {
    if (missing(Regel  )) {stop ("Keine Regel angegeben")}
    if (missing(Daten  )) {stop ("Keine Daten angegeben")}

    cRegel _ class (Regel) [length (class (Regel))]

      if (cRegel=="lda"  ) {predict (Regel, Daten)$class}  else {
      if (cRegel=="qda"  ) {predict (Regel, Daten)$class}  else {
      if (cRegel=="tree" ) {predict (Regel, Daten,
                                     type=type)}         else {
      if (cRegel=="data.frame") {kNN(Regel, Daten, k=k)}   else {
      if (cRegel=="nnet" ) {as.factor (predict (Regel, Daten,
                                                type=type))
       } else {
         stop (paste ("Konnte Regel nicht identifizieren: ",cRegel))
           }}}}}
 }

# k - Nearest Neighbours
# ----------------------------------------------------------------------
# - k-NN-Regel -
kNN _ function(tsample, newsample, k=1){
 dt _ dim(tsample  )[1]
 dn _ dim(newsample)[1]
 p  _ dim (tsample) [2]  # Klassenmerkmal
 class _ tsample [,p, drop=T]
 distmat _ dist2full (dist (rbind(as.matrix (tsample [, -p, drop=F]),
                                 as.matrix(newsample [, -p, drop=F])) )
                    ) [dt+(1:dn), 1:dt, drop=F]
 index _ apply (distmat,1,order)[1:k,,drop=F]  # k naechste Punkte
 result _ numeric (dn)
 for( i in 1:dn) {
   result[i] _ as.numeric (names (
                 table (class [index[,i]])) [
                   table (class[index[,i]]) ==
                     max (table (class [index[,i]]))
                  ]
                 ) [1]
  }
 as.factor (result)
}

# - Uebersetzung einer unteren Dreiecksmatrix in eine volle -
dist2full <- function(dis) {
  n <- attr(dis, "Size")           # Anzahl der Punkte
  full <- matrix(0, n, n)          # Matrix definieren
  full[lower.tri(full)] <- dis     # Belegung nur des unteren Dreiecks
  full + t(full)                   # volle Matrix durch Add. der Transpon.
 }
```

Visualisierung von Trennflächen

Der Code lehnt sich zum Teil an VENABLES & RIPLEY [218], S. 487-493, an.

```
# Zeichnen der Trennflaeche(n) mehrerer bestimmter Regeln
# -------------------------------------------------------------
Trennflaeche _ function (Regeln, Daten,
                         Titel="", UnterTitel="",
                         Feinheit=50, ...) {
    if (missing(Regeln)) {stop ("Keine Regeln angegeben")}
    if (missing(Daten   )) {stop ("Keine Daten angegeben")}

  # malt den Rahmen (ohne Punkte)
```

```
  plot (Daten[,1], Daten[,2], type="n",
        xlab=names (Daten) [1], ylab=names (Daten) [2],
        main=Titel, sub=UnterTitel,
        ... )
  # beschriftet die nicht gezeichneten Punkte mit den Klassen-Labels
  text (Daten[,1], Daten[,2],
        as.character (GraphikSymbol( Daten[,"group"])))
  # Erzeugung des Gitters fuer den Contour-Plot
  x1.seq _ seq (min (Daten[,1]), max (Daten[,1]), length=Feinheit)
  x2.seq _ seq (min (Daten[,2]), max (Daten[,2]), length=Feinheit)
  Gitter _ expand.grid (X1=x1.seq, X2=x2.seq)
  # Erzeugung der Contour-Plots
  for (i in 1:length (Regeln)) {
    # Bestimmung der Klassenzugehoerigkeit
    Werte _ unclass (VerwendeRegel (Regeln[[i]], Gitter))
    contour (x1.seq, x2.seq, matrix (Werte, Feinheit),
               add=T, levels=2, labex=0, lty=i)
   }
  invisible ()
 }

# Graphiksymbole fuer die Klassen
# -------------------------------------------
GraphikSymbol _ function (vektor) {
   v _ character (length (vektor))
   klasse1 _ as.numeric(vektor)==1
   v [ klasse1] _ "+"
   v [!klasse1] _ "o"
   v
 }

# EPS-Datei oeffnen als Graphik-Ausgabe
# ---------------------------------------------------------------
PS.On _ function (Name="unbenannt.eps",
                    Onefile=F,
                    Print.it=F,
                    Append=F,
                    Paper="a4",
                    Horizontal=T,
                    Maximize=T) {
  postscript (file=Name,
              onefile=Onefile,
              print.it=Print.it,
              append=Append,
              paper=Paper,
              horizontal=Horizontal,
              maximize=Maximize
             )
  invisible ()
 }

# EPS-Datei als Graphik-Ausgabe schliessen
# ---------------------------------------------------------------
PS.Off _ function () {
  dev.off ()
  invisible ()
 }

# Bereits am Bildschirm ausgegebenes Bild in EPS-Datei sichern
# ---------------------------------------------------------------
Ausgabe.PS _ function (Name="unbenannt.eps") {
  printgraph(file=Name,
             method="postscript",
             paper="a4", horizontal=F, dpi=300,
             print=F,
             maximize=T)
  invisible ()
 }
```

A.2 Demonstrationen

Tabelle 7.1

```
TabRisikoschaetzung _ function (Methoden=c ("lda","qda","cart", "nnet")) {
    # Namen der zu analysierenden Datensaetze
    Namen _         c ("Beispiel1",
                       "Beispiel2",
                       "Beispiel3",
                       "Beispiel4",
                       "Beispiel5",
                       "Liver disorders"
                      )
    TabSpalten _ c ("AER","2-CV","10-CV", "25-CV", "LOO", "Test")
    FKWs _ data.frame (matrix (NA, nrow=length (Namen),
                                   ncol=length (TabSpalten)),
                         row.names=Namen)
    names (FKWs) _ TabSpalten
    cat ("\n\n")
    cat ("Angegebene Datensaetze: ", length (Namen)   , "\n", sep="")
    cat ("Angegebene Verfahren  : ", length (Methoden), "\n", sep="")
    for (j in 1:length (Methoden)) {
      FKWs [,] _ NA
      for (i in 1:length (Namen)) {
        cat (".")
        AktDaten _ switch (i,
                             Beispiel1,
                             Beispiel2,
                             Beispiel3,
                             Beispiel4,
                             Beispiel5,
                             bupa
                           )
        TestDaten _ switch (i,
                             Beispiel1.Test,
                             Beispiel2.Test,
                             Beispiel3.Test,
                             Beispiel4.Test,
                             Beispiel5.Test
                           )
        FKWs[i,1] _ Risk.AER (Methoden[j], AktDaten)
        FKWs[i,2] _ Risk.CV  (Methoden[j], AktDaten,d=2)
        FKWs[i,3] _ Risk.CV  (Methoden[j], AktDaten)
        FKWs[i,4] _ Risk.CV  (Methoden[j], AktDaten,d=25)
        FKWs[i,5] _ Risk.Loo (Methoden[j], AktDaten)
        if (i <= length (TestDaten)) {
          FKWs[i,6] _ Risk.Test (Methoden[j], AktDaten, TestDaten)
         }
       }
      cat ("\n\n","  Methode = ", Methoden[j], "\n\n", sep="")
      print (FKWs)
     }
    cat ("\n\n")
 }
```

Tabelle 7.2

```
# -- Bestimmung des Risikos ueber 10-CV --
TabVariablenauswahl _ function (Methoden=c ("lda","qda")) {
    # Namen der zu analysierenden Datensaetze
    Namen _         c ("Beispiel1",
                       "Beispiel2",
                       "Beispiel3",
```

```
                    "Beispiel4",
                    "Beispiel5"
                   )
    TabSpalten _ c ("gesamt", "Fehler.G", "beste.Teilmenge", "Fehler.T")
    SatzZahl _ length (Namen)
    MethZahl _ length (Methoden)
    Zahlen _ data.frame (matrix (0, nrow=SatzZahl,
                                ncol=length (TabSpalten)),
                         row.names=Namen )
    names (Zahlen) _ TabSpalten
    cat ("\n\n", "Angegebene Datensaetze: ", SatzZahl , "\n", sep="")
    cat (        "Angegebene Verfahren  : ", MethZahl , "\n", sep="")
    for (j in 1:MethZahl) {
      Zahlen [,] _ 0
      for (i in 1:SatzZahl) {
        AktDaten _ switch (i,
                           Beispiel1,
                           Beispiel2,
                           Beispiel3,
                           Beispiel4,
                           Beispiel5
                         )
        Optimum _ Best.Subset (Methoden[j], AktDaten)
        Zahlen[i,1] _ dim (AktDaten) [2] - 1
        Zahlen[i,2] _ Risk.CV (Methoden[j], AktDaten)
        Zahlen[i,3] _ length ( Optimum$Variablen [[1]] )
        Zahlen[i,4] _ Optimum$Risiko [1]
       }
      cat ("\n\n","  Methode = ", Methoden[j], "\n\n", sep="")
      print (Zahlen)
     }
    cat ("\n\n")
 }

# -- Bestimmung des Risikos ueber Testdatensatz --
TabVariablenauswahl.Test _ function (Methoden=c ("lda","qda")) {
    # Namen der zu analysierenden Datensaetze
    Namen _         c ("Beispiel1",
                    "Beispiel2",
                    "Beispiel3",
                    "Beispiel4",
                    "Beispiel5"
                   )
    TabSpalten _ c ("gesamt", "Fehler.G", "beste.Teilmenge", "Fehler.T")
    SatzZahl _ length (Namen)
    MethZahl _ length (Methoden)
    Zahlen _ data.frame (matrix (0, nrow=SatzZahl,
                                ncol=length (TabSpalten)),
                         row.names=Namen )
    names (Zahlen) _ TabSpalten
    cat ("\n\n", "Angegebene Datensaetze: ", SatzZahl , "\n", sep="")
    cat (        "Angegebene Verfahren  : ", MethZahl , "\n", sep="")
    for (j in 1:MethZahl) {
      Zahlen [,] _ 0
      for (i in 1:SatzZahl) {
        AktDaten _ switch (i,
                           Beispiel1,
                           Beispiel2,
                           Beispiel3,
                           Beispiel4,
                           Beispiel5
                         )
```

```
        TestDaten _ switch (i,
                             Beispiel1.Test,
                             Beispiel2.Test,
                             Beispiel3.Test,
                             Beispiel4.Test,
                             Beispiel5.Test
                           )
        Optimum _ Best.Subset.Test (Methoden[j], AktDaten, TestDaten)
        Zahlen[i,1] _ dim (AktDaten) [2] - 1
        Zahlen[i,2] _ Risk.Test (Methoden[j], AktDaten, TestDaten)
        Zahlen[i,3] _ length ( Optimum$Variablen [[1]] )
        Zahlen[i,4] _ Optimum$Risiko [1]
       }
      cat ("\n\n","  Methode = ", Methoden[j], "\n\n", sep="")
      print (Zahlen)
     }
    cat ("\n\n")
 }

# Index des minimalen Elements eines Vektors
argmin _ function (Vektor){
  ((1:length(Vektor) ) [Vektor==min(Vektor)]) [1]
 }
```

Tabelle 7.3

```
TabBiasVarianz _ function (kmin=1, kmax=10, imax=6) {
    # Namen der zu analysierenden Datensaetze
    Namen _         c ("Beispiel1",
                       "Beispiel2",
                       "Beispiel3",
                       "Beispiel4",
                       "Beispiel5",
                       "Liver disorders"
                      )
    Methode _ "knn"
    TabSpalten _ paste ("k=",kmin:kmax, sep="")
    imax _ min (imax, length (Namen))
    FKWs _ data.frame (matrix (0, nrow=imax,
                               ncol=length (TabSpalten)),
                         row.names=Namen[1:imax])
    names (FKWs) _ TabSpalten
    cat ("\n\n", "Angegebene Datensaetze: ", imax   , "\n", sep="")
    for (j in kmin:kmax) {
      for (i in 1:imax) {
        cat (".")
        AktDaten _ switch (i,
                             Beispiel1,
                             Beispiel2,
                             Beispiel3,
                             Beispiel4,
                             Beispiel5,
                             bupa
                           )
        FKWs[i,(j-kmin+1)] _ Risk.CV  (Methode, AktDaten, k=j)
       }
     }
    cat ("\n\n")
    print (FKWs)
    cat ("\n\n")
 }
```

Tabelle 7.4

```
# -- Bestimmung des Risikos ueber Testdatensatz --
TabOptima _ function (Methoden=c ("lda", "qda", "nnet")) {
    # Namen der zu analysierenden Datensaetze
```

```
    Namen _         c ("Beispiel1",
                       "Beispiel2",
                       "Beispiel3",
                       "Beispiel4",
                       "Beispiel5"
                      )
    TabSpalten _ Methoden
    FKWs _ data.frame (matrix (NA, nrow=length (Namen),
                                   ncol=length (TabSpalten)),
                        row.names=Namen)
    names (FKWs) _ TabSpalten
    cat ("\n\n")
    cat ("Angegebene Datensaetze: ", length (Namen)   , "\n", sep="")
    cat ("Angegebene Verfahren  : ", length (Methoden), "\n", sep="")
    FKWs [,] _ NA
    for (i in 1:length (Methoden)) {
      for (j in 1:length (Namen)) {
        cat (".")
        AktDaten _ switch (j,
                             Beispiel1,
                             Beispiel2,
                             Beispiel3,
                             Beispiel4,
                             Beispiel5
                           )
        TestDaten _ switch (j,
                             Beispiel1.Test,
                             Beispiel2.Test,
                             Beispiel3.Test,
                             Beispiel4.Test,
                             Beispiel5.Test
                           )
        FKWs[j,i] _ Best.Subset.Test (Methoden [i],
                                      AktDaten,
                                      TestDaten)$Risiko
       }
     }
    cat ("\n\n")
    print (FKWs)
    cat ("\n\n")
    FKWs
 }
```

Graphiken im Text

```
# Graphik-Beispiel fuer LDA (Beispiel 6)
# -------------------------------
Graphik.LDA.1 _ function (drucken=F,
                          Name="Graphik.LDA.1.eps",
                          Feinheit=50,
                          Maximize=F) {
  RegelName _ "lda"
  Regel _ BestimmeRegel (RegelName, GraphBsp.LDA.200)
  mUeber _ "LDA"
  sUeber _ ""
  if (drucken) PS.On (Name, Horizontal=F, Maximize=F)
  Trennflaeche (list (xx=Regel),
                GraphBsp.LDA.200,
                Titel=mUeber, UnterTitel= sUeber, Feinheit=Feinheit,
                pty="s")
  if (drucken) PS.Off ()
  cat ("\n", "Geschaetzte FKW: ",
             Risk.CV  (RegelName, GraphBsp.LDA.200), "%", "\n\n",
           sep="")
  invisible ()
 }

# Graphik-Beispiel fuer QDA (Beispiel 7)
# -------------------------------
Graphik.QDA.1 _ function (drucken=F,
                          Name="Graphik.QDA.1.eps",
                          Feinheit=50,
```

```
                         Maximize=F) {
  RegelName _ "qda"
  Regel _ BestimmeRegel (RegelName, GraphBsp.QDA.200)
  mUeber _ "QDA"
  sUeber _ ""
  if (drucken) PS.On (Name, Horizontal=F, Maximize=Maximize)
  Trennflaeche (list (xx=Regel),
               GraphBsp.QDA.200,
               Titel=mUeber, UnterTitel= sUeber, Feinheit=Feinheit,
               pty="s")
  if (drucken) PS.Off ()
  cat ("\n", "Geschaetzte FKW: ",
            Risk.CV  (RegelName, GraphBsp.QDA.200), "%", "\n\n",
          sep="")
  invisible ()
 }
```

Graphiken Verschiedene Trennflächen

```
# Graphik: Vergleich der Trennflaechen (Beispiel 7)
# ------------------------------
Graphik.Vergleich.4 _ function (drucken=F,
                                Name="Graphik.Vergleich.4.eps",
                                Feinheit=50,
                                Maximize=F) {
  Regel.lda  _ BestimmeRegel ("lda" , GraphBsp.LDA.200)
  Regel.qda  _ BestimmeRegel ("qda" , GraphBsp.LDA.200)
  Regel.cart _ BestimmeRegel ("cart", GraphBsp.LDA.200)
  Regel.nnet _ BestimmeRegel ("nnet", GraphBsp.LDA.200)
  if (drucken) PS.On (Name, Horizontal=F, Maximize=Maximize)
  split.screen (figs=c(2,2))
  screen (1);  Trennflaeche (list (r1=Regel.lda),
               GraphBsp.LDA.200,
               Titel="LDA", UnterTitel= "", Feinheit=Feinheit,
               pty="s")
  screen (2);  Trennflaeche (list (r2=Regel.qda),
               GraphBsp.LDA.200,
               Titel="QDA", UnterTitel= "", Feinheit=Feinheit,
               pty="s")
  screen (3);  Trennflaeche (list (r3=Regel.cart),
               GraphBsp.LDA.200,
               Titel="CART", UnterTitel= "", Feinheit=Feinheit,
               pty="s")
  screen (4);  Trennflaeche (list (r4=Regel.nnet),
               GraphBsp.LDA.200,
               Titel="One-Hidden-Layer", UnterTitel= "",
               Feinheit=Feinheit,
               pty="s")
  if (drucken) PS.Off ()
  close.screen (all=T)
  cat ("\n")
  cat ("."); r.lda  _ Risk.CV  ("lda" , GraphBsp.LDA.200)
  cat ("."); r.qda  _ Risk.CV  ("qda" , GraphBsp.LDA.200)
  cat ("."); r.cart _ Risk.CV  ("cart", GraphBsp.LDA.200)
  cat ("."); r.nnet _ Risk.CV  ("nnet", GraphBsp.LDA.200)
  cat ("\n", "Geschaetzte Fehlklassifikationswahrscheinlichkeiten: ",
       "\n", "\n", sep="")
  cat ("LDA : ", r.lda , "%", "\n", sep="")
  cat ("QDA : ", r.qda , "%", "\n", sep="")
  cat ("CART: ", r.cart, "%", "\n", sep="")
  cat ("NNet: ", r.nnet, "%", "\n", sep="")
  cat ("\n")
  invisible ()
 }

# Graphik: Vergleich der Trennflaechen (Beispiel 10)
# ------------------------------
Graphik.Vergleich.6 _ function (drucken=F,
                                Name="Graphik.Vergleich.6.eps",
                                Feinheit=50,
                                Maximize=F) {
```

```
  Regel.lda  _ BestimmeRegel ("lda" , GraphBsp.QDA.200.s3)
  Regel.qda  _ BestimmeRegel ("qda" , GraphBsp.QDA.200.s3)
  Regel.cart _ BestimmeRegel ("cart", GraphBsp.QDA.200.s3)
  Regel.nnet _ BestimmeRegel ("nnet", GraphBsp.QDA.200.s3)
  if (drucken) PS.On (Name, Horizontal=F, Maximize=Maximize)
  split.screen (figs=c(2,2))
  screen (1);  Trennflaeche (list (r1=Regel.lda),
                GraphBsp.QDA.200.s3,
                Titel="LDA", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (2);  Trennflaeche (list (r2=Regel.qda),
                GraphBsp.QDA.200.s3,
                Titel="QDA", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (3);  Trennflaeche (list (r3=Regel.cart),
                GraphBsp.QDA.200.s3,
                Titel="CART", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (4);  Trennflaeche (list (r4=Regel.nnet),
                GraphBsp.QDA.200.s3,
                Titel="One-Hidden-Layer", UnterTitel= "",
                Feinheit=Feinheit,
                pty="s")
  if (drucken) PS.Off ()
  close.screen (all=T)
  cat ("\n")
  cat ("."); r.lda  _ Risk.CV  ("lda" , GraphBsp.QDA.200.s3)
  cat ("."); r.qda  _ Risk.CV  ("qda" , GraphBsp.QDA.200.s3)
  cat ("."); r.cart _ Risk.CV  ("cart", GraphBsp.QDA.200.s3)
  cat ("."); r.nnet _ Risk.CV  ("nnet", GraphBsp.QDA.200.s3)
  cat ("\n", "Geschaetzte Fehlklassifikationswahrscheinlichkeiten: ",
         "\n", "\n", sep="")
  cat ("LDA : ", r.lda , "%", "\n", sep="")
  cat ("QDA : ", r.qda , "%", "\n", sep="")
  cat ("CART: ", r.cart, "%", "\n", sep="")
  cat ("NNet: ", r.nnet, "%", "\n", sep="")
  cat ("\n")
  invisible ()
 }

# Graphik: Vergleich der Trennflaechen (Beispiel 8)
# --------------------------------
Graphik.Vergleich.2 _ function (drucken=F,
                                Name="Graphik.Vergleich.2.eps",
                                Feinheit=50,
                                Maximize=F) {
  Regel.lda  _ BestimmeRegel ("lda" , GraphBsp.QDA.200.s)
  Regel.qda  _ BestimmeRegel ("qda" , GraphBsp.QDA.200.s)
  Regel.cart _ BestimmeRegel ("cart", GraphBsp.QDA.200.s)
  Regel.nnet _ BestimmeRegel ("nnet", GraphBsp.QDA.200.s)
  if (drucken) PS.On (Name, Horizontal=F, Maximize=Maximize)
  split.screen (figs=c(2,2))
  screen (1);  Trennflaeche (list (r1=Regel.lda),
                GraphBsp.QDA.200.s,
                Titel="LDA", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (2);  Trennflaeche (list (r2=Regel.qda),
                GraphBsp.QDA.200.s,
                Titel="QDA", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (3);  Trennflaeche (list (r3=Regel.cart),
                GraphBsp.QDA.200.s,
                Titel="CART", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (4);  Trennflaeche (list (r4=Regel.nnet),
                GraphBsp.QDA.200.s,
                Titel="One-Hidden-Layer", UnterTitel= "",
                Feinheit=Feinheit,
                pty="s")
  if (drucken) PS.Off ()
  close.screen (all=T)
```

```
  cat ("\n")
  cat ("."); r.lda  _ Risk.CV  ("lda" , GraphBsp.QDA.200.s)
  cat ("."); r.qda  _ Risk.CV  ("qda" , GraphBsp.QDA.200.s)
  cat ("."); r.cart _ Risk.CV  ("cart", GraphBsp.QDA.200.s)
  cat ("."); r.nnet _ Risk.CV  ("nnet", GraphBsp.QDA.200.s)
  cat ("\n", "Geschaetzte Fehlklassifikationswahrscheinlichkeiten: ",
         "\n", "\n", sep="")
  cat ("LDA : ", r.lda , "%", "\n", sep="")
  cat ("QDA : ", r.qda , "%", "\n", sep="")
  cat ("CART: ", r.cart, "%", "\n", sep="")
  cat ("NNet: ", r.nnet, "%", "\n", sep="")
  cat ("\n")
  invisible ()
 }

# Graphik: Vergleich der Trennflaechen (Beispiel 9)
# ------------------------------
Graphik.Vergleich.5 _ function (drucken=F,
                                Name="Graphik.Vergleich.5.eps",
                                Feinheit=50,
                                Maximize=F) {
  Regel.lda  _ BestimmeRegel ("lda" , GraphBsp.QDA.200.s2)
  Regel.qda  _ BestimmeRegel ("qda" , GraphBsp.QDA.200.s2)
  Regel.cart _ BestimmeRegel ("cart", GraphBsp.QDA.200.s2)
  Regel.nnet _ BestimmeRegel ("nnet", GraphBsp.QDA.200.s2)
  if (drucken) PS.On (Name, Horizontal=F, Maximize=Maximize)
  split.screen (figs=c(2,2))
  screen (1);  Trennflaeche (list (r1=Regel.lda),
                GraphBsp.QDA.200.s2,
                Titel="LDA", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (2);  Trennflaeche (list (r2=Regel.qda),
                GraphBsp.QDA.200.s2,
                Titel="QDA", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (3);  Trennflaeche (list (r3=Regel.cart),
                GraphBsp.QDA.200.s2,
                Titel="CART", UnterTitel= "", Feinheit=Feinheit,
                pty="s")
  screen (4);  Trennflaeche (list (r4=Regel.nnet),
                GraphBsp.QDA.200.s2,
                Titel="One-Hidden-Layer", UnterTitel= "",
                Feinheit=Feinheit,
                pty="s")
  if (drucken) PS.Off ()
  close.screen (all=T)
  cat ("\n")
  cat ("."); r.lda  _ Risk.CV  ("lda" , GraphBsp.QDA.200.s2)
  cat ("."); r.qda  _ Risk.CV  ("qda" , GraphBsp.QDA.200.s2)
  cat ("."); r.cart _ Risk.CV  ("cart", GraphBsp.QDA.200.s2)
  cat ("."); r.nnet _ Risk.CV  ("nnet", GraphBsp.QDA.200.s2)
  cat ("\n", "Geschaetzte Fehlklassifikationswahrscheinlichkeiten: ",
         "\n", "\n", sep="")
  cat ("LDA : ", r.lda , "%", "\n", sep="")
  cat ("QDA : ", r.qda , "%", "\n", sep="")
  cat ("CART: ", r.cart, "%", "\n", sep="")
  cat ("NNet: ", r.nnet, "%", "\n", sep="")
  cat ("\n")
  invisible ()
 }

# Graphik: Vergleich der Trennflaechen (Beispiel 11)
# ------------------------------
Graphik.Vergleich.7 _ function (drucken=F,
                                Name="Graphik.Vergleich.7.eps",
                                Feinheit=50,
                                Maximize=F) {
  Regel.lda  _ BestimmeRegel ("lda" , GraphBsp.QDA.200.s4)
  Regel.qda  _ BestimmeRegel ("qda" , GraphBsp.QDA.200.s4)
  Regel.cart _ BestimmeRegel ("cart", GraphBsp.QDA.200.s4)
```

```
 Regel.nnet _ BestimmeRegel ("nnet", GraphBsp.QDA.200.s4)
 if (drucken) PS.On (Name, Horizontal=F, Maximize=Maximize)
 split.screen (figs=c(2,2))
 screen (1);  Trennflaeche (list (r1=Regel.lda),
               GraphBsp.QDA.200.s4,
               Titel="LDA", UnterTitel= "", Feinheit=Feinheit,
               pty="s")
 screen (2);  Trennflaeche (list (r2=Regel.qda),
               GraphBsp.QDA.200.s4,
               Titel="QDA", UnterTitel= "", Feinheit=Feinheit,
               pty="s")
 screen (3);  Trennflaeche (list (r3=Regel.cart),
               GraphBsp.QDA.200.s4,
               Titel="CART", UnterTitel= "", Feinheit=Feinheit,
               pty="s")
 screen (4);  Trennflaeche (list (r4=Regel.nnet),
               GraphBsp.QDA.200.s4,
               Titel="One-Hidden-Layer", UnterTitel= "",
               Feinheit=Feinheit,
               pty="s")
 if (drucken) PS.Off ()
 close.screen (all=T)
 cat ("\n")
 cat ("."); r.lda  _ Risk.CV  ("lda" , GraphBsp.QDA.200.s4)
 cat ("."); r.qda  _ Risk.CV  ("qda" , GraphBsp.QDA.200.s4)
 cat ("."); r.cart _ Risk.CV  ("cart", GraphBsp.QDA.200.s4)
 cat ("."); r.nnet _ Risk.CV  ("nnet", GraphBsp.QDA.200.s4)
 cat ("\n", "Geschaetzte Fehlklassifikationswahrscheinlichkeiten: ",
       "\n", "\n", sep="")
 cat ("LDA : ", r.lda , "%", "\n", sep="")
 cat ("QDA : ", r.qda , "%", "\n", sep="")
 cat ("CART: ", r.cart, "%", "\n", sep="")
 cat ("NNet: ", r.nnet, "%", "\n", sep="")
 cat ("\n")
 invisible ()
}
```

Anhang B

Symbole und Abkürzungen

Symbole

Nachfolgend sind die benutzten Symbole in ihrer häufigsten Bedeutung aufgeführt. Abweichungen in Teilkapiteln sind jedoch möglich.

Symbol	*Bedeutung*	*Seite*
Ω	Merkmalsgrundraum	3
δ	Diskriminanzregel	3
k	Anzahl der Klassen	3
i	Klassenindex	
p	Dimensionalität (Anzahl der Merkmale)	3
$\{\Omega_i\}_i$	Zerlegung des Merkmalsgrundraumes durch δ	4
Π_i	Population, Klasse, Kategorie	3
π_i	A-priori-Wahrscheinlichkeit der Klasse Π_i	4
τ_i	A-posteriori-Verteilung der Klasse Π_i	8
f_i	Dichtefunktion der Beobachtungen aus Π_i	4
α_{ij}	Fehlklassifikationswahrscheinlichkeit $\Pi_i \to \Pi_j$	5
c_{ij}	Kosten einer Fehlklassifikation $\Pi_i \to \Pi_j$	6
$\mathcal{R}_i$	Gesamtkosten der Allokationen aller Beobachtungen der Klasse Π_i durch δ	6
$\mathcal{R}$	BAYESsches Risiko	6
δ_0	BAYESsche Regel	7
$\mathcal{S}$	Stichprobe	8
n	Umfang der Stichprobe	8
n_i	Anzahl der Beobachtungen aus Π_i in $\mathcal{S}$	9
j	Index der Beobachtungen	
$-j$	ausgelassene Beobachtung	
x	p-variater Merkmalsvektor eines Objekts	8
x_j	j-te Beobachtung aus $\mathcal{S}$	
x_{i_j}	j-te Beobachtung aus $\Pi_i \cap \mathcal{S}$	9
$\mathcal{X}$	Gruppe von Beobachtungen	14
c	Zellenindex bei Kontingenztafeln diskreter Merkmale	
(j)	Index der Merkmale	
$x_{(j)}$	j-tes Merkmal von x	

Symbol	*Bedeutung*	*Seite*
$\mathcal{M}$	Modell	23
$\mathcal{L}$	Likelihood-Funktion	
μ	Mittelwertsvektor	
σ	Varianz	
Σ, S	Kovarianzmatrix bzw. deren Schätzung	
ω	Gewicht	
$\mathbb{R}$	reelle Achse $(-\infty, \infty)$	
I_p	p-variate Einheitsmatrix diag $(1, \ldots, 1)$	
$\mathcal{M}_{p \times p}$	Raum der $p \times p$-Matrizen	
log	natürlicher Logarithmus $\log_{\mathrm{e}}$	
Φ	Standardnormalverteilung N (0,1)	
ϕ	Dichte der Standardnormalverteilung N (0,1)	
Δ^2	MAHALANOBIS-Abstand	13
ρ	MINKOWSKI-Metrik	21
I_η	Indikatorfunktion	13
$\delta_{x,y}$	KRONECKER-Symbol: $\delta_{x,y} = 1 \iff x = y$, sonst 0	36
Δ	Abstand oder Änderung einer Größe	

Abkürzungen

Im folgenden sind die in der Arbeit benutzten Abkürzungen für Schätzungen und Diskriminanzanalyse-Verfahren mit den Kapiteln ihrer Erläuterung aufgeführt.

Name	*steht für*	*Kapitel*
FKW	Fehlklassifikationswahrscheinlichkeit	2.1
ML	Maximum-Likelihood-Schätzung	2.2.1
AER	Apparent Error Rate (Resubstitutionsschätzung)	2.2.3
CV	Cross-validation-Schätzung	2.2.3
LOO	Leave-one-out-Schätzung	2.2.3
BS	Bootstrap-Schätzung	2.2.3
MDS	Mehrdimensionale Skalierung	2.4.2
CCA	Kanonische Korrelationsanalyse	2.4.2
LDA	Lineare Diskriminanzanalyse	3.2
PDA	Penalized Discriminant Analysis	3.2.3
QDA	Quadratische Diskriminanzanalyse	3.3
CPC	Common Principal-Component Model	3.3.3
SIMCA	Soft Independent Modeling of Class Analogy	3.3.3
DASCO	Discriminant Analysis with Shrunken Covariances	3.3.3
RDA	Regularisierte Diskriminanzanalyse	3.3.3
LGA	Logistische Diskriminanzanalyse	3.4
MNA	Multinomialmodell	4.1.1
UMA	Unabhängigkeitsmodell	4.1.1
LOM	Lokationsmodell	4.1.2
k-NN	k-Nearest-Neighbours	4.3
FMNN	Flexible Metric Nearest Neighbour	4.3.3
DANN	Discriminant Adaptive Nearest Neighbour	4.3.3
CART	Classification and Regression Trees	4.4
FACT	Fast Algorithm for Classification Trees	4.4.3
DART		4.4.3
NNet	Neuronale Netze	4.5
GLM	Verallgemeinertes Lineares Modell	5.2
AM	Additives Modell	5.2
GAM	Verallgemeinertes Additives Modell	5.2
FDA	Flexible Diskriminanzanalyse	5.2.2
MDA	Mixture Discriminant Analysis	5.2.2
PP	Projection Pursuit	5.3
SVM	Support Vector Machine	5.4
ECOC	Error Coding Output Correction	6.3
PICT	Plug-In Classification Technique	6.3

Literaturverzeichnis

[1] ABRAMSON, IAN S. (1982). On bandwidth variation in kernel estimates — A square root law. *Ann. Statist.*; **10**: 4, 1217-1223.

[2] AITCHISON, J.; DUNSMORE, I.R. (1975). *Statistical Prediction Analysis.* Cambridge University Press: Cambridge.

[3] AITCHISON, J.; AITKEN, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*; **63**: 3, 413-420.

[4] AITCHISON, J.; BEGG, C.B. (1976). Statistical diagnosis when basic cases are not classified with certainty. *Biometrika*; **63**: 1, 1-12.

[5] AITKEN, C.G.G. (1978). Methods of discrimination in multivariate binary data. In *Compstat 1978, Proc. Computational Statistics.* Physika-Verlag: Wien, pp. 155-161.

[6] ANDERSON, J.A. (1972). Separate sample logistic discrimination. *Biometrika*; **59**: 1, 19-35.

[7] ANDERSON, J.A. (1974). Diagnosis by logistic discrimination function: Further practical problems and results. *Appl. Statist.*; **23**, 397-404.

[8] ANDERSON, J.A. (1975). Quadratic logistic discrimination. *Biometrika*; **62**: 1, 149-154.

[9] ANDERSON, J.A. (1982). Logistic discrimination. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 169-191.

[10] ANDERSON, T.W. (1984). *An Introduction to Multivariate Analysis.* Wiley: New York.

[11] BAUER, ERIC; KOHAVI, RON (1998). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning*, to appear. (available at `www.cs.stanford.edu`)

[12] BAUER, HEINZ (1991). *Wahrscheinlichkeitstheorie.* de Gruyter: Berlin.

[13] BAUER, HEINZ (1992). *Maß- und Integrationstheorie.* de Gruyter: Berlin.

[14] BELLMAN, R.E. (1961). *Adaptive Control Processes.* Princeton University Press: Princeton.

[15] BENZÉCRI, J.-P. (1992). *Correspondence Analysis Handbook.* Statistics: Textbooks and Monographs. Dekker: New York.

[16] BHATTACHARYYA, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*; **35**, 99-109.

[17] BISHOP, CHRISTOPHER M. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press: Oxford.

[18] BISHOP, YVONNE M.M.; FIENBERG, STEPHEN E.; HOLLAND, PAUL W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press: Cambridge, Massachusetts.

[19] Borg, Ingwer; Groenen, Patrick (1997). *Modern Multidimensional Scaling: Theory and Applications.* Springer Series in Statistics. Springer-Verlag: New York.

[20] Bowman, A.W. (1985). A comparative study of some kernel-based nonparametric density estimators. *J. Statist. Comput. Simulation*; **21**, 313-327.

[21] Breiman, L.; Friedman, J.H.; Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth: Belmont, California.

[22] Breiman, Leo (1991). The π-method for estimating multivariate functions from noisy data. *Technometrics*; **33**: 2, 125-160.

[23] Breiman, Leo (1996). Bagging predictors. *Machine Learning*; **24**: 2, 123-140.

[24] Breiman, Leo (1997a). Arcing the edge. Technical Report 486, Statistics Dept., University of California, Berkeley. (available at `www.stat.berkeley.edu`)

[25] Breiman, Leo (1997b). Prediction games and arcing algorithms. Technical Report 504, Statistics Dept., University of California, Berkeley. (available at `www.stat.berkeley.edu`)

[26] Breiman, Leo (1998). Arcing classifiers (with discussion). *Ann. Statist.*; **26**: 3, 801-849.

[27] Breiman, Leo (1999). Using adaptive bagging to debias regressions. Technical report, Statistics Dept., University of California, Berkeley. (available at `www.stat.berkeley.edu`)

[28] Breiman, Leo; Friedman, Jerome H. (1988). Contribution to the discussion of paper by Loh & Vanichsetakul. *J. Amer. Statist. Assoc.*; **83**: 2, 725-727.

[29] Breiman, L.; Meisel, W.; Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*; **19**, 135-144.

[30] Bridle, J.S. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in Neural Information Processing Systems* (Vol. 2), D. Touretzky (Ed.).

[31] Bunke, Helga; Droge, Bernd (1985). A stepwise procedure for the selection of nonlinear regression models. *Statistics*; **16**: 1, 35-45.

[32] Bunke, Olaf (1985a). An adaptive smoothing estimator for probabilities in contingency tables. *Statistics*; **16**: 1, 55-62.

[33] Bunke, Olaf; Fischer, Klaus (1985b). Some fundamentals and procedures of parametric, distribution-free and discrete discrimininant analysis. *Statistics*; **16**: 2, 185-201.

[34] Bunke, Olaf; Fischer, Klaus; Grund, Birgit (1988). On parametric and nonparametric procedures in discriminant analysis for mixed continuous and categorical variables. *Statistische Methoden der Modellbildung*; 47-67.

[35] Bunke, Olaf; Droge, Bernd; Polzehl, Jörg (1995a). Selection of regression and variance models in nonlinear regression. Discussion paper, SFB 373, Humboldt-Universität zu Berlin. (available at `sfb.wiwi.hu-berlin.de`)

[36] Bunke, Olaf; Droge, Bernd; Polzehl, Jörg (1995b). Model selection, transformations and variance estimation in nonlinear regression. *Statistics*, to appear. (available at `sfb.wiwi.hu-berlin.de`)

[37] Bunke, Olaf; Droge, Bernd; Polzehl, Jörg (1995c). Splus tools for model selection in nonlinear regression. *Computational Statistics*; **13**, 257-281.

[38] Buntine, W.L. (1992). Learning classification trees. *Statistics and Computing*; **2**, 63-73.

[39] Burman, P. (1993b). A comparative study of ordinary cross-validation, v hold cross-validation and the repeated learning-testing methods. *Biometrika*; **76**, 503-514.

[40] Chambers, J.M.; Hastie, J.H. (Eds.) (1992). *Statistical Models in S*. Wadsworth: Pacific Grove, California.

[41] Chipman, H.A.; George, E.I.; McCulloch, R.E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.*; **93**: 443, 935-948.

[42] Cleveland, W.S.; Devlin, S.J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.*; **83**, 596.

[43] Cox, T.F.; Ferry, G. (1991). Robust logistic discrimination. *Biometrika*; **78**: 4, 841-849.

[44] Cwik, J.; Mielniczuk, J. (1995). Nonpararmetric rank discrimination method. *J. Comput. Stat. Data Anal.*; **19**: 1, 59-74.

[45] Cybenko, G. (1988). Continuous valued neural networks with two hidden layers are sufficient. Technical Report, Dept. of Computer Science, Tuffs University.

[46] Davison, A.C.; Hall, P. (1992). On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems. *Biometrika*; **79**: 2, 279-284.

[47] Denison, D.G.T.; Mallick, B.K.; Smith, A.F.M. (1998). A bayesian CART algorithm. *Biometrika*; **85**: 2, 363-377.

[48] Devroye, L.; Wagner, T.J. (1982). Nearest neighbor methods in discrimination. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 193-197.

[49] Devroye, Luc (1985). A note on the L_1 consistency of variable kernel estimates. *Ann. Statist.*; **13**: 3, 1041-1049.

[50] Dietterich, Thomas G. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, to appear. (available at `ftp.cs.orst.edu/pub/tgd/papers`)

[51] Dietterich, T.; Bakiri, G. (1991). Error-correcting output codes: A general method for improving multiclass inductive learning problems. In *Proc. AAAI-91*: AAAI Press / MIT Press, pp. 572-577.

[52] Dietterich, T.; Bakiri, G. (1995a). Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res.*; **2**, 263-286.

[53] Dietterich, T.G.; Kong, E.B. (1995b). Machine learning bias, statistical bias and statistical variance of decision trees: algorithms. Technical report. Department of Computer Science, Oregon State University, Corvallis, Oregon. (available at `ftp.cs.orst.edu/pub/tgd/papers`)

[54] Dillon, William R.; Goldstein, Matthew (1984). *Multivariate Analysis: Methods and Applications*. Wiley series in Probability and Mathematical Statistics. Wiley: New York.

[55] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*; **7**, 1-26.

[56] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM: Philadelphia.

[57] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvements on cross-validation. *J. Amer. Statist. Assoc.*; **78**, 316-331.

[58] Efron, B.; Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall: London & New York.

[59] Enas, G.G.; Choi, S.C. (1986). Choice of the smoothing parameter and efficiency of k-nearest neighbour classification. *Comput. Math. Applic.*; **12A**, 235-244.

[60] Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density. *Theor. Prob. Appl.*; **14**, 153-158.

[61] Everitt, Brian (1974). *Cluster Analysis.* Heinemann Educational Books: London.

[62] Fahlman, S.E.; Lebiere, C. (1990). The cascade-correlation learning algorithm. In *Advances in Neural Information Processing 2*, D.S. Touretzky (Ed.). Morgan Kaufmann Publishers, San Mateo, CA.

[63] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*; **7**, 179-188.

[64] Fix, E.; Hodges, J. (1951). Discriminatory analysis: nonparametric discrimination: consistency properties. Rept. No. 4, USAF School of Aviation Medicine, Randolph Field, TX.

[65] Flury, B. (1984). Common principal components in k groups. *J. Amer. Statist. Assoc.*; **79**, 892-898.

[66] Flury, B.; Boukai, B.; Flury, B.D. (1997). The discrimination subspace model. *J. Amer. Statist. Assoc.*; **92**: 438, 758-766.

[67] Freund, Yoav; Schapire, Robert E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, Saitta, L. (Ed.). Morgan Kaufmann: San Francisco, pp. 148-156. (available at `www.research.att.com`)

[68] Freund, Yoav; Schapire, Robert E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*; **55**, 119-139.

[69] Freund, Y.; Iyer, R.; Schapire, R.E.; Singer, Y. (1998). A efficient boosting algorithm for combining preferences. In *Proceedings of the Fifteenth International Conference on Machine Learning.* (available at `www.research.att.com`)

[70] Friedman, Jerome H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*; **84**, 165-175.

[71] Friedman, Jerome H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*; **19**: 1, 1-141.

[72] Friedman, Jerome H. (1993). An overview of predictive learning and function approximation. In *From Statistics to Neural Networks: Theory and Pattern Recognition Applications.* V. Cherkassky, J.H. Friedman and H. Wechsler (Eds.). NATO ASI Series. Series F, Computer and System Sciences; vol. 136. Springer: Berlin, Heidelberg 1994, pp. 1-61.

[73] Friedman, Jerome H. (1994). Flexible metric nearest neighbor classification. Technical Report, Dept. of Statistics and Stanford Linear Accelerator Center, Stanford University. (available at `www.stat.stanford.edu/~jhf`)

[74] Friedman, Jerome H. (1996a). On bias, variance, 0/1-loss, and the curse of dimensionality. Technical Report, Dept. of Statistics and Stanford Linear Accelerator Center, Stanford University. (available at `www.stat.stanford.edu/~jhf`)

[75] Friedman, Jerome H. (1996b). Local learning based on recursive covering. Technical Report, Dept. of Statistics and Stanford Linear Accelerator Center, Stanford University. (available at `www.stat.stanford.edu/~jhf`)

[76] Friedman, Jerome H. (1996c). Another approach to polychotomous classification. Technical Report, Dept. of Statistics and Stanford Linear Accelerator Center, Stanford University. (available at `www.stat.stanford.edu/~jhf`)

[77] Friedman, Jerome H. (1999a). Greedy function approximation: a gradient boosting machine. Technical Report, CSIRO CMIS. (available at `www.stat.stanford.edu/~jhf`)

[78] Friedman, Jerome H. (1999b). Stochastic gradient boosting. Technical Report, CSIRO CMIS. (available at `www.stat.stanford.edu/~jhf`)

[79] Friedman, Jerome H.; Hastie, Trevor; Tibshirani, Robert (1998). Additive logistic regression: a statistical view of boosting. Technical Report, Dept. of Statistics, Stanford University. (available at `www.stat.stanford.edu/~jhf`)

[80] Friedman, J.; Stuetzle, W.; Schroeder, A. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*; **C23**, 881-890.

[81] Friedman, J.H.; Tukey, J.W. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.*; **79**, 599-608.

[82] Geisser, S. (1967). Estimation associated with linear discriminants. *Ann. Math. Statist.*; **38**, 807-817.

[83] Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*; **70**, 320-328.

[84] Geisser, S. (1982). Bayesian discrimination. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 101-120.

[85] Gelfand, S.B.; Ravishankar, C.S.; Delp, E.J. (1991). An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*; **13**, 163-174.

[86] Glick, N. (1972). Sample-based classification procedures derived from density estimators. *J. Amer. Statist. Assoc.*; **67**, 116-122.

[87] Glick, N. (1973a). Sample-based multinomial classification. *Biometrics*; **29**, 241-256.

[88] Gnanadesikan, R.; Kettenring, J. (Eds.) (1989). Discriminant analysis and clustering. *Statistical Science*; **4**, 34-69.

[89] Göhler, Wilhelm (Zsmst.) (1987). *Höhere Mathematik: Formeln und Hinweise.* Deutscher Verlag für Grundstoffindustrie: Leipzig.

[90] Goldstein, M.; Rabinowitz, M. (1975). Selection of variates for the two-group multinomial classification problem. *J. Amer. Statist. Assoc.*; **70**, 776-781.

[91] Green, P.J.; Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman & Hall: London.

[92] Grove, Adam J.; Schuurmans, Dale (1998). Boosting in the limit: Maximizing the margin of learned ensembles. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, to appear.

[93] Grund, Birgit (1986). Schätzungen für Zellwahrscheinlichkeiten in multinomialverteilten Kontingenztafeln. Dissertation (A); Humboldt-Universität zu Berlin.

[94] Grund, Birgit (1993). Kernel estimates for cell probabilities. *J. Multivariate Analysis*; **46**, 283-308.

[95] Grund, Birgit; Hall, Peter (1993). On the performance of kernel estimators for high-dimensional, sparse binary data. *J. Multivariate Analysis*; **44**, 321-344.

[96] Habbema, J.D.F.; Hermans, J.; Burgt, A.T. van der (1974). Cases of doubt in allocation problems. *Biometrika*; **61**: 2, 313-324.

[97] Hall, Peter (1981a). On nonparametric multivariate binary discrimination. *Biometrika*; **68**, 287-294.

[98] Hall, Peter (1981b). Optimal near neighbour estimator for use in discriminant analysis. *Biometrika*; **68**, 572-575.

[99] Hall, Peter (1990). On the bias of variable bandwidth curve estimators. *Biometrika*; **77**: 3, 529-535.

[100] Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Ann. Statist.*; **15**, 1491-1519.

[101] Hall, P.; Marron, J.S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.*; **15**: 1, 163-181.

[102] Hall, P.; Marron, J.S. (1988). Choice of kernel order in density estimation. *Ann. Statist.*; **16**: 1, 161-173.

[103] Hall, P.; Sheather, S.J.; Jones, M.C.; Marron, J.S. (1992). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*; **78**: 2, 263-269.

[104] Hall, Peter; Wand, Matthew P. (1988). On parametric discrimination using density differences. *Biometrika*; **75**: 3, 541-547.

[105] Hand, D.J. (1981). *Discrimination and Classification.* Wiley series in Probability and Mathematical Statistics. Wiley: Chichester.

[106] Hand, D.J. (1982). *Kernel Discrimination Analysis.* Wiley: New York.

[107] Hand, D.J. (1983). A comparison of two methods of discriminant analysis applied to binary data. *Biometrics*; **39**, 683-694.

[108] Hand, D.J. (1986a). Recent advances in error rate estimation. *Pattern Recognition Letters*; **4**, 335-346.

[109] Härdle, Wolfgang (1990). *Applied Nonparametric Regression.* Cambridge University Press: Cambridge.

[110] Härdle, Wolfgang (1991). *Smoothing techniques: with implementation in S.* Springer Series in Statistics. Springer-Verlag: New York.

[111] Hartung, Joachim (1995). *Statistik. Lehr- und Handbuch der angewandten Statistik; mit zahlreichen, vollständig durchgerechneten Beispielen.* Oldenbourg: München.

[112] Hassibi, B.; Stork, D.G.; Solla, S.A. (1993). Second order derivates for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing 5*, S.J. Hanson, J.D. Cowan and C.L. Giles (Eds.). Morgan Kaufmann Publishers: San Mateo, CA, pp. 164-171.

[113] Hastie, Trevor (1989). Discussion of „Flexible parsimonious smoothing and additive modelling“ by Friedman & Silverman. *Technometrics*; **3**, 3-39.

[114] Hastie, Trevor; Buja, Andreas; Tibshirani, Robert (1995). Penalized discriminant analysis. *Ann. Statist.*; **23**: 1, 73-102.

[115] Hastie, Trevor; Tibshirani, Robert (1990). *Generalized additive models.* Monographs on Statistics and Applied Probability 43. Chapman & Hall: London, New York.

[116] Hastie, Trevor; Tibshirani, Robert (1993). Nonparametric regression und classification. In *From Statistics to Neural Networks: Theory and Pattern Recognition Applications.* V. Cherkassky, J.H. Friedman and H. Wechsler (Eds.). NATO ASI Series. Series F, Computer and System Sciences; vol. 136. Springer: Berlin, Heidelberg 1994, pp. 62-82.

[117] Hastie, Trevor; Tibshirani, Robert (1994a). Discriminant analysis by Gaussian mixtures. AT & T Bell Laboratories and University of Toronto.

[118] Hastie, Trevor; Tibshirani, Robert (1994b). Discriminant adaptive nearest neighbor classification. Dept. of Statistics & Div. of Biostatistics, Stanford University.

[119] Hastie, Trevor; Tibshirani, Robert (1998). Classification by pairwise coupling. *Ann. Statist.*; **26**: 2, 451-471.

[120] Hastie, Trevor; Tibshirani, Robert; Buja, Andreas (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.*; **89**: 428, 1255-1270.

[121] Hellman, M.E. (1970). The nearest neighbour classification rule with a reject option. *IEEE Trans. Syst. Science Cybern.*; **SSC-6**, 179-185.

[122] Helmbold, David P.; Schapire, Robert E. (1997). Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*; **27**: 1, 51-68. (available at `www.research.att.com`)

[123] Highleyman, W.H. (1962). The design and analysis of pattern recognition experiments. *Bell Syst. Tech. J*; **41**, 723-744.

[124] Hills, M. (1966). Allocation rules and their error rates (with discussion). *J. R. Statist. Soc.*; **B 28**, 1-31.

[125] Huber, P.J. (1985). Projection pursuit (with discussion). *Ann. Statist.*; **13**: 2, 435-525.

[126] Humak, K.M.S. (1977). *Statistische Methoden der Modellbildung, Band I: Statistische Inferenz für lineare Parameter.* Akademie-Verlag: Berlin.

[127] Izenman, A.J. (1991). Recent developments in nonparametric density estimation (review paper). *J. Amer. Statist. Assoc.*; **86**: 413, 205-224.

[128] Jain, A.K.; Chandrasekaran, B. (1982). Dimensionality and sample size considerations in pattern recognition practice. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 835-855.

[129] James, Gareth; Hastie, Trevor (1997). Generalizations of the bias / variance decomposition for prediction error. Dept. of Statistics, Stanford University.

[130] James, Gareth; Hastie, Trevor (1998). The error coding method and PICTs. *J. Comput Graph Statistics*; **7**: 3, 377-387.

[131] Johnson, W.H.; Wichern D.W. (1988). *Applied Multivariate Statistical Analysis.* Prentice Hall Series in Statistics. Prentice-Hall: Englewood Cliffs NJ.

[132] Jones, M.C.; Marron, J.S.; Sheater S.J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*; **11**, 337-381.

[133] Kaatsch, Simone (1992). Ein C-Programm zur Diskriminanzanalyse für gemischt stetige und diskrete Merkmale. Diplomarbeit; Humboldt-Universität zu Berlin.

[134] Klinke, S.; Grassmann, J. (1998). Projection pursuit regression and neural networks. Discussion paper 980017, SFB 373, Humboldt-Universität zu Berlin. (available at `sfb.wiwi.hu-berlin.de`)

[135] Kong, Eun Bae; Dietterich, Thomas G. (1995). The error-correcting output coding corrects bias and variance. In *Submitted to the International Conference on Machine Learning*. (available at `ftp.cs.orst.edu/pub/tgd/papers`)

[136] Kopka, Helmut (1996). *LaTeX: Einführung, Band 1.* 2. überarb. Aufl. Addison-Wesley: Bonn.

[137] Krishnaiah, P.R. (1982). Selection of variables in discriminant analysis. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 883-892.

[138] Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Assoc.*; **70**, 782-790.

[139] Krzanowski, W.J. (1977). The performance of Fishers linear discriminant function under non-optimal conditions. *Technometrics*; **19**: 2, 191-200.

[140] Krzanowski, W.J. (1979). Some linear transformations for mixtures of binary and continuous variables, with particular reference to linear discriminant analysis. *Biometrika*; **66**, 33-39.

[141] Krzanowski, W.J. (1983a). Distance between populations using mixed continuous and categorical variables. *Biometrika*; **70**, 235-243.

[142] Kullback, S.; Leibler, A. (1951). On information and sufficiency. *Ann. Math. Statist.*; **22**, 79-86.

[143] Lachenbruch, P.A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics*; **24**, 823-834.

[144] Lachenbruch, P.A. (1975). *Discriminant Analysis.* Hafner Press: New York.

[145] Lachenbruch, P.A. (1975). Zero-mean difference discrimination and the absolute linear discriminant function. *Biometrika*; **62**: 2, 397-401.

[146] Läuter, Henning (1985). An efficient estimator for the error rate in discriminant analysis. *Statistics*; **16**: 1, 107-119.

[147] Läuter, H.; Thiele, H. (1985). Optimal combinations of classification procedures. *Statistics*; **16**, 389-406.

[148] Läuter, Jürgen (1985). Discriminant analysis under parameter restrictions — Statistical and computational aspects. *Statistics*; **16**: 1, 125-137.

[149] Läuter, Jürgen (1992). *Stabile multivariate Verfahren: Diskriminanzanalyse – Regressionsanalyse – Faktoranalyse.* Akademie-Verlag: Berlin.

[150] Laird, Nan (1982). The EM algorithm. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 509-520.

[151] LeCan, Y.; Denker, J.S.; Solla, S.A. (1990). Optimal brain damage. In *Advances in Neural Information Processing 2*, D.S. Touretzky (Ed.). Morgan Kaufmann Publishers, San Mateo, CA.

[152] Lindtner, Oliver (1998). Aspekte des Lernens mit qualitativen und gemischten Attributen im Data-Mining. Diplomarbeit; Humboldt-Universität zu Berlin.

[153] Loftsgaarden, D.O.; Quesenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*; **36**, 1049-1051.

[154] Loh, W.-Y.; Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *J. Amer. Statist. Assoc.*; **83**: 2, 715-728.

[155] Mammen, Enno; Tsybakov, Alexandre B. (1998). Smooth discrimination analysis. Discussion paper, SFB 373, Humboldt-Universität zu Berlin. (available at `sfb.wiwi.hu-berlin.de`)

[156] Mardia, K.V.; Kent, J.T.; Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press: London.

[157] Marron, J.S. (1987). A comparison of cross-validation techniques in density estimation. *Ann. Statist.*; **15**: 1, 152-162.

[158] McKay, R.J. (1977). Simultaneous procedures for variable selection in multiple discriminant analysis. *Biometrika*; **64**, 283-290.

[159] McLachlan, Geoffrey J. (1974). The asymptotic distributions of the conditional error rate in discriminant analysis. *Biometrika*; **61**: 1, 131-135.

[160] McLachlan, Geoffrey J. (1976). The bias of the apparent error rate in discriminant analysis. *Biometrika*; **63**, 2: 239-44.

[161] McLachlan, Geoffrey J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley: New York.

[162] Michie, D.; Spiegelhalter, D.J.; Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood series in artificial intelligence. Ellis Horwood: New York.

[163] Miller, Rupert G. (1974). The jackknife - a review. *Biometrika*; **61**: 1, 1-15.

[164] MÖHNER, M. (1986). A comparative study of estimator for probabilities in contingency tables. *Statistics*; **17**: 4, 557-568.

[165] MOODY, JOHN (1993). Prediction risk and architecture selection of neural networks. In *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*. V. Cherkassky, J.H. Friedman and H. Wechsler (Eds.). NATO ASI Series. Series F, Computer and System Sciences; vol. 136. Springer: Berlin, Heidelberg 1994, pp. 147-165.

[166] MORRIS, J.E.; LAYCOCK, P.J. (1974). Discriminant analysis of directional data. *Biometrika*; **61**: 2, 335-341.

[167] MÜLLER, P.H. (Hrsg.) (1975). *Wahrscheinlichkeitsrechnung und Mathematische Statistik: Lexikon der Stochastik*. Akademie-Verlag: Berlin.

[168] NARENDA, P.M.; FUKUNAGA, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Trans. Comput.*; **26**: 9, 917-922.

[169] OTT, J.; KRONMAL, R.A. (1976). Some classification procedures for multivariate binary data using orthogonal functions. *J. Amer. Statist. Assoc.*; **71**, 391-399.

[170] POLZEHL, JÖRG (1993). Projection pursuit discriminant analysis. Discussion paper 9320, Center for Operations Research & Econometrics (CORE), Universite Catholique du Louvain, Belgium.

[171] POLZEHL, JÖRG. Vorlesung „Multivariate Analyse". Sommersemester 1997 an der Humboldt-Universität zu Berlin. Private Mitschriften.

[172] POSSE, C. (1992). Projection pursuit discriminant analysis for two groups. *Communications in Statistics, A - Theory and Methods*; **21**, 1-19.

[173] PRESS, S.J.; WILSON, S. (1978). Choosing between logistic regression and discriminant analysis. *J. Amer. Statist. Assoc.*; **73**, 699-705.

[174] QUINLAN, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann: San Mateo, CA.

[175] QUINLAN, J.R. (1996). Bagging, boosting and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*. AAAI Press / MIT Press: Cambridge, MA, pp. 725-730.

[176] RAO, C.R. (1949). On some problems arising out of discrimination with multiple characters. *Sankhyã*; **9**, 343-366.

[177] REED, R. (1993). Pruning algorithms — a survey. *IEEE Transactions on Neural Networks*; **4**: 740-747.

[178] REMME, J.; HABBEMA, J.D.F.; HERMANS, J. (1980). A simulative comparison of linear, quadratic and kernel discrimination. *J. Statist. Comput. Simulation*; **11**, 87-106.

[179] RIGBY, ROBERT A. (1997). Bayesian discrimination between two multivariate normal populations with equal covariance matrices. *J. Amer. Statist. Assoc.*; **92**: 439, 1151-1154.

[180] RIPLEY, BRIAN D. (1994). Neural networks and related methods for classification. *J. R. Statist. Soc.*; **B 56**: 3, 409-456.

[181] RIPLEY, BRIAN D. (1996). *Pattern recognition and neural networks*. Cambridge University Press: Cambridge.

[182] ROJAS, R. (1996). *Theorie der neuronalen Netze: Eine systematische Einführung*. 4. korrig. Nachdruck. Springer: Berlin.

[183] SACHS, LOTHAR (1992). *Angewandte Statistik. Anwendung statistischer Methoden; mit 291 Tabellen*. Springer: Berlin, Heidelberg.

[184] Sain, Stephan R.; Scott, David W. (1996). On locally adaptive density estimation. *J. Amer. Statist. Assoc.*; **91**: 436, 1525-1534.

[185] Schaafsma, W. (1982). Selecting variables in discriminant analysis for improving upon classical procedures. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 857-881.

[186] Schapire, Robert E. (1997). Using output codes to boost multiclass learning problems. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pp. 313-321. (available at `www.research.att.com`)

[187] Schapire, Robert E. (1999). Theoretical views of boosting. In *Computational Learning Theory: Fourth European Conference, EuroCOLT 99*, to appear. (available at `www.research.att.com`)

[188] Schapire, Robert E.; Singer, Yoram (1998a). Improved boosting algorithm using confidence-rated predictions. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory.* (available at `www.research.att.com`)

[189] Schapire, Robert E.; Freund, Yoav; Bartlett, Peter; Lee, Wee Sun (1998b). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.*; **26**: 5, 1651-1686.

[190] Schimek, Michael G.; Turlach, Berwin A. (1998). Additive and generalized additive models: a survey. Discussion paper 980097, SFB 373, Humboldt-Universität zu Berlin. (available at `sfb.wiwi.hu-berlin.de`)

[191] Schmitz, P.I.M.; Habbema, J.D.F.; Hermans, J. (1983a). The performance of four discriminant analysis methods for mixtures of continuous and binary variables. *Statist. Med.*; **2**, 199-205.

[192] Schmitz, P.I.M.; Habbema, J.D.F.; Hermans, J.; Raatgever, J.W. (1983b). Comparative performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods. *Commun. Statist.-Simula. Computa.*; **12**, 727-751.

[193] Schmitz, P.I.M.; Habbema, J.D.F.; Hermans, J. (1985). A simulation study of the performance of five discriminant analysis methods for mixtures of continuous and binary variables. *J. Statist. Comput. Simulation*; **23**, 69-95.

[194] Schulze, Katrin (1993). Simulationsstudien zum Vergleich verschiedener Diskriminanzregeln für gemischt stetige und diskrete Daten unter Einbeziehung einer adaptiven Regel. Diplomarbeit; Humboldt-Universität zu Berlin.

[195] Schwetlick, Hubert (1979). *Numerische Lösung nichtlinearer Gleichungen.* Deutscher Verlag der Wissenschaften: Berlin.

[196] Scott, David W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley: New York.

[197] Scott, D.W.; Thompson, J.R. (1983). Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, J.E. Gentle (Ed.). North-Holland: Amsterdam, pp. 173-179.

[198] Seber, G.A.F. (1984). *Multivariate Observations.* Wiley series in Probability and Mathematical Statistics. Wiley: New York.

[199] Shao, J. (1993b). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*; **88**, 486-494.

[200] Shao, J.; Tu, D. (1995). *The Jackknife and Bootstrap.* Springer series in statistics. Springer: New York.

[201] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall: London.

[202] Siotani, Minoru; Haykawa, Takesi; Fujikoshi, Yasunori (1985). *Modern Multivariate Statistical Analysis: A graduate Course and Handbook.* American Sciences Press: Columbus, Ohio.

[203] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.*; **B 36**, 111-147.

[204] Thorburn, Daniel (1976). Some asymptotic properties of jackknife statistics. *Biometrika*; **63**: 2, 305-313.

[205] Thiele, C. (1978). Ein Verfahren zur Diskrimination zwischen zwei mehrdimensionalen Grundgesamtheiten, das nur von den Erwartungswerten und Kovarianzmatrizen der zugrundeliegenden Verteilungen ausgeht. Diplomarbeit; Humboldt-Universität zu Berlin.

[206] Titterington, D.M. (1977). Analysis of incomplete multivariate binary data by the kernel method. *Biometrika*; **64**, 259-268.

[207] Titterington, D.M. (1980). A comparative study of kernel-based density estimators for categorical data. *Technometrics*; **22**, 259-268.

[208] Titterington, D.M.; Bowman, A.W. (1985). A comparative study of smoothing procedures for ordered categorical data. *J. Statist. Comput. Simulation*; **21**, 291-312.

[209] Titterington, D.M.; Mill, G.M. (1983). Kernel-based density estimates from incomplete data. *J. R. Statist. Soc.*; **B 45**, 258-266.

[210] Titterington, D.M.; Murray, G.D.; Murray, L.S.; Spiegelhalter, D.J.; Skene, A.M.; Habbema, J.D.F.; Gelpke, G.J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients (with discussion). *J. R. Statist. Soc.*; **A 144**, 145-175.

[211] Tutz, G.E. (1985). Smoothed additive estimators for non-error rates in multiple discriminant analysis. *Pattern Recognition*; **18**, 151-159.

[212] Tutz, G.E. (1986). An alternative choice of smoothing for kernel-based denstity estimates in discrete discriminant analysis. *Biometrika*; **73**, 405-411.

[213] Tutz, G.E. (1988). Smoothing for discrete kernels in discrimination. *Biom. J.*; **6**, 729-739.

[214] Van Campenhout, J.M. (1982). Topics in measurement selection. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). North-Holland: Amsterdam, pp. 793-803.

[215] Van Ness, J.W. (1979). On the effects of dimension in discriminant analysis for unequal covariance populations. *Technometrics*; **21**, 119-127.

[216] Van Ness, J.W.; Simpson, C. (1976). On the effects of dimension in discriminant analysis. *Technometrics*; **18**, 175-187.

[217] Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* Springer: New York.

[218] Venables, W.N.; Ripley, B.D. (1997). *Modern Applied Statistics with S-PLUS.* 2. Aufl. Springer: New York.

[219] Vlachonikolis, I.G.; Marriott, F.H.C. (1982). Discrimination with mixed binary and continuous data. *Appl. Statist.*; **31**, 23-31.

[220] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Statist.*; **15**, 145-162.

[221] Wand, M.P.; Jones, M.C. (1995). *Kernel Smoothing.* Monographs on Statistics and Applied Probability 60. Chapman & Hall: London.

[222] Wand, M.P.; Marron, J.S.; Ruppert, D. (1991). Transformations in density estimation (with discussion). *J. Amer. Statist. Assoc.*; **86**, 343-361.

[223] WERBIS, P.J. (1991). Links between artificial neural networks (ANN) and statistical pattern recognition. In *Artificial Neural Networks and Pattern Recognition: Old and New Connections*, Ishwar K. Sethi, Anil K. Jain (Eds.). Machine Intelligence and Pattern Recognition; vol. 11. North-Holland: Amsterdam 1991, pp. 11-32.

[224] WOJCIECHOWSKI, T.J. (1987). Nearest neighbor classification rule for mixtures of discrete and continuous random variables. *Biometrical J.*; **29**, 953-959.

[225] YANG, L.; MARRON, J.S. (1996). Iterated transformation-kernel density estimation. Discussion paper, SFB 373, Humboldt-Universität zu Berlin. (available at `sfb.wiwi.hu-berlin.de`)

[226] ZHANG, HEPING (1998). Classification trees for multiple binary response. *J. Amer. Statist. Assoc.*; **93**: 441, 180-193.

[227] ZHANG, P. (1993b). Model selection via multifold cross validation. *Ann. Statist.*; **21**, 299-313.

Medizinische & biologische Literatur

[228] ANDERSON KM, ODELL PM, WILSON PWF, KANNEL WB (1991). Cardiovascular disease risk profiles. *Am Heart J*; **121**, 293-8.

[229] ASSMANN G, CULLEN P (1995). Nationale Cardiovaskuläre Initiative: Erkennung und Behandlung von Fettstoffwechselstörungen. *Dt. Ärzteblatt*; **51/52**, Beilage.

[230] ASSMANN G, SCHULTE H, VON ECKARDSTEIN A (1996). Hypertriglyceridemia and elevated lipoprotein(a) are risk factors for major coronary events in middle-aged men. *Am J Cardiol*; **77**, 1179-1184.

[231] ASSMANN G, SCHULTE H (1992). Relation of high density lipoprotein cholesterol and triglycerides to incidence of atherosclerotic coronary artery disease (the PROCAM experience). *Am J Cardiol*; **70**, 733-737.

[232] BIELKA, HEINZ; BÖRNER, THOMAS (1995). *Molekulare Biologie der Zelle*. Fischer: Jena, Stuttgart.

[233] BÖRNER, THOMAS. Vorlesung „Grundlagen der Genetik und Molekularbiologie“. Sommersemester 1996 an der Humboldt-Universität zu Berlin. Private Mitschriften.

[234] Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel II) (1998). Summary of the second report of the National Cholesterol Education Program (NCEP). *JAMA*; **269**, 3015-3023.

[235] FRIEDLANDER Y, SISCOVICK DS, WEINMANN S, AUSTIN MA, PSATY BM, LEMAITRE RN, ARBOGAST P, RAGHUNATHAN TE, COBB LA (1998). Familiy History as a Risk Factor for Primary Cardiac Arrest. *Circulation*; **97**, 155-160.

[236] GENEST JR. JJ, MARTIN-MUNLEY SS, MCNAMARA JR, ORDOVAS JM, JENNER J, MYERS RH, SILBERMAN SR,WILSON PWF, SALEM DN, SCHAEFER EJ (1992). Familial Lipoprotein Disorders in Patients With Premature Coronary Artery Disease. *Circulation*; **85**, 2025-2033.

[237] GOLDSTEIN JL, HAZZARD WR, SCHROTT HG, MOTULSKY AG, BIERMAN EL (1973). Hyperlipidemia in coronary heart disease I.. Lipid levels in 500 survivors of myocardial infarction. *J Clin Invest*; **52**, 1533-1543.

[238] GOLDSTEIN JL, SCHROTT HG, HAZZARD WR, BIERMAN EL, MOTULSKY AG (1973). Hyperlipidemia in coronary heart disease II. Genetic analysis of lipid levels in 176 families and delination of a new inherited disorder, combined hyperlipidemia. *J Clin Invest*; **52**, 1544-1568.

[239] GOLDMAN RR, HUNT SC, SCHUMACHER C, HEGELE RA, LEPPERT MF, LUDWIG EH, HOPKINS PN (1993). Diagnosing Heterozygous Familial Hypercholesterolemia Using New Praktical Criteria Validated by Molecular Genetics. *Am J Cardiol*; **72**, 171-176.

[240] Hazzard WR, Goldstein JL, Schrott HG, Motulsky AG, Bierman EL (1973). Hyperlipidemia in coronary heart disease III. Evaluation of lipoprotein phenotypes of 156 genetically defined survivors of myocardial infarction. *J Clin Invest*; **52**, 1569-1577.

[241] Libbert, Eike (Hrsg.); bearb. von Günther, Elisabeth (1991). *Allgemeine Biologie*. UTB für Wissenschaft: Uni-Taschenbücher; 1197. Fischer: Jena.

[242] Schuster H (1995). Familienanamnese — der vergessene Risikofaktor. Fettstoffwechselstörungen als Beispiel. *Münch. med. Wschr.*; **137**: 16.

[243] Schuster H, Lamprecht A, Junghans C, Dietz B, Baron H, Nothnagel M, Mueller-Myhsok B, Luft FC (1998). Approaches to the genetics of cardiovascular disease through genetic field work. *Kidney Intern*; **53**, 1449-1454.

[244] Wood D, De Backer G, Faergeman O, Graham I, Mancia G, Pyörälä K (1998). Prevention of coronary heart disease in clinical practice. Summary of Recommendations of the Second Joint Task Force of European and other Societies on Coronary Prevention. *European Heart Journal, Atherosclerosis and Journal of Hypertension.*

Sonstige Literatur

[245] Messinger, Heinz (1991). *Langenscheidt's New English College German Dictionary: German–English / English–German*. Langenscheidt: New York, Berlin.

[246] Ueding, Gerd (1996). *Klassische Rhetorik*. Becksche Reihe, Bd. 2000. C.H. Beck: München.

Thesen

1. Die Diskriminanzanalyse ist eine statistische Verfahrensklasse, die sich um die Zuordnung oder Unterscheidung von Personen oder Objekten anhand einiger ihrer Merkmale zu einer bzw. in eine von mehreren vorher bekannten Klassen oder Populationen bemüht. Dieser allgemeine Ansatz ermöglicht den Einsatz dieser Verfahren in vielen, inhaltlich sehr unterschiedlichen Gebieten.

2. Die Entscheidungstheorie mittels der A-posteriori-Wahrscheinlichkeiten und des BAYESschen Risikos als Verlustfunktion bietet einen allgemeinen Rahmen, um die einzelnen Verfahren vorzustellen und zu vergleichen. Es können so außerdem Kosten und verschiedene Anteile der Populationen am Grundraum in die Analyse miteinbezogen werden.

3. Die Verfahren verwenden sehr unterschiedliche Ansätze und machen verschiedene Grundannahmen, um eine Diskriminanzregel aufzustellen. Grundsätzlich können sie danach unterschieden werden, ob Annahmen globaler oder lokaler Art gemacht werden. Innerhalb dieser Ansätze werden Parameter über Maximum-Likelihood- oder Plug-In-Schätzungen, durch BAYES-Zugänge oder durch iterative Optimierung verschiedener Kriterien bestimmt.

4. Parametrische Verfahren machen globale parametrische Annahmen über die A-posteriori- oder klassenspezifischen Verteilungen, deren Quotienten oder die Trennfläche. Unter diesen Annahmen können theoretisch optimale Schätzungen und Regeln abgeleitet werden. Das Finden der Regel besteht dann im Schätzen der globalen Parameter. Exponenten diesen Ansatzes sind Normalverteilungen in den Klassen.

5. Nichtparametrische Verfahren verzichten auf globale Annahmen und unterstellen lokale Eigenschaften wie Glattheit der A-posteriori- oder klassenspezifischen Verteilungen in einer Umgebung; es zählen so alle Verfahren dazu, die nicht parametrisch sind. Die Modellierung der Verteilungen erfolgt z.B. über lokale Mittelung in einer Umgebung, als rekursive Partitionierung des Merkmalsraumes oder auch funktional als nichtparametrische Regression.

6. Die letzten Jahre haben die Entwicklung weiterer nichtparametrischer Verfahren gesehen, die durch einen starken Gebrauch der Rechentechnik gekennzeichnet und ohne diese nicht denkbar sind. Sie sind Zeichen einer methodischen Revolution in der Diskriminanzanalyse — weg von rein theoretischen Kriterien, hin zu sehr rechenintensiven Verfahren, die sehr große Anzahlen von Fällen evaluieren. Sie bemühen sich z.B. um eine Varianzreduktion, verwenden Verfahren der nichtparametrischen Regression, um klassische Verfahren zu verallgemeinern, oder benutzen einen projektiven Zugang für die Dichteschätzung.

7. Die Anwendung der meisten Diskriminanzanalyse-Verfahren ist auf den Fall von drei oder mehr unterschiedenen Klassen erweiterbar: entweder durch die allgemeine BAYESsche Regel oder die Zurückführung der Mehr-Klassen- auf wiederholte Zwei-Klassen-Probleme.

8. Durch den begrenzten Stichprobenumfang sind theoretische Optimalitätskriterien nur bedingt aussagefähig. Die Begrenztheit und Zufälligkeit der Stichprobe bedingt die Zufälligkeit der daraus gewonnenen Diskriminanzregel. Die durch den Vergleich verschiedener Regeln gewonnene optimale Regel ist daher ebenfalls (in einem gewissen Maße) zufällig.

9. Das primäre Interesse bei einer Diskriminanzregel besteht in ihrem minimalen BAYESschen Risiko . Darum sollte eine möglichst genaue Schätzung dieses Risikos als Vergleichskriterium dienen. Dies wird für nicht zu geringe Stichprobenumfänge durch Cross-validation-, Bootstrap- und Testdatensatz-Schätzungen geleistet. Die Resubstitutionsmethode und die Benutzung von Kriterien, die nicht an das BAYESsche Risiko gekoppelt sind, sollten vermieden werden.

10. Das Risiko einer Diskriminanzregel kann analytisch in einen verfälschenden Teil (Bias) und einen zufälligen Teil (Varianz) zerlegt werden. Im konkreten Anwendungsfall sollte eine Bias-Varianz-Abwägung erfolgen mit der Risikoschätzung als Maß, um das Risiko zu minimieren. Die Abwägung ist motiviert durch die Hoffnung, die Varianz verringern zu können bei gleichzeitig nur moderater Erhöhung des Bias, um so das Risiko insgesamt zu senken.

11. Der Fehler in der Diskriminanzanalyse ist im Gegensatz zum Regressionskontext in einer Beobachtung von qualitativer Natur. Bias und Varianz wirken darum in anderer Weise auf den Verlust (Risiko). Die Abwägung zwischen beiden, um den Verlust zu minimieren, wird daher oft eine andere sein als in Regressionsproblemen. Optimalitätskriterien der Regression gelten somit nicht unbedingt in der Diskriminanzanalyse.

12. Mittel einer Bias-Varianz-Abwägung sind die Variablenauswahl, regularisierende (einschränkende) Annahmen über Parameter, Strafterme in Optimierungsproblemen für komplexe Strukturen oder die Verwendung bestimmter Verfahren auch bei invaliden Modellannahmen.

13. Geeignete Transformationen von Variablen können die Voraussetzungen oder Optimalitätskriterien von Verfahren erfüllen helfen und so deren Risiko senken. Der Einschluß abgeleiteter Variablen aus vielen Klassen von Transformationen birgt damit das Potential einer Verbesserung der Diskriminanzregeln.

14. Die Verfahren legen unterschiedliche Annahmen zugrunde, auf denen sie die Diskriminanzregel errichten, und verursachen dadurch in unterschiedlichem Maße eine Verfälschung bzw. eine Varianz der Regel. Im Vornherein lassen sich nur bedingt Aussagen über die Zulässigkeit der Verfahren treffen. Die vergleichende Anwendung solch unterschiedlicher Verfahren ist damit ebenfalls ein Weg für die Abwägung der beiden Fehlerkomponenten.

15. Die theoretische und praktische Erfahrung zeigt, daß kein Verfahren den anderen in allen Situationen überlegen ist. Stattdessen wird es für jede Situation ein jeweils bestes Verfahren geben. Dieses ist jedoch meist unbekannt und muß geschätzt werden. Einschränkungen auf gewisse Verfahrensklassen, Schätzansätze oder Methoden der Variablenauswahl — falls nicht durch wiederholte Anwendung oder fachspezifisches A-priori-Wissen begründet — können verfälschend wirken und das Risiko (unbemerkt) erhöhen. Darum sollte die Suche nach dem optimalen Verfahren über möglichst viele Verfahrensklassen und Variablenteilmengen — inklusive abgeleiteter Variablen — erfolgen, um so die Diskriminanzregel bestmöglich an die Daten anzupassen (Prinzip der Adaptiven Diskriminanzanalyse).

16. Die als Vergleichskriterium zu benutzende Risikoschätzung ist aufgrund des begrenzten Umfanges der Stichprobe ebenfalls fehlerbehaftet. Die eigentliche Aufgabe besteht so darin, unter vielen möglichen ein oder mehrere sehr gute Verfahren zu finden. Die endgültige Entscheidung über die Wahl der Diskriminanzregel aus diesem Kreis guter Regeln sollte nicht allein von deren geschätzten Risiken abhängen, sondern weitere anwendungsrelevante bzw. analytische Erfordernisse, wie Geschwindigkeit oder Interpretierbarkeit, miteinbeziehen.