Kapitel 1

Kategoriale Prognose und Diskriminanzanalyse

In den bisherigen Kapiteln lag der Schwerpunkt auf der Analyse des Zusammenhangs zwischen unabhängigen EinfluSSgröSSen und abhängigen kategorialen Variablen. Von Interesse war, welche Form dieser Zusammenhang hat und welche EinfluSSgröSSen wie stark die abhängige GröSSe bestimmen. In diesem Kapitel wird dargestellt, wie sich die entwickelten Modelle konkret zur Prognose einsetzen lassen.

Das Grundproblem besteht darin, aus einem Merkmalsvektor $\boldsymbol{x}' = (x_1, \ldots, x_p)$ die kategoriale Variable $Y \in \{1, \ldots, k\}$ zu prognostizieren. Dazu gilt es, den Zusammenhang zwischen Y und $\boldsymbol{x}$ herzustellen. Ein direkter Weg ist die Betrachtung der

- Verteilung von Y, gegeben $\boldsymbol{x}$, kurz $Y|\boldsymbol{x}$.

Dies ist der in den bisherigen Kapiteln verfolgte Ansatz, wenn die Verteilung $Y|\boldsymbol{x}$ beispielsweise durch ein Logit-Modell parametrisiert wird. Für die Schätzung der Parameter eines kategorialen Regressionsmodells ist allerdings im Regelfall eine entsprechende Form der Stichprobe notwendig. Entweder $(Y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n$, sind unabhängige Ziehungen aus der *gemeinsamen* Verteilung von $(Y, \boldsymbol{x})$, d.h. Y *und* $\boldsymbol{x}$ sind Zufallsvariablen, oder man betrachtet Realisationen der Zufallsvariablen Y bei gegebenem $\boldsymbol{x}$. Im letzteren Fall ist $\boldsymbol{x}$ eine vorgegebene Designvariable, beispielsweise die an eine Anzahl von Mäusen verabreichte Giftdosis in einem geplanten toxikologischen Experiment. Völlig analog lassen sich im folgenden Beispiel die Bedingungen zum Zeitpunkt der Unternehmensgründung als zu Beginn des "Experiments" festgelegte EinfluSSgröSSen verstehen.

Beispiel 1.1 : Unternehmensgründungen
In einer umfangreichen Studie um Erfolg von neugegründeten Unternehmen (Brüderl, Preisendörfer & Ziegler 1992) werden die Überlebenschancen der Unternehmen bewertet. Eine Vielzahl von potentiellen EinfluSSgröSSen wurden zum Zeitpunkt der Unternehmensgründung erhoben, darunter die Rechtsform des Unternehmens, der Wirtschaftsbereich, das Startkapital und der Zielmarkt (vgl. Appendix **??***). Als abhängige Variable wird das Scheitern des Unternehmens innerhalb der ersten drei Jahre ($Y = 1$: Scheitern, $Y = 2$: Überleben) betrachtet.*

□

In vielen Problemen der Praxis ist eine andere Form der Datenerhebung adäquater. Anstatt Realisationen der gemeinsamen Verteilung $(Y, \boldsymbol{x})$ bzw. der bedingten Verteilung $Y|\boldsymbol{x}$, erhält man häufig einfacher Beobachtungen aus den

- Verteilungen von $\boldsymbol{x}$, gegeben $Y = r$, kurz $\boldsymbol{x}|Y = r$.

Die Verteilungen des Merkmals $\boldsymbol{x}$ bei gegebener Kategorie $Y = r$ sind immer dann einfacher zu erhalten, wenn ein Diagnoseinstrument validiert werden soll. Soll $\boldsymbol{x}$ ein Indikator für eine Krankheit sein, läSSt sich die Verteilung von $\boldsymbol{x}$ jeweils in einer Population Gesunder und einer Population Erkrankter bestimmen. Analog ist es bei der Beurteilung von Kreditrisiken häufig adäquater, die Charakteristiken von Problemkunden und von unproblematischen Kunden als getrennte Stichproben zu erfassen, da Problemkunden weit seltener auftreten als das Pendant.

Beispiel 1.2 : Kredit-Scoring
In einer von Fahrmeir, Hamerle & Tutz (1996) betrachteten Untersuchung zur Kreditwürdigkeit wurde eine geschätzte Stichprobe von 300 guten und 700 schlechten Konsumentenkrediten betrachtet. Als Merkmale standen unter anderem die Laufzeit in Monaten, bisherige Zahlungsmoral, Verwendungszweck und Darlehenshöhe zur Verfügung. Eine ausführliche Darstellung der 20 betrachteten Merkmale findet sich in Fahrmeir, Hamerle & Tutz (1996).

□

Die prinzipielle Struktur des in diesen Beispielen gegebenen Entscheidungsproblems ist das einer Diagnose, die bestimmt ist durch

- einen unbeobachtbaren, möglicherweise erst in der Zukunft realisierten Zustand (Konkurs / Nicht-Konkurs bzw. kreditwürdig / nicht kreditwürdig),
- einen bzw. mehrere Indikatoren oder Prädiktoren, die Auskunft über den zugrundeliegenden Zustand geben sollen.

Analoge Entscheidungprobleme treten in vielen Bereichen auf, beispielsweise in der Medizin, wenn auf Grund von Symptomen bestimmt werden soll, ob eine bestimmte Krankheit vorliegt. Bei der Zeichenerkennung (pattern recognition) kann das Ziel darin bestehen, Buchstaben automatisiert zu erkennen, Wetterprognose zielt darauf ab, Zustände wie "Sonne" oder "Regen" zu unterscheiden. Ein Problem einfachster Struktur mit dichotomen Indikator und bekannten Auftretenswahrscheinlichkeiten ist das folgende.

Beispiel 1.3 : Drogenkonsum
In manchen US-Firmen werden bei Stellenbewerbern Tests durchgeführt, um zu eruieren, ob die Bewerber Drogenkonsumenten sind bzw. waren – für Angehörige der US-Regierung ist ein Test sogar obligatorisch. Marylin von Savant, die Frau mit dem derzeit höchsten gemessenen Intelligenzquotienten, diskutiert in der Gainesville Sun einen Test mit den folgenden bedingten Wahrscheinlichkeiten.

	Test positiv	*Test negativ*
Konsument	*0.95*	*0.05*
Nicht-Konsument	*0.05*	*0.95*

Auf Grund des Testergebnisses soll bestimmt werden, ob ein Bewerber Konsument oder Nicht-Konsument ist. Welchen Fehler begeht man, wenn man einen Bewerber mit positivem Testergebnis zum Konsumenten erklärt? □

Im einfachen Fall eines binären Indikators, d.h. eines Tests der positiv oder negativ ist, wird die Wahrscheinlichkeit eines positiven Ergebnisses, bei Vorliegen des Zustandes als *Sensitivität* bezeichnet. Die Sensitivität in Beispiel 1.3 ist also durch

$$P(\text{Test positiv}|\text{Konsument}) = 0.95$$

bestimmt. Die Wahrscheinlichkeit, daSS das Testergebnis negativ ist, wenn der Zustand nicht vorliegt, wird als *Spezifität* bezeichnet. Für den Drogenkonsum ist die Spezifität bestimmt durch

$$P(\text{Test negativ}|\text{Nicht-Konsument}) = 0.95.$$

1.1 Bayes-Zuordnung als diskriminanzanalytisches Verfahren

1.1.1 Grundkonzept

Ausgangspunkt ist hier der zweite Verteilungstyp, d.h. die Verteilung der Merkmale in den einzelnen Klassen. Die Transformation der Merkmalsverteilungen $\boldsymbol{x}|Y = r$ in die prognostisch interessante Verteilung $Y|\boldsymbol{x}$ erfolgt im Rahmen der auf der Bayes-Zuordnung aufbauenden Diskriminanzanalyse.

In etwas allgemeinerer Form als in Beispiel 1.3 betrachtet man mehrere unbeobachtbare Zustände $A_1, \ldots, A_k$ und Indikatoren bzw. Testergebnisse $T_1, \ldots, T_m$. Die bedingten Wahrscheinlichkeiten für spezifische Testergebnisse, gegeben ein latenter Zustand, besitzen die im folgenden gegebene Form, wobei zusätzlich die absoluten Wahrscheinlichkeiten für die zugrundeliegenden Zustände gegeben sind. Diese absoluten Wahrscheinlichkeiten bestimmen, mit welcher Wahrscheinlichkeit mit den einzelnen Zuständen zu rechnen ist, wenn das Testergebnis *noch nicht* berücksichtigt ist. Sie heiSSen daher auch *a priori-Wahrscheinlichkeiten*. Darin drückt sich das Vorwissen über den zu diagnostizierenden Zustand aus, d.h. mit welcher Wahrscheinlichkeit im Kreditscoring mit Problemkunden zu rechnen ist bzw. mit einer bestimmten Krankheit in medizinischen Problemen.

		Testergebnisse				A priori Wahrscheinlichkeiten
		T_1	$\ldots$	T_m		
latente Zustände	A_1	$P(T_1\|A_1)$	$\ldots$	$P(T_m\|A_1)$	1	$p(A_1)$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
	A_k	$P(T_1\|A_k)$	$\ldots$	$P(T_m\|A_k)$	1	$p(A_k)$

Welchen Zustand soll man diagnostizieren, wenn das Testergebnis T (d.h. $T_1, T_2, \ldots$ oder T_m) vorliegt. Eine naheliegende Regel besagt, denjenigen Zustand zu diagnostizieren, für den die bedingte Wahrscheinlichkeit, gegeben das Testergebnis T, maximal ist. Diese Zuordnungsregel wird als Bayes-Zuordnung bezeichnet.

Bayes-Zuordnung

Bei Vorliegen des Testergebnisses T wird derjenige Zustand A_i diagnostiziert, für den gilt

$$P(A_i|T) = \max_{j=1,\ldots,k} P(A_j|T)$$

Die Wahrscheinlichkeiten $P(A_1|T), \ldots, P(A_k|T)$ heiSSen *a posteriori Wahrscheinlichkeiten*, da sie sich erst aus der Beobachtung des Testergebnisses ergeben.

Der Zusammenhang zwischen den bedingten Wahrscheinlichkeiten der Testergebnisse $P(T_j|A_i)$, den a priori Wahrscheinlichkeiten $p(A_i)$ und den a posteriori-Wahrscheinlichkeiten wird durch den *Satz von Bayes* hergestellt, der gegeben ist durch

$$P(A_i|T) = \frac{P(T|A_i)p(A_i)}{\sum_{j=1}^{k} P(T|A_j)p(A_j)}.$$

Daraus läSSt sich unmittelbar die Bayes-Zuordnung bestimmen, wobei die a priori-Wahrscheinlichkeiten und die Verteilung der Testergebnisse gegeben die latenten Zustände eingehen.

Beispiel 1.4 : Drogenkonsum
Mit einer a priori-Wahrscheinlichkeit von $p(\text{Konsument}) = 0.10$ geht man davon aus, daSS 10% der Bewerber Drogenkonsumenten sind. Man erhält damit die folgende Zusammenfassung.

	T_1 *(positiv)*	T_2 *(negativ)*	*A priori*
A_1 *(Konsument)*	0.95	0.05	0.10
A_2 *(Nicht-Konsument)*	0.05	0.95	0.90

Nach der Regel von Bayes erhält man

$$P(A_1|T_1) = \frac{P(T_1|A_1)p(A_1)}{P(T_1|A_1)p(A_1) + P(T_1|A_2)p(A_2)} = \frac{0.95 \cdot 0.10}{0.95 \cdot 0.10 + 0.05 \cdot 0.90} = 0.68$$
$$P(A_2|T_1) = 0.32.$$

Somit ist die a posteriori Wahrscheinlichkeit $P(A_1|T_1)$, d.h. die Wahrscheinlichkeit, daSS ein Bewerber Drogenkonsument ist, wenn der Test positiv ausfällt, nur 0.68. Insbesondere heiSSt das, daSS unter den Bewerbern mit positiven Testergebnis auch 32% Nicht-Konsumenten zu erwarten sind, die fälschlicherweise nicht berücksichtigt werden, wenn man sich ausschlieSSlich am Testergebnis orientiert. Weiter ergibt sich für negatives Testergebnis T_2

$$P(A_1|T_2) = \frac{P(T_2|A_1)p(A_1)}{P(T_2|A_1)p(A_1) + P(T_2|A_2)p(A_2)} = \frac{0.05 \cdot 0.10}{0.05 \cdot 0.10 + 0.95 \cdot 0.9} = 0.006$$
$$P(A_2|T_2) = 0.994.$$

Unter den Bewerbern mit negativem Testergebnis befinden sich somit nur noch 0.6% Konsumenten, aber 99.4% Nicht-Konsumenten. Während das positive Testergebnis Konsumenten von Nicht-Konsumenten relativ schlecht trennt, ist das negative Testergebnis wesentlich trennschärfer. □

1.1.2 Bayes-Zuordnung und Fehlerraten

Im folgenden werden der Einfachheit halber die zugrundeliegenden, zu diagnostizierenden Zustände durch die Indikatorfunktion $Y \in \{1, \ldots, k\}$ ausgedrückt, wobei $Y = r$ dem Vorliegen des Zustands A_r entspricht. Die Indikatoren oder Tests

für den Zustand lassen sich allgemeiner durch Merkmale $\boldsymbol{x}' = (x_1, \ldots, x_p)$ ausdrücken, d.h. an den zu klassifizierenden Personen oder Objekten werden Merkmale $x_1, \ldots, x_p$ beobachtet, die AufschluSS über den zugrundeliegenden Zustand geben sollen. Durch die Auswahl einer Person aus der Population werden Y und $\boldsymbol{x}$ naturgemäSS zu Zufallsvariablen. Das *Zuordnungs- oder Klassifikationsproblem* ergibt sich nun daraus, daSS Y, d.h. der zugrundeliegende Zustand (oder die Klasse), nicht beobachtbar ist, der Merkmalsvektor $\boldsymbol{x}$ hingegen beobachtet wird. Man erhält somit das Zuordnungs- oder Entscheidungsproblem, dem Vektor $\boldsymbol{x}$ einen Zustand zuzuordnen. Gesucht ist also eine Zuordnungsregel

$$\begin{aligned} \delta : \mathbb{R}^n &\mapsto \{1, \ldots, k\} \\ \boldsymbol{x} &\mapsto \delta(\boldsymbol{x}), \end{aligned}$$

die möglichst "optimal" ist. Dabei wird dem Merkmalsvektor $\boldsymbol{x}$ die Klasse $\delta(\boldsymbol{x})$ zugeordnet. δ läSSt sich in vielen Fällen als Prognoseregel verstehen. Ein Zustand wie "Problemkunde" beim Kreditscoring wird erst in der Zukunft manifest, insofern ist die Diagnose eine Aussage über einen Zustand, der erst in der Zukunft relevant wird.

Wichtige GröSSen für das Klassifikationsproblem sind

- die *a priori-Wahrscheinlichkeiten* $p(r) = P(Y = r),\ r = 1, \ldots, k$,
- die *a posteriori-Wahrscheinlichkeiten* $P(r|\boldsymbol{x}) = P(Y = r|\boldsymbol{x}),\ r = 1, \ldots, k$
- die *Verteilung der Merkmale*, gegeben die Klasse, bestimmt durch die Dichten $f(\boldsymbol{x}|1), \ldots, f(\boldsymbol{x}|k)$.
- die *Mischverteilung* der Population $f(\boldsymbol{x}) = p(1)f(\boldsymbol{x}|1) + \cdots + p(k)f(\boldsymbol{x}|k)$.

Die Dichten $f(\boldsymbol{x}|r)$ geben dabei an, wie das Merkmal $\boldsymbol{x}$ verteilt ist, wenn die $r$te Klasse zugrundeliegt. Für diskrete Mermale $\boldsymbol{x}$ sind die Dichten einfache Auftretenswahrscheinlichkeiten, d.h. $f(\boldsymbol{x}|r)$ für $x \in \{1, \ldots, m\}$ entspricht dann den Wahrscheinlichkeiten $P(x = 1|Y = r), \ldots, P(x = m|Y = r)$. Für metrische Merkmale $\boldsymbol{x}$ ist es hingegen meist sinnvoller, für $f(\boldsymbol{x}|r)$ stetige Verteilungen anzunehmen, beispielsweise eine Normalverteilung.

Die Mischverteilungsdichte $f(\boldsymbol{x})$ bestimmt, wie das Merkmal $\boldsymbol{x}$ verteilt ist, wenn aus der Population zufällig gezogen wird, d.h. aus einer Mischpopulation, in der die Klasse r mit der Wahrscheinlichkeit $p(r)$ vorkommt. Für diskretes Merkmal $x \in \{1, \ldots, m\}$ reduziert sich die Formel für $f(x)$ auf den Satz von der totalen Wahrscheinlichkeit. Man erhält $f(x) = P(x|Y = 1)p(1) + \cdots + P(x|Y = k)p(k)$.

Die *Bayes-Regel* besitzt nun die einfache Form

$$\delta^*(\boldsymbol{x}) = r \quad \Longleftrightarrow \quad P(r|x) = \max_{i=1,\ldots,k} P(i|\boldsymbol{x}) \tag{1.1}$$

d.h. zu $\boldsymbol{x}$ wird die Klasse gewählt, für die die a posteriori-Wahrscheinlichkeit gegeben $\boldsymbol{x}$ maximal ist. Dabei bezeichnet $\delta^*(\boldsymbol{x})$ die Entscheidung bei Vorliegen des beobachteten Indikators $\boldsymbol{x}$.

1.1.3 Fehlklassifikationswahrscheinlichkeiten

Es lassen sich verschiedene Formen der Fehlklassifikationen durch die Zuordnungsregel δ unterscheiden. Die *Wahrscheinlichkeit einer Fehlklassifikation, gegeben der feste Merkmalsvektor* $\boldsymbol{x}$, ist bestimmt durch

$$\begin{aligned}\varepsilon(\boldsymbol{x}) &= P(\delta(\boldsymbol{x}) \neq Y|\boldsymbol{x}) = 1 - P(\delta(\boldsymbol{x}) = Y|\boldsymbol{x}) \\ &= 1 - P(\delta(\boldsymbol{x})|\boldsymbol{x}).\end{aligned}$$

Die *Verwechslungswahrscheinlichkeit* oder *individuelle Fehlerrate* ist gegeben durch

$$\varepsilon_{rs} = P(\delta(\boldsymbol{x}) = s|Y = r) = \int_{\boldsymbol{x}:\delta(\boldsymbol{x})=s} f(\boldsymbol{x}|r)d\boldsymbol{x}$$

und bestimmt die Wahrscheinlichkeit, ein Objekt, das aus Klasse r kommt, der Klasse s zuzuordnen.

Die *globale Fehlklassifikationswahrscheinlichkeit* oder *Gesamt-Fehlerrate* ist bestimmt durch

$$\varepsilon = P(\delta(\boldsymbol{x}) \neq Y).$$

Sie gibt die Wahrscheinlichkeit an, daSS die Zuordnungsregel eine Fehlklassifikation liefert.

Die *Wahrscheinlichkeit einer Fehlklassifikation, gegeben das Objekt entstammt der Klasse* r, ergibt sich durch

$$\varepsilon_r = P(\hat{\delta}(\boldsymbol{x}) \neq r|Y = r) = \sum_{s \neq r} \varepsilon_{rs}.$$

Die Gesamtfehlerrate läSSt sich auf einfache Weise sowohl aus dem Verwechslungswahrscheinlichkeiten als auch aus den bedingten Fehlklassifikationraten, gegeben $\boldsymbol{x}$, bestimmen.

$$\begin{aligned}\varepsilon = P(\delta(\boldsymbol{x}) \neq Y) &= \sum_{r=1}^{k} P(\delta(\boldsymbol{x}) \neq r|Y = r)p(r) = \sum_{r=1}^{k} \varepsilon_r p(r) \\ &= \sum_{r=1}^{k} \sum_{s \neq r} \varepsilon_{rs} p(r) \qquad (1.2)\end{aligned}$$

$$\varepsilon = P(\delta(\boldsymbol{x}) \neq Y) = \int P(\delta(\boldsymbol{x}) \neq Y|\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \int \varepsilon(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} \qquad (1.3)$$

Aus der Darstellung (1.3) ergibt sich eine wichtige Konsequenz: um die Gesamtfehlerrate ε zu minimieren, genügt es, für jedes $\boldsymbol{x}$ die bedingte Fehlerrate, gegeben $\boldsymbol{x}$, zu minimieren. Da die bedingte Fehlerrate durch $\varepsilon(\boldsymbol{x}) = 1 - P(\delta(\boldsymbol{x})|\boldsymbol{x})$ bestimmt ist, wird diese minimal, wenn zu gegebenem $\boldsymbol{x}$ die Zuordnung $\delta(\boldsymbol{x})$ so gewählt wird, daSS $P(\delta(\boldsymbol{x})|\boldsymbol{x})$ maximal ist. Diese Zuordnung entspricht genau der Bayes-Zuordnung δ^*.

Optimalität der Bayes-Zuordnung

Die Bayes-Zuordnung

$$\delta^*(\boldsymbol{x}) = r \quad \Longleftrightarrow \quad P(r|\boldsymbol{x}) = \max_{i=1,\ldots,k} P(i|\boldsymbol{x})$$

minimiert die Gesamtfehlerrate ε.

Die damit erreichte optimale Fehlklassifikationswahrscheinlichkeit ist durch

$$\varepsilon_{\text{opt}} = \int \min_{r=1,\ldots,k} \{1 - P(r|\boldsymbol{x})\} f(\boldsymbol{x}) d\boldsymbol{x}$$

gegeben. Anstatt Fehlerraten lassen sich auch Trefferraten betrachten. Die *Trefferrate, gegeben* $\boldsymbol{x}$, ist bestimmt durch $\tau(\boldsymbol{x}) = P(\delta(\boldsymbol{x})|\boldsymbol{x})$, die *Trefferrate, gegeben* $Y = r$, ist bestimmt durch $\tau_r = 1 - \varepsilon_r$ und die *Gesamttrefferrate* durch $\tau = 1 - \varepsilon$.

Beispiel 1.5 : Drogenkonsum
Mit $x = 1$ für positives und $x = 0$ für negatives Testergebnis ergeben sich die diskreten Dichten $f(x|r)$ durch

		$x = 1$	$x = 0$	*a priori* $p(r)$
Klasse	*1*	0.95	0.05	*0.10*
	2	0.05	0.95	*0.90*

Als a posteriori-Wahrscheinlichkeiten werden im Beispiel 1.4 (Seite 5) bestimmt

	$x = 1$	$x = 0$
$P(1\|x)$	0.68	0.006
$P(2\|x)$	0.32	0.994

Die Bayes-Zuordnung hat demnach die Form $\delta^*(x) = 1$ wenn $x = 1$ und $\delta^*(x) = 2$ wenn $x = 0$. Damit ergeben sich

$$\varepsilon(x = 1) = 1 - P(1|x = 1) = 1 - 0.68 = 0.32$$
$$\varepsilon(x = 0) = 1 - P(2|x = 0) = 1 - 0.994 = 0.006.$$

Die individuellen Fehlerraten sind bestimmt durch

$$\varepsilon_{12} = \varepsilon_1 = P(\delta^*(x) = 2|Y = 1) = P(x = 0|Y = 1) = 0.05$$
$$\varepsilon_{21} = \varepsilon_2 = P(\delta^*(x) = 1|Y = 2) = P(x = 1|Y = 2) = 0.05.$$

Als Trefferraten ergeben sich entsprechend $\tau_1 = 0.95$ *und* $\tau_2 = 0.95$. *Während die Verwechslungswahrscheinlichkeiten gleich sind, sind die Fehler, gegeben* x, *d.h.* $\varepsilon(x)$, *sehr unterschiedlich für* $x = 1$ *und* $x = 0$. *Als Gesamtfehlerrate erhält man*

$$\begin{aligned}\varepsilon &= P(\delta^*(x) \neq Y) = p(1)\varepsilon_{12} + p(2)\varepsilon_{21} = 0.10 \cdot 0.05 + 0.90 \cdot 0.05 \\ &= 0.05,\end{aligned}$$

entsprechend beträgt die Trefferrate $\tau = 0.95$.

Es empfiehlt sich prinzipiell, zu einer Zuordnungsregel die damit vorhandene Fehlklassifikationswahrscheinlichkeit anzugeben. Es ist wenig hilfreich, einen x*-Wert nach Bayes der optimalen Klasse zuzuordnen ohne sich Rechenschaft darüber abzugeben, wie gut diese Zuordnung ist. Für dieses einfache Problem erhält man*

	$x = 1$	$x = 0$
Zuordnung in Klasse	1	2
$\varepsilon(x)$	0.32	0.006

□

1.1.4 Bayes-Regel und Diskriminanzfunktionen

Eine alternative Darstellung erhält man mit Hilfe von Diskriminanzfunktionen. Man ordnet jedem Merkmalsvektor $\boldsymbol{x}$ Werte $d_r(\boldsymbol{x}),\ r = 1, \ldots, k$ zu, die angeben, wie stark die Beobachtung $\boldsymbol{x}$ für die Klasse r spricht. Wählt man $d_r(\boldsymbol{x}) = P(r|\boldsymbol{x})$, ergibt sich die Bayes-Regel als

$$\delta(\boldsymbol{x}) = r \quad \Longleftrightarrow \quad d_r(\boldsymbol{x}) = \max_{i=1,\ldots,k} d_i(\boldsymbol{x}). \tag{1.4}$$

Was hier nur wie eine neue Bezeichnung aussieht , läSSt sich als Zuordnung mit Diskriminanzfunktionen verstehen. Dabei werden jedem Merkmalsvektor $\boldsymbol{x}$ den Klassen zugehörende Werte $d_r(\boldsymbol{x}),\ r = 1, \ldots, k$, zugeordnet, und die Zuordnung erfolgt durch Maximierung der Diskriminanzfunktion entsprechend (1.4).

Das Konzept führt dazu, daSS anstatt $d_r(\boldsymbol{x}) = P(r|\boldsymbol{x})$ alternative Diskriminanzfunktionen betrachtet werden können. Die Verallgemeinerung des Satzes von Bayes hat die Form

$$P(r|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|r)p(r)}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|r)p(r)}{\sum_{i=1}^{k} p(i)P(\boldsymbol{x}|i)}.$$

Daraus ergibt sich unmittelbar für zwei Klassen r und s

$$\begin{aligned}P(r|\boldsymbol{x}) \geq P(s|\boldsymbol{x}) \Longleftrightarrow \quad & \frac{f(\boldsymbol{x}|r)p(r)}{f(\boldsymbol{x})} \geq \frac{f(\boldsymbol{x}|s)p(s)}{f(\boldsymbol{x})} \\ \Longleftrightarrow \quad & f(\boldsymbol{x}|r)p(r) \geq f(\boldsymbol{x}|s)p(s) \\ \Longleftrightarrow \quad & \log(p(\boldsymbol{x})) + \log(f(\boldsymbol{x}|r)) \geq \log(f(\boldsymbol{x}|s)) + \log(p(s)).\end{aligned}$$

Die Maximierung von $P(s|\boldsymbol{x})$ über $s = 1, \ldots, k$ führt demnach zum selben Ergebnis, wie die Maximierung von $f(\boldsymbol{x}|s)p(s)$ über $s = 1, \ldots, k$. Daraus ergibt sich, daSS die Bayes-Zuordnungsregel (1.4) mit äquivalentem Ergebnis formulierbar ist durch die Diskriminanzfunktionen

(a) $d_r(\boldsymbol{x}) = P(r|\boldsymbol{x})$,

(b) $d_r(\boldsymbol{x}) = f(\boldsymbol{x}|r)p(r)/f(\boldsymbol{x})$,

(c) $d_r(\boldsymbol{x}) = f(\boldsymbol{x}|r)p(r)$,

(d) $d_r(\boldsymbol{x}) = \log(f(\boldsymbol{x}|r)) + \log(p(r))$.

Bayes-Zuordnung

Bei Vorliegen des Beobachtungsvektors $\boldsymbol{x}$ erfolgt die Zuordnung in die Klasse r, für die $d_r(\boldsymbol{x})$ maximal ist, d.h.

$$\delta^*(\boldsymbol{x}) = r \quad \Longleftrightarrow \quad d_r(\boldsymbol{x}) = \max_{i=1,\ldots,k} d_i(\boldsymbol{x}),$$

wobei $d_r(\boldsymbol{x}) = P(r|\boldsymbol{x})$ bzw. $d_r(\boldsymbol{x}) = f(\boldsymbol{x}|r)p(r)$ bzw. $d_r(\boldsymbol{x}) = \log(f(\boldsymbol{x}|r)) + \log(p(r))$.

Die Darstellung durch äquivalente Diskriminanzfunktionen bietet die Möglichkeit einer alternativen Veranschaulichung der Bayes-Zuordnung. Die Diskriminanzfunktion $d_r(\boldsymbol{x}) = f(\boldsymbol{x}|r)p(r)$ entspricht bis auf den Faktor $p(r)$ der Merkmalsverteilung in der r-ten Klasse. In Abbildung 1.1 ist die zugehörige Zuordnungsregel dargestellt, wenn die Merkmale in den Klassen normalverteilt sind. Die Veränderung der a priori-Wahrscheinlichkeiten hat ein Aufblähen bzw. Schrumpfen der Merkmalsdichten zur Folge, die zu einer Verschiebung des Trennpunktes zwischen den Klassen führt.

Die alternativen Darstellungen durch Diskriminanzfunktionen sind vor allem auch deswegen relevant, weil sie verschiedene Wege weisen zur Bestimmung *geschätzter* Zuordnungsregeln. Wenn in der Anwendung die wahren GröSSen durch entsprechende Schätzungen ersetzt werden müssen, bieten sich für die Variante (a) der a posteriori-Wahrscheinlichkeiten direkt die in Kapitel **??** bzw. **??** behandelten parametrischen Modelle wie Logit- oder Probit-Modelle an. Eine nonparametrische Alternative sind die Methoden der nonparametrischen kategorialen Regression in Kapitel **??**. In den Varianten (b) bis (d) ist es notwendig, die Merkmalsdichten $f(\boldsymbol{x}|r)$ und die a priori-Wahrscheinlichkeiten $p(\boldsymbol{x})$ zu schätzen. Parametrische

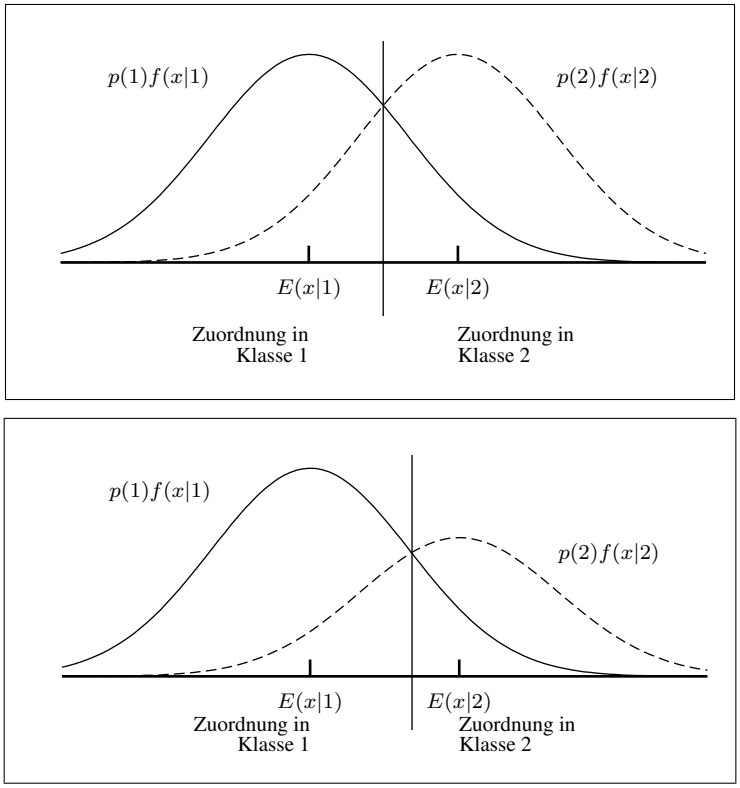

Abbildung 1.1: Zuordnung in die Klassen für gleiche a priori-Wahrscheinlichkeiten $p(1) = p(2) = 0.5$ (oberes Bild) und unterschiedliche a priori-Wahrscheinlichkeiten $p(1) = 0.6,\ p(2) = 0.4$ (unteres Bild)

Verfahren unterstellen einen Verteilungstyp für $f(\boldsymbol{x}|r)$, beispielsweise die Normalverteilung und schätzen die notwendigen Parameter, wie Erwartungswert und Varianz. Alternative Verfahren basieren auf nonparametrischen Dichteschätzern. Die a priori-Wahrscheinlichkeit – so sie nicht bekannt sind – werden durch die relativen Häufigkeiten der Klasse bestimmt. Geschätzte Diskriminanzregeln werden in Abschnitt 1.2 ausführlich behandelt.

1.1.5 Logit-Modell und normalverteilte Merkmale

Geht man bei zwei Klassen von klassenweise normalverteilten Merkmalen mit identischen Kovarianzmatrizen aus, d.h. $\boldsymbol{x}|Y = r \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$, ergeben sich interessante Spezialfälle. Die a posteriori-Wahrscheinlichkeit läSSt sich nach dem Satz von

Bayes in einer an die logistische Funktion angelehnten Form darstellen durch

$$P(1|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|1)p(1)}{f(\boldsymbol{x}|1)p(1) + f(\boldsymbol{x}|2)p(2)} = \frac{\exp(a)}{1+\exp(a)}, \tag{1.5}$$

wobei $a = \log[f(x|1)p(1)]/[f(x|2)p(2)]$. Setzt man die Dichten der Normalverteilung

$$f(\boldsymbol{x}|r) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_r)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_r)\right\}$$

ein, erhält man für a die Linearform

$$a = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta}$$

mit

$$\beta_0 = -\frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \log\left(\frac{p(1)}{p(2)}\right), \tag{1.6}$$

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{1.7}$$

Für die a posteriori-Wahrscheinlichkeit erhält man somit das logistische Modell

$$P(1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\beta_0 + \boldsymbol{x}'\boldsymbol{\beta})}.$$

Damit läSSt sich die Klassenzuordnung unmittelbar an einer logistisch modellierten a posteriori-Wahrscheinlichkeit festmachen. Anstatt der Diskrimininanzfunktion $d_r(\boldsymbol{x}) = P(r|\boldsymbol{x})$ läSSt sich alternativ auch die Differenz $d_1(\boldsymbol{x}) - d_2(\boldsymbol{x})$ betrachten mit der Zuordnungsregel

$$\text{Ordne } \boldsymbol{x} \text{ der Klasse 1 zu, wenn } d(\boldsymbol{x}) = d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) \geq 0.$$

Ansonsten wird der Klasse 2 zugeordnet. Mit der Diskriminanzfunktion $d_r(\boldsymbol{x}) = \log(P(r|\boldsymbol{x}))$ ergibt sich unmittelbar

$$\begin{aligned} & d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) \geq 0 \\ \iff \quad & \log(P(1|\boldsymbol{x})) - \log(P(2|\boldsymbol{x})) = \log\left(\frac{P(1|\boldsymbol{x})}{P(2|\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta} \geq 0. \end{aligned}$$

Man erhält damit die einfache lineare Zuordnungsregel

$$\text{Ordne } \boldsymbol{x} \text{ der Klasse 1 zu, wenn } \beta_0 + \boldsymbol{x}'\boldsymbol{\beta} \geq 0,$$

wobei die Parameter $\beta_0, \boldsymbol{\beta}$ durch (1.6) und (1.7) bestimmt sind.

Auch der allgemeinere Fall unterschiedlicher Kovarianzmatrizen normalverteilter Merkmale läSSt sich als logistisches Modell darstellen. Man erhält mit $\boldsymbol{x}|Y = r \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ für die GröSSe a aus (1.5)

$$a = \beta_0 + \boldsymbol{x}'\beta + \boldsymbol{x}'\boldsymbol{M}\boldsymbol{x},$$

wobei

$$\begin{aligned}\beta_0 &= -\frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \log\left(\frac{p(1)\,|\boldsymbol{\Sigma}_2|^{1/2}}{p(2)\,|\boldsymbol{\Sigma}_1|^{1/2}}\right),\\ \boldsymbol{\beta} &= \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2,\\ \boldsymbol{M} &= (\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})/2.\end{aligned}$$

Das entsprechende logistische Modell ist wiederum *linear in den Parametern*, nicht allerdings für die Prädiktoren. Mit $\boldsymbol{M} = (m_{ij})$ ist der hinzukommende Term von der Form $\boldsymbol{x}'\boldsymbol{M}\boldsymbol{x} = \sum_{i,j} m_{ij}x_ix_j$, enthält also eine Linearkombination von quadratischen Termen $m_{11}x_1^2, \ldots, m_{pp}x_p^2$ und "Interaktionen" $m_{ij}x_ix_j, i \neq j$. Betrachtet man als Prädiktor $\boldsymbol{z} = (x_1, \ldots, x_p, x_1^2, \ldots, x_1x_2, \ldots)$, ergibt sich wiederum ein lineares Logit-Modell.

1.1.6 Logit-Modell und binäre Merkmale

Für den Fall zweier Klassen und binärer Merkmale $x_1, \ldots, x_p$ gelte

$$P(x_1, \ldots, x_p|Y = r) = \pi^{(r)}_{x_1,\ldots,x_p}, \qquad r = 1, 2.$$

Aus der logistischen Form (1.5) erhält man

$$\begin{aligned}a &= \log[P(\boldsymbol{x}|1)p(1)/P(\boldsymbol{x}|2)p(2)]\\ &= \log \pi^{(1)}_{\boldsymbol{x}} - \log \pi^{(2)}_{\boldsymbol{x}} + \log(p(1)/p(2)).\end{aligned}$$

Aus der Theorie der loglinearen Modelle (vgl Kap. **??**) weiSS man, daSS sich der Logarithmus einer Auftretenswahrscheinlichkeit linear darstellen läSSt, d.h. es gibt Parameter, so daSS gilt

$$\log \pi^{(r)}_{\boldsymbol{x}} = \alpha_0^{(r)} + x_1\alpha_1^{(r)} + \cdots + x_p\alpha_p^{(r)} + x_1x_2\alpha_{12}^{(r)} + \cdots + x_1x_2\cdots x_p\alpha_{12\ldots p}^{(r)}.$$

Daraus erhält man für a

$$\begin{aligned}a = {} & \alpha_0^{(1)} - \alpha_0^{(2)} + \log(p(1)/p(2))\\ & + x_1(\alpha_1^{(1)} - \alpha_1^{(2)}) + \cdots + x_p(\alpha_p^{(1)} - \alpha_p^{(2)})\\ & + x_1x_2(\alpha_{12}^{(1)} - \alpha_{12}^{(2)}) + \cdots + x_1x_2\cdots x_p(\alpha_{12\ldots p}^{(1)} - \alpha_{12\ldots p}^{(2)}).\end{aligned}$$

Mit $\beta_0 = \alpha_0^{(1)} - \alpha_0^{(2)} + \log(p(1)/p(2))$ und $\beta_r = \alpha_r^{(1)} - \alpha_r^{(2)}$ ergibt sich das logistische Modell

$$P(1|x) = \frac{\exp(\boldsymbol{z}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}'\boldsymbol{\beta})},$$

wobei $\boldsymbol{z}' = (1, x_1, \ldots, x_p, x_1x_2, \ldots, x_1x_2\cdots x_p)$,
$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_2, \beta_{12}, \ldots, \beta_{12\ldots p})$.
Man erhält somit wiederum eine lineare Zuordungsregel

$$\delta(x) = 1 \qquad \Leftrightarrow \qquad \beta_0 + \boldsymbol{z}'\boldsymbol{\beta} \geq 0,$$

die allerdings nicht linear in $\boldsymbol{x}$ ist, sondern linear für den erweiterten Vektor $\boldsymbol{z}$, der auch alle Produkte zwischen Variablen enthält. Für den Spezialfall innerhalb der Klassen unabhängiger Merkmale erhält man $\beta_{12} = \beta_{13} = \cdots = \beta_{12\ldots p} = 0$, so daSS sich der erweiterte Vektor $\boldsymbol{z}$ auf $\boldsymbol{x}$ reduzieren läSSt (vgl. Abschnitt **??**).

1.1.7 Grenzen der Bayes Zuordnung: Maximum-Likelihood-Regel

Die Bayes-Zuordnung ist optimal in dem Sinne, daSS durch sie die globale Fehlklassifikationswahrscheinlichkeit, ebenso wie die lokale für gegebenes $\boldsymbol{x}$ minimiert wird. In die Zuordnungsregel gehen wesentlich die a priori-Wahrscheinlichkeiten, d.h. das Vorwissen über das Auftreten der einzelnen Klassen ein. Besitzt nun eine Klasse eine sehr hohe a priori-Wahrscheinlichkeit, kann die Bayes-Regel dahingehend entarten, daSS alle beobachteten Merkmalsvektoren $\boldsymbol{x}$ eben dieser Klasse zugeordnet werden. Das Diagnoseinstrument verliert damit jegliche Differenzierungskraft. Man vergleiche das folgende Beispiel.

Beispiel 1.6 : Drogenkonsum
Betrachtet wird wie in Beispiel 1.5 (Seite 8) nur das Auftreten ($x = 1$) oder Nicht-Auftreten ($x = 0$) eines Indikators für Klasse 1, allerdings hier mit der Sensitivität 0.95 und der Spezifität 0.9,

		$x = 1$	$x = 0$
Klasse	1	0.95	0.05
	2	0.10	0.90

Die a priori-Wahrscheinlichkeiten $p(1),\ p(2) = 1 - p(1)$ seien noch nicht festgelegt. Es läSSt sich einfach ableiten, wie die a posteriori-Wahrscheinlichkeiten von den a priori-Wahrscheinlichkeiten bestimmt werden. Man erhält die Funktionen

$$P(1|x = 1) = \frac{0.95p(1)}{0.85p(1) + 0.10}, \qquad p(1|x = 0) = \frac{0.05p(1)}{0.90 - 0.85p(1)},$$

die in Abbildung 1.2 dargestellt sind. Man sieht, daSS für $x = 1$ die a posteriori-Wahrscheinlichkeit für kleine a priori-Wahrscheinlichkeit sehr schnell, für $x = 0$ hingegen wesentlich langsamer steigt. Liegt die a priori-Wahrscheinlichkeit $p(1)$ jedoch unter 0.095 wird sowohl $x = 1$ als auch $x = 0$ der Klasse 2 zugeordnet, die Zuordnungsregel hängt nicht mehr vom Testergebnis x ab. Analog gilt für $p(1) > 0.947$, daSS die Zuordnung für jedes Beobachtungsergebnis in Klasse 1 erfolgt. Für diese Extrembereiche liefert die Bayes-Zuordnung zwar eine minimale Gesamtfehlerrate, allerdings wird diese vorwiegend durch

die a priori-Wahrscheinlichkeit bestimmt. Die Beobachtung selbst wird nicht mehr herangezogen. Man erhält für $p(1) < 0.095$ *die Verwechslungswahrscheinlichkeiten*

$$\varepsilon_{12} = 1, \qquad \varepsilon_{21} = 0,$$

für $p(1) > 0.947$

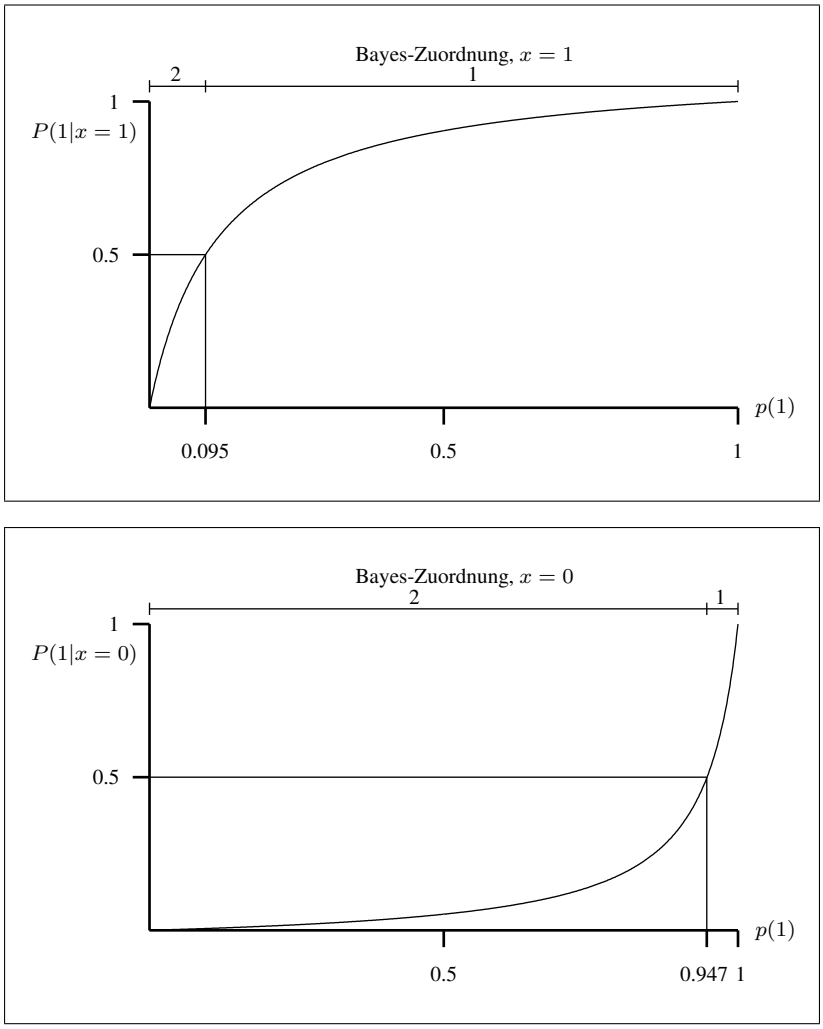

Abbildung 1.2: A posteriori-Wahrscheinlichkeiten für das Beispiel Drogenkonsum in Abhängigkeit von der a priori-Wahrscheinlichkeit $p(1)$, die Bayes-Zuordnung ist ebenfalls in Abhängigkeit von $p(1)$ zu verstehen

$$\varepsilon_{12} = 0, \qquad \varepsilon_{21} = 1.$$

□

Eine alternative Zuordnungsregel, die sich anbietet, wenn keine a priori-Wahrscheinlichkeiten vorliegen oder die Bayes-Zuordnung dahingehend entartet ist, daSS die Beobachtung irrelevant ist, ist die *Maximum-Likelihood- (ML)-Zuordnung* δ_{ML}. Sie entspricht der Bayes-Zuordnung mit gleichen a priori-Wahrscheinlichkeiten $p(1) = \cdots = p(k) = 1/k$. In der entsprechenden Diskriminanzfunktion $d_r(x) = f(\boldsymbol{x}|r)p(r) = f(\boldsymbol{x}|r)(1/k)$ läSSt sich der Faktor $1/k$ vernachlässigen und man erhält

Maximum-Likelihood- (ML)-Zuordnungsregel

$$\delta_{ML}(\boldsymbol{x}) = r \quad \Longleftrightarrow \quad f(\boldsymbol{x}|r) = \max_{i=1,\ldots,k} f(\boldsymbol{x}|i)$$

Die Regel läSSt sich analog zum Maximum-Likelihood-Schätzverfahren anschaulich interpretieren: man wählt zu gegebenem $\boldsymbol{x}$ diejenige Klasse r, die am ehesten dafür spricht, daSS gerade $\boldsymbol{x}$ beobachtet wird. Für diskretes $\boldsymbol{x}$ heiSSt das, man wählt diejenige Klasse, für die die Wahrscheinlichkeit, daSS $\boldsymbol{x}$ auftritt, maximal wird, für stetiges $\boldsymbol{x}$ wird entsprechend die Dichte maximiert.

Die ML-Regel im Beispiel 1.6 Drogenkonsum liefert $\delta_{ML}(x = 1) = 1$, $\delta_{ML}(x = 0) = 2$. Die dabei auftretenden Fehlklassifikationswahrscheinlichkeiten lassen sich nur in Abhängigkeit von den potentiell unbekannten a priori-Wahrscheinlichkeiten bestimmen. Man erhält die Fehlerraten

$$\begin{aligned}
\varepsilon_{12} &= P(x = 0|Y = 1) = 0.05, \\
\varepsilon_{21} &= P(x = 1|Y = 2) = 0.10, \\
\varepsilon(x = 1) &= 1 - P(1|x = 1), \\
\varepsilon(x = 0) &= 1 - P(2|x = 0) = P(1|x = 0), \\
\varepsilon &= 0.10 - 0.05\, p(1).
\end{aligned}$$

Man beachte, daSS die Gesamtfehlerrate von der (möglicherweise) unbekannten a priori-Wahrscheinlichkeit abhängt. In Abbildung 1.3 ist die Gesamtfehlderrate für die ML-Zuordnung in Abhängigkeit von $p(1)$ dargestellt. Zusätzlich ist die Gesamtfehlerrate nach Bayes eingezeichnet. Für $0.095 \leq p(1) \leq 0.947$ sind die Zuordnungsregeln und damit die Fehlerraten identisch (vgl. Beispiel 1.6). Für $p(1) \leq 0.095$ ergibt sich mit der Bayes-Zuordnung nach (1.2) $\varepsilon = p(1)$, für $p(1) \geq 0.947$ $\varepsilon = 1 - p(1)$. Die ML-Zuordnung liefert für sehr kleine und sehr groSSe Werte eine schlechtere Gesamtfehlerrate, liefert dafür aber immer einen Hinweis auf die Klasse, der nicht von den a priori-Wahrescheinlichkeiten beeinfluSSt wird. Insbesondere unter dem Aspekt, daSS die a priori-Wahrscheinlichkeiten meist

nicht bekannt sind oder nur grob approximiert werden, sollten die Hinweise der ML-Regel in einer Analyse mit berücksichtigt werden.

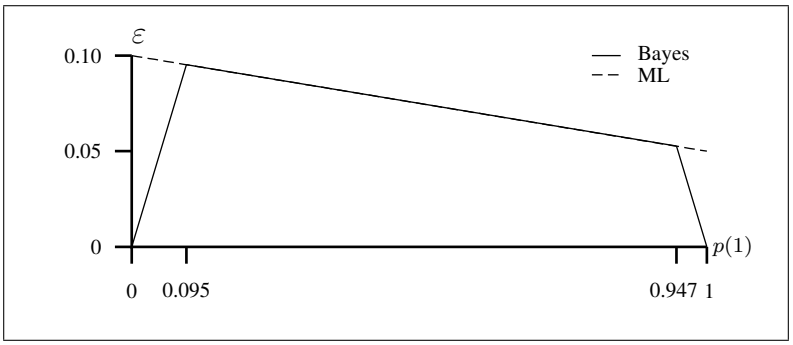

Abbildung 1.3: Gesamtfehlerrate in Abhängigkeit von der a priori Wahrscheinlichkeit $p(1)$

1.1.8 Kostenoptimale Bayes-Zuordnung

Etwas genereller lassen sich die bei falschen Zuordnungen anfallenden Kosten in die Entscheidungsregel einbeziehen. Dies führt zu allgemeineren kostenoptimalen Zuordnungsregeln. Seien die Kosten bestimmt durch

$$c(i,j) = c_{ij} = \text{Kosten einer Zuordnung eines Objekts aus Klasse } i \text{ in die Klasse } j.$$

Für $i = j$ gelte $c_{ii} = 0$, d.h. die Kosten einer richtigen Zuordnung sind Null, während $c_{ij} \geq 0$ für $i \neq j$.

Die zu erwartenden Kosten für gegebenes $\boldsymbol{x}$ ergeben das *bedingte Risiko, gegeben* $\boldsymbol{x}$, durch

$$r(\boldsymbol{x}) = \sum_{i=1}^{k} c_{i,\delta(\boldsymbol{x})} P(i|\boldsymbol{x}).$$

Das *individuelle Risiko* ist

$$r_{ij} = c_{ij} P(\delta(\boldsymbol{x}) = j | Y = i) = c_{ij} \int_{\boldsymbol{x}:\delta(\boldsymbol{x})=j} f(\boldsymbol{x}|i) d\boldsymbol{x}$$

und das *Risiko, gegeben Klasse* i, erhält man durch

$$r_i = \sum_{j=1}^{k} r_{ij}.$$

Das *gesamte Bayes Risiko*, d.h. die zu erwartenden Kosten lassen sich darstellen durch

$$R = E(c(Y, \delta(\boldsymbol{x}))) = \sum_{i=1}^{k} r_i p(i) = \int r(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}.$$

Wie im Fall der Minimierung der globalen Fehlerrate, wird das Bayes-Risiko minimal, wenn zu jedem $\boldsymbol{x}$ das bedingte Risiko $r(\boldsymbol{x})$ minimiert wird. Daraus ergibt sich unmittelbar die Bayes-optimale Zuordnungsregel.

Bayes-Zuordnung mit Kosten

Bei Vorliegen des Beobachtungsvektors $\boldsymbol{x}$ wird das Bayes-Risiko minimiert durch die Zuordnung

$$\delta^*(\boldsymbol{x}) = r \Longleftrightarrow \sum_{i=1}^{k} P(i|\boldsymbol{x}) c_{ir} = \min_{j=1,\ldots,k} \sum_{i=1}^{k} P(i|\boldsymbol{x}) c_{ij}.$$

Mit den Diskrimininanzfunktionen

$$d_r(\boldsymbol{x}) = -\sum_{i=1}^{k} P(i|\boldsymbol{x}) c_{ir},$$

erhält man

$$\delta^*(x) = r \quad \Longleftrightarrow \quad d_r(x) = \max_{i=1,\ldots,k} d_i(\boldsymbol{x}).$$

Einige Spezialfälle der Bayes-optimalen Zuordnung mit Kosten sind von Interesse:

(1) Wählt man die zwischen der Art der Fehlklassifikation nicht differenzierenden Kosten

$$c_{ij} = \begin{cases} c & i \neq j \\ 0 & i = j, \end{cases}$$

so ergibt sich die Bayes-Zuordnung durch

$$\begin{aligned} \delta^*(\boldsymbol{x}) = r \qquad &\Longleftrightarrow \qquad \sum_{i \neq r} P(i|\boldsymbol{x}) c = \min_{j=1,\ldots,k} \sum_{i \neq j} P(i|\boldsymbol{x}) c \\ &\Longleftrightarrow \qquad 1 - P(r|\boldsymbol{x}) = \min_{j=1,\ldots,k} (1 - P(j|\boldsymbol{x})) \\ &\Longleftrightarrow \qquad P(r|\boldsymbol{x}) = \max_{j=1,\ldots,k} P(j|\boldsymbol{x}). \end{aligned}$$

Die Zuordnungsregel entspricht der Bayes-Zuordnung (1.1), die ohne Berücksichtigung von Kosten entwickelt wurde. In diesem Fall entsprechen die Risiken den Fehlklassifikationswahrscheinlichkeiten, d.h.: $r(x) = \varepsilon(x), r_i = \varepsilon_i,\ R = \varepsilon$.

(2) Eine Alternative ist die umgekehrt proportionale Kostenfunktion

$$c_{ij} = \begin{cases} \frac{c}{p(i)} & i \neq j \\ 0 & i = j. \end{cases}$$

Damit werden die Kosten einer falschen Zuordnung eines aus Klasse i stammenden Objekts umso höher bewertet, desto kleiner die a priori-Wahrscheinlichkeit ist. Wegen $P(i|\boldsymbol{x}) = f(\boldsymbol{x}|i)p(i)/f(\boldsymbol{x})$ ergibt sich für $i \neq j$ $P(i|\boldsymbol{x})c_{ij} = P(i|\boldsymbol{x})c/p(i) = cf(\boldsymbol{x}|i)/f(\boldsymbol{x})$. Daraus erhält man

$$\begin{aligned} \delta_r(\boldsymbol{x}) = r \qquad &\Longleftrightarrow \qquad \sum_{i \neq r} cf(\boldsymbol{x}|i)/f(\boldsymbol{x}) = \min_{j=1,\ldots,k} \sum_{i \neq j} cf(\boldsymbol{x}|j)/f(\boldsymbol{x}) \\ &\Longleftrightarrow \qquad f(\boldsymbol{x}|r) = \max_{j=1,\ldots,k} f(\boldsymbol{x}|j). \end{aligned}$$

Die Zuordnungsregel ist damit äquivalent zur ML-Regel.

1.2 Geschätzte Zuordnungsregeln

1.2.1 Stichproben und geschätzte Zuordnungsregeln

Geschätzte Zuordnungsregeln $\hat{\delta}$ erhält man aus entsprechend geschätzten Diskriminanzfunktionen $\hat{d}_1, \ldots, \hat{d}_k$ durch

$$\hat{\delta}(\boldsymbol{x}) = r \qquad \Longleftrightarrow \qquad \hat{d}_r(\boldsymbol{x}) = \max_i \hat{d}_i(\boldsymbol{x}).$$

Die Schätzung der Zuordnungsregel beruht auf einer sogenannten *Lernstichprobe*, die ihren Namen aus der Tatsache bezieht, daSS sie dazu dient, ein Zuordnungsverfahren zu lernen. Welche Schätzungen sinnvoll sind, hängt von der Art der Lernstichprobe ab. Man vergleiche dazu auch die Einleitung des Kapitels. Unterschieden wird

- die *Gesamtstichprobe* $\{Y_i, \boldsymbol{x}_i\},\ i = 1, \ldots, n$ mit $(Y_i, \boldsymbol{x}_i)$ als unabhängigen Wiederholungen,
- die *nach $\boldsymbol{x}$ geschichtete Stichprobe* $Y_i^{(\boldsymbol{x})}|\boldsymbol{x},\ i = 1, \ldots, n(\boldsymbol{x})$, in der unabhängige Responses Y_i zu festen $\boldsymbol{x}$ beobachtet werden, und

– die *nach Y geschichtete Stichprobe* $\boldsymbol{x}_i^{(r)}|Y = r,\ i = 1, \ldots, n_r$, in der unabhängige Merkmalsvektoren zu fester Klasse gezogen werden.

Die hier vorwiegend betrachteten Verfahren basieren auf der Schätzung von $P(r|\boldsymbol{x})$, d.h. auf Diskriminanzfunktionen $\hat{d}_r(\boldsymbol{x}) = \hat{P}(r|\boldsymbol{x})$ bzw. $\hat{d}_r(\boldsymbol{x}) = -\Sigma_i c_{ir} P(i|\boldsymbol{x})$ im Falle unterschiedlicher Kosten. Die Schätzung von $\hat{P}(r|\boldsymbol{x})$ kann entweder durch die parametrischen Modelle der Kapitel **??–??** erfolgen oder durch nonparametrische Alternativen, wie sie in den Kapiteln **??** und **??** dargestellt sind. Die Verfahren beruhen entweder auf einer Gesamtstichprobe oder auf einer nach $\boldsymbol{x}$ geschichteten Stichprobe.

Die Alternative $\hat{d}_r(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}|r)p(r)$ beruht auf der Schätzung der Merkmalsverteilungen in jeder Klasse. Hier ist zu unterscheiden, ob $\boldsymbol{x}$ kategoriale, stetige oder beide Typen von Merkamlen enthält. Für kategoriale Merkmale ist eine Parametrisierung beispielsweise im Rahmen der log-linearen Modelle möglich, für stetige Merkmale wird häufig die Normalverteilung zugrundegelegt. Nonparametrische Variablen beruhen z.B. auf nonparametrischen Dichteschätzern für $f(\boldsymbol{x}|r)$. Verfahren des Typs $\hat{d}_r(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}|r)$, parametrische als auch nonparametrische, werden ausführlich in Fahrmeir, HäuSSler & Tutz (1996, Kapitel 8) betrachtet. Eine zunehmend wichtige Rolle spielen verteilungsfreie Ansätze, die von einer bestimmten Form der Diskriminanzfunktion ausgehen. Das Konzept der Fisherschen Diskriminanzanalyse sucht die beste lineare Trennung, d.h. im Zwei-Klassen-Fall wird von einer linearen Trennfunktion $d(x) = d_1(x) - d_2(x) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta}$ ausgegangen. Für normalverteilte und unabhängige binäre Merkmale ist die optimale Trennung tatsächlich linear, für andere Verteilungen kann die postulierte Linearität theoretisch suboptimal sein aber durchaus zu stabil geschätzten Zuordnungsverfahren führen. Der viel weitere Ansatz der neuronalen Netze läSSt auch nichtlineare Trennfunktionen zu, die hinsichtlich eines Zielkriteriums, optimiert werden. Einen guten Überblick über neuronale Netze und die verwendeten Algorithmen zur Klassentrennung findet sich bei Bishop (1995), Ripley (1996).

1.2.2 Prognosefehler – Direkte Prognose der Klassenzugehörigkeit

Sei $(y, \boldsymbol{x})$ mit $y \in \{0, 1\}$ eine neue Beobachtung, von der nur $\boldsymbol{x}$, nicht aber der Response bekannt ist. Wird die Wahrscheinlichkeit $\pi = P(y = 1|\boldsymbol{x})$ geschätzt, beispielsweise durch ein binäres Logit-Modell, $\pi = h(\boldsymbol{x}'\boldsymbol{\beta})$, läSSt sich die zugehörige Schätzung $\hat{\pi} = h(\boldsymbol{x}'\hat{\boldsymbol{\beta}})$ als eine Prognose für das Auftreten von $y = 1$ verstehen. Eine *direkte* Prognose des Response erhält man jedoch erst durch Anwendung der Bayes-Regel mit oder ohne Kosten. Im einfachsten Fall symmetrischer Kostenfunktion ergibt sich $\hat{y} = 1$, wenn $\hat{\pi} \geq 0.5$, und $\hat{y} = 0$, wenn $\hat{\pi} < 0.5$. Die "wei-

	Parametrisch	Nonparametrisch	Stichprobe
A posteriori direkt $\hat{d}_r(\boldsymbol{x}) = \hat{P}(r\|\boldsymbol{x})$	Logit-Modell und Alternativen (Kap. **??**–**??**)	Nonparametrische Responsemodelle Kapitel **??** und **??**	Gesamtstichprobe oder nach $\boldsymbol{x}$ geschichtet
Merkmalsschichten $d_r(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}\|r)p(r)$ bzw. $d_r(\boldsymbol{x}) = \log(\hat{f})(\boldsymbol{x}\|r) + \log(p(r))$	Parametrisierung von $\hat{f}(\boldsymbol{x}\|r)$ $\boldsymbol{x}$ kategorial: z.B. Log-lineare Modelle $\boldsymbol{x}$ stetig: z.B. Normalverteilung	Nonparametrische Dichteschätzung von $\hat{f}(\boldsymbol{x}\|r)$, neuronale Netze	Gesamtstichprobe oder nach Y geschichtet

Tabelle 1.1: Geschätzte Zuordnungsregeln

chere" Prognose $\hat{\pi}$ erhält naturgemäSS mehr Information über die Genauigkeit der Prognose, während $\hat{y}$ nur noch wiedergibt, welcher Response bzw. welche Klasse vorausgesagt wird. Ähnlich wie in der metrischen Regressionsanalyse läSSt sich die Prognose $\hat{\pi}$ bzw. $\hat{y}$ mit dem tatsächlichen y vergleichen. Man erhält für die zu erwartende Differenz bei festgelegter Prognose $\hat{y}$

$$E(\hat{y} - y) = \pi\,(\hat{y} - 1) + (1 - \pi)\hat{y} = \hat{y} - \pi.$$

Für die Prognose $\hat{y} = 1$ ist mit $1 - \pi$ eine tendenzielle Überschätzung, für $\hat{y} = 0$ mit $-\pi$ eine tendenzielle Unterschätzung impliziert. Für die weichere Prognose $\hat{\pi}$ gilt

$$E(\hat{\pi} - y) = \pi\,(\hat{\pi} - 1) + (1 - \pi)\hat{\pi} = \hat{\pi} - \pi.$$

Diese Differenz hängt naturgemäSS auch von der Güte der Schätzung $\hat{\pi}$ ab.

Im folgenden wird die direkte Prognose ausführlicher behandelt. Sei genereller $(Y, \boldsymbol{x})$ mit $Y \in \{1, \ldots, k\}$ eine neue Beobachtung, von der nur $\boldsymbol{x}$ bekannt ist. Die wahre Klasse Y wird nicht beobachtet, sondern durch eine Zuordnungsfunktion $\hat{Y} = \hat{\delta}(\boldsymbol{x})$ geschätzt. Mit $\pi = P(Y = 1|\boldsymbol{x})$ erhält man im Falle $k = 2$ für *die zu erwartende absolute Abweichung*

$$\begin{aligned} E(|Y - \hat{Y}|) &= \pi(1 - \hat{Y}) + (1 - \pi)|2 - \hat{Y}| \\ &= \begin{cases} 1 - \pi & \hat{Y} = 1 \\ \pi & \hat{Y} = 2. \end{cases} \end{aligned} \tag{1.8}$$

Wegen $\pi = P(1|\boldsymbol{x}),\ 1-\pi = P(Y=2|\boldsymbol{x})$ läSSt sich der Prognosefehler darstellen durch

$$E(|Y-\hat{Y}|) = 1 - P(\hat{\delta}(\boldsymbol{x})|\boldsymbol{x}) = \varepsilon(\boldsymbol{x})$$

und ist damit äquivalent zur Fehlklassifikationswahrscheinlichkeit, gegeben $\boldsymbol{x}$, bei Verwendung der Prognoseregel δ. Sie wird auch als *tatsächliche Irrtumsrate* bei Verwendung von δ bezeichnet.

Im mehrkategorialen Fall geht man zu der vektoriellen Darstellung über, wobei der Kodierungsvektor $\boldsymbol{y}' = (y_1, \ldots, y_k)$ mit $y_r = 1$, wenn $Y = r,\ y_r = 0$, wenn $Y \neq r$, die wahre Klasse kodiert, und $\hat{\boldsymbol{y}}' = (\hat{y}_1, \ldots, \hat{y}_k)$, mit $\hat{y}_r = 1$, wenn $\hat{\delta}(\boldsymbol{x}) = r,\ \hat{y}_r = 0$, wenn $\hat{\delta}(\boldsymbol{x}) \neq r$, den Prognosevektor darstellt. Die zu erwartende absolute Abweichung ergibt sich nun durch

$$\begin{aligned} &E\left(\sum_{r=1}^{k}|y_r - \hat{y}_r|\right) \\ &\quad= \sum_{r=1}^{k} E(|y_r - \hat{y}_r|) = \sum_{r=1}^{k}\{\pi_r|1-\hat{y}_r| + (1-\pi_r)|\hat{y}_r|\} \\ &\quad= (1-\pi_{\hat{Y}}) + \sum_{r\neq\hat{Y}}\pi_r = 2(1-\pi_{\hat{Y}}) = 2(1-P(\hat{\delta}(\boldsymbol{x})|\boldsymbol{x})) \end{aligned} \tag{1.9}$$

Für $k = 2$ ergibt sich damit $2(1 - P(\hat{\delta}(\boldsymbol{x})|\boldsymbol{x})$, was sich von (1.8) nur um den Faktor 2 unterscheidet. Der Faktor 2 ist darauf zurückzuführen, daSS in (1.9) über alle Kategorien (nicht nur über die erste wie im dichotomen Fall üblich) summiert wird.

Aus (1.8) und (1.9) ergibt sich, daSS die tatsächliche Fehlerrate (actual error rate), die der Fehlklassifikationswahrscheinlichkeit $\varepsilon(\boldsymbol{x})$ entspricht, sich bis auf den Faktor 2 als erwartete absolute Abweichung darstellen läSSt, d.h. es gilt

$$\varepsilon(\boldsymbol{x}) = \frac{1}{2}E\left(\sum_{r=1}^{k}|y_r - \hat{y}_r|\right) = 1 - P(\hat{\delta}(\boldsymbol{x})|\boldsymbol{x}).$$

Für die Güte der Regel $\hat{\delta}$ bei zufällig gezogenem $(Y, \boldsymbol{x})$ erhält man

$$\varepsilon = \int \varepsilon(\boldsymbol{x})\, f(\boldsymbol{x})\, d\boldsymbol{x} = E_{\boldsymbol{x}}(\varepsilon(\boldsymbol{x})).$$

Die tatsächliche Fehlerrate ist ein GütemaSS für eine *gegebene Zuordnungsregel*. Will man die Güte des gesamten Verfahrens, d.h. Schätzung der Zuordnungsregel *und* resultierende Treffsicherheit, beurteilen, ist zu berücksichtigen, daSS die geschätzte Zuordnungsregel auf einer Stichprobe beruht, $\varepsilon(\boldsymbol{x})$ wird dann zu einer Zufallsvariablen. Ein adäquates MaSS dafür wäre der Erwartungswert $E_S(\varepsilon(\boldsymbol{x}))$

bzw. $E_S(\varepsilon)$, wobei die Erwartungswertbildung über die Stichprobenverteilung erfolgt.

Da die tatsächliche Fehlerrate die unbekannte a posteriori-Wahrscheinlichkeit enthält, ist sie selbst zu schätzen. Ausgangspunkt sei eine Gesamtstichprobe $(Y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n$. Sei $\hat{\boldsymbol{y}}(\boldsymbol{x}_i) = (\hat{y}_1(\boldsymbol{x}_i), \ldots, \hat{y}_k(\boldsymbol{x}_i))$ die geschätzte Zuordnung mit den Komponenten

$$\hat{\boldsymbol{y}}_r(\boldsymbol{x}_i) = \begin{cases} 1 & \text{wenn } \hat{\delta}(\boldsymbol{x}_i) = r, \\ 0 & \text{sonst.} \end{cases}$$

Sollte die Zuordnung nicht eindeutig sein, wird der Klasse mit dem kleinsten Index zugeordnet oder zum Vektor $\hat{\boldsymbol{y}}(\boldsymbol{x}_i)/\sum_r \hat{\boldsymbol{y}}_r(\boldsymbol{x}_i)$ übergegangen. Eine einfache Schätzung ergibt sich, indem man in $\varepsilon(\boldsymbol{x})$ die wahre Wahrscheinlichkeit durch die Schätzung ersetzt. Dieses *plug-in Verfahren* liefert die Fehlerrate

$$\varepsilon_{PI} = \frac{1}{n}\sum_{i=1}^{n}(1 - \hat{P}(\hat{\delta}(\boldsymbol{x}_i)|\boldsymbol{x}_i)),$$

die auch als *apparent error rate* bezeichnet wird. Eine weitere naive Schätzung des Gesamtfehlers ist die *Resubstitutions- bzw. Reklassifikationsfehlerrate*

$$\varepsilon_R = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\sum_{r=1}^{k}|y_r(\boldsymbol{x}_i) - \hat{y}_r(\boldsymbol{x}_i)|,$$

wobei $y_r(\boldsymbol{x}_i) = 1$ wenn $Y_i = r$, 0 sonst. ε_R zählt nur die Anzahl der Fehlklassifikationen in der Lernstichprobe. Da die Lernstichprobe hier in zweifacher Hinsicht verwendet wird, einmal zur Generierung der Zuordnungsregel und sodann zur Güte der Zuordnungsregel, wird durch sie die tatsächlich zu erwartende Fehlklassifikationswahrscheinlichkeit systematisch unterschätzt. Sie gibt wieder, wie gut die *Lernstichprobe* trennbar ist, nicht aber wie gut die Zuordnungsregel für zukünftige Beobachtungen ist. Eine nahezu unverzerrte Schätzung der Fehlerrate beruht darauf, "zukünftige" Beobachtungen aus der Lernstichprobe zu generieren. Die einfachste, aber rechenintensive Methode besteht darin, *jeweils* eine Beobachtung aus der Lernstichprobe zu entfernen, die Zuordnungsregel aus dem restlichen $n-1$ Beobachtungen zu bestimmen und dann für diese Beobachtung festzustellen, ob sie richtig oder falsch klassifiziert ist. Das Verfahren wird als Kreuz-Klassifizierung (cross classification) bezeichnet. Die entsprechende *Kreuzklassifizierungsfehlerrate* ist durch

$$\varepsilon_{CV} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\sum_{r=1}^{k}|y_r(\boldsymbol{x}_i) - \hat{y}_r^{-i}(\boldsymbol{x}_i)|$$

gegeben, wobei $\hat{y}_r^{-i}(\boldsymbol{x}_i)$ die Zuordnungsregel ist, die aus der Lernstichprobe geschätzt ist ohne die $i$te Beobachtung $(Y_i, \boldsymbol{x}_i)$.

In Tabelle 1.2 sind diese einfachen Fehlerraten zusammengefaSSt. Zusätzlich angegeben sind die Fehlerraten aus der nach Klassen geschichteten Stichprobe $\boldsymbol{x}_i^{(r)}|Y = r,\ i = 1,\ldots,n_r$. Für ihre Berechnung ist die a priori-Wahrscheinlichkeit oder zumindest eine adäquate Schätzung notwendig.

Optimale Fehlerrate

$$\varepsilon_{\text{opt}} = \int \min_{r=1,\ldots,k}\{1 - P(r|\boldsymbol{x})\} f(\boldsymbol{x}) d\boldsymbol{x}$$

Plug-in-Schätzung (apparent error rate)

$$\varepsilon_{PI} = \frac{1}{n}\sum_{i=1}^{n}(1 - \hat{P}(\hat{\delta}(\boldsymbol{x})|\boldsymbol{x}))$$

Reklassifikationsfehlerrate (Resubstitutionsfehlerrate)

$$\varepsilon_R = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\sum_{r=1}^{k}|y_r(\boldsymbol{x}_i) - \hat{y}_r(\boldsymbol{x}_i)| \qquad \text{Gesamtstichprobe}$$

$$\varepsilon_R = \sum_{s=1}^{k} p(s)\left\{\frac{1}{n_s}\frac{1}{2}\sum_{r=1}^{k}|y_r(\boldsymbol{x}_i^{(s)}) - \hat{y}_r(\boldsymbol{x}_i^{(s)})|\right\} \qquad \begin{array}{l}\text{Nach Klassen}\\ \text{geschichtete}\\ \text{Stichprobe}\end{array}$$

Kreuzvalidierungs-Fehlerrate (Jackknife-Fehlerrate)

$$\varepsilon_{CV} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\sum_{r=1}^{k}|y_r(\boldsymbol{x}_i) - \hat{y}_r^{-i}(\boldsymbol{x}_i)| \qquad \text{Gesamtstichprobe}$$

$$\varepsilon_{CV} = \sum_{s=1}^{k} p(s)\left\{\frac{1}{n_s}\frac{1}{2}\sum_{r=1}^{k}|y_r(\boldsymbol{x}_i^{(s)}) - \hat{y}_r^{-i(s)}(\boldsymbol{x}_i)|\right\} \qquad \begin{array}{l}\text{Nach Klassen}\\ \text{geschichtete}\\ \text{Stichprobe}\end{array}$$

Tabelle 1.2: Einfache Fehlerklassifikationsraten für Gesmatstichprobe und nach Klassen geschichtete Stichprobe

1.2.3 Prognosefehler – alternative Schadensfunktionen

In diesem Abschnitt wird der Zusammenhang hergestellt zwischen Schadensfunktionen, die bei der Schätzung von Wahrscheinlichkeiten eine Rolle spielen, und Schadensfunktionen, die bei der Prognose zukünftiger Ereignisse Anwendung finden. Dabei werden die Schätzungen $\hat{\pi}_i$ als Prognose für y_i verstanden. Als erstes sei die Fehlklassifikationswahrscheinlichkeit, d.h. das Bayes-Risiko mit (0–1)-Schadensfunktion, betrachtet.

Zu festem $\boldsymbol{x}$ bezeichne $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_k)$ mit $\pi_r = P(r|\boldsymbol{x})$ den Vektor der a posteriori-Wahrscheinlichkeiten, $\hat{\boldsymbol{\pi}}' = (\hat{\pi}_1, \ldots, \hat{\pi}_k)$ mit $\hat{\pi}_r = \hat{P}(r|\boldsymbol{x})$ den geschätzten Vektor. Die Fehlklassifikationswahrscheinlichkeit $\varepsilon(\boldsymbol{x})$ bzw. das Bayes-Risiko läSSt sich als direkte Funktion von $\boldsymbol{\pi}$ und $\hat{\boldsymbol{\pi}}$ darstellen durch

$$L_B(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k} \pi_r (1 - \text{Ind}_r(\hat{\boldsymbol{\pi}})),$$

wobei

$$\text{Ind}_r(\boldsymbol{\pi}) = \begin{cases} 1 & \pi_r = \max\limits_{i=1,\ldots,k} \pi_i,\ \pi_r > \pi_i \text{ für } i < r \\ 0 & \text{sonst} \end{cases}$$

die Indikatorfunktion bezeichnet, die kodiert, ob die Zuordnung in Klasse r erfolgt. Die Indikatorfunktion ist so definiert, daSS im Zweifelsfall, d.h. wenn $\pi_r = \pi_s = \max \pi_i$ gilt, die Zuordnung in die Klasse mit der kleineren Klassennummer erfolgt. Es ergibt sich unmittelbar

$$L_B(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k} \pi_r (1 - \text{Ind}_r(\hat{\boldsymbol{\pi}})) = \sum_{r \neq \hat{\delta}(\boldsymbol{x})} \pi_r = 1 - \pi_{\hat{\delta}(\boldsymbol{x})} = 1 - P(\hat{\delta}(\boldsymbol{x})|\boldsymbol{x}) = \varepsilon(\boldsymbol{x}).$$

Versteht man den Vektor der geschätzten Wahrscheinlichkeiten $\hat{\boldsymbol{\pi}}' = (\hat{\pi}_1, \ldots, \hat{\pi}_k)$ als Prognose für die tatsächliche Klassenzugehörigkeit $\boldsymbol{y}' = (y_1, \ldots, y_k)$, läSSt sich der zu erwartende Schaden betrachten in der Form

$$E(L_B(\boldsymbol{y}, \hat{\boldsymbol{\pi}})) = E\left\{ \sum_{r=1}^{k} y_r (1 - \text{Ind}_r(\hat{\boldsymbol{\pi}})) \right\} = \sum_{r=1}^{k} \pi_r (1 - \text{Ind}_r(\hat{\boldsymbol{\pi}})) = L_B(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}).$$

Das Bayes-Risiko $L_B(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}})$ ist somit identisch dem zu erwartenden Schaden für eine zukünftige Beobachtung. Die zugrundeliegende (0–1) Schadensfunktion läSSt sich wie im vorhergehenden Abschnitt auch als absolute Abweichung darstellen. Es gilt

$$L_B(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k} y_r (1 - \text{Ind}_r(\hat{\boldsymbol{\pi}})) = \frac{1}{2} \sum_{r=1}^{k} |y_r - \text{Ind}_r(\hat{\boldsymbol{\pi}})|.$$

$L_B(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$ ist Null, wenn die geschätzte Bayes-Prognose richtig ist und Eins, wenn sie falsch ist. Durch diese spezielle Schadensfunktion wird $\hat{\boldsymbol{\pi}}$ als Schätzung von $\boldsymbol{y}$ unmittelbar in direkte Prognosewerte $\hat{y}_1, \ldots, \hat{y}_k$ umgesetzt. Die Erwartungswertbildung führt zur tatsächlichen Fehlerrate. Der Unterschied zu (1.9) liegt darin, daSS hier die Zuordnung als Bayes-Zuordnung formuliert ist, während in (1.9) eine beliebige feste Zuordungsregel $\hat{\delta}$ betrachtet wird.

Eine für die ML-Schätzung relevante Schadensfunktion ist die Kullback-Leibler-Distanz

$$L_{KL}(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k} \pi_r \log\left(\frac{\pi_r}{\hat{\pi}_r}\right).$$

Das Kullback-Leibler-Risiko, d.h. die für eine zukünftige Beobachtung zu erwartende Distanz, ist bestimmt durch

$$E(L_{KL}(\boldsymbol{y}, \hat{\boldsymbol{\pi}})) = E\left\{\sum_{r=1}^{k} y_r \log\left(\frac{y_r}{\hat{\pi}_r}\right)\right\} = -\sum_{r=1}^{k} \pi_r \log(\hat{\pi}_r).$$

Diese läSSt sich durch Erweitern mit der *Entropie* $\text{Ent}(\boldsymbol{\pi}) = -\sum \pi_r \log(\pi_r)$ darstellen durch

$$E(L_{KL}(\boldsymbol{y}, \hat{\boldsymbol{\pi}})) = L_{KL}(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) + \text{Ent}(\boldsymbol{\pi}).$$

Die zu erwartende Distanz einer zukünftigen Beobachtung setzt sich somit zusammen aus der Kullback-Leibler-Distanz zwischen $\boldsymbol{\pi}$ und der Schätzung von $\boldsymbol{\pi}$ und einer GröSSe die nur von $\boldsymbol{\pi}$ abhängt. Das Minimum des Kullback-Leibler-Risikos wird erreicht, wenn $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$, d.h. wenn $KL(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = 0$. Das optimal zu erreichende Risiko entspricht dann der Entropie der zugrundeliegenden Wahrscheinlichkeitsverteilung $\boldsymbol{\pi}$.

Eine weitere relevante Schadensfunktion ist die quadratische

$$L_Q(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k} (\pi_r - \hat{\pi}_r)^2.$$

Das quadratische Risiko (mean squared error), d.h. der zu erwartende Schaden für eine zukünftige Beobachtung ist gegeben durch

$$\begin{aligned} E(L_Q(\boldsymbol{y}, \hat{\boldsymbol{\pi}})) &= E\left\{\sum_{r=1}^{k} (y_r - \hat{\pi}_r)^2\right\} = \sum_{r=1}^{k} (\pi_r - \hat{\pi}_r)^2 + \sum_{r=1}^{k} \pi_r(1 - \pi_r) \\ &= L_Q(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) + \sum_{r=1}^{k} \mathit{var}(y_r). \end{aligned}$$

Man erhält wiederum ein Risiko, das sich als Distanz zwischen $\boldsymbol{\pi}$ und $\hat{\boldsymbol{\pi}}$ und einem nur von den zugrundeliegenden Wahrscheinlichkeiten abhängigen Term ergibt. Es wird minimal, wenn $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$ und damit $L_Q(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = 0$ gilt. In Tabelle 1.3 sind die Schadensfunktionen für $\hat{\boldsymbol{\pi}}$ als Schätzung von $\boldsymbol{\pi}$, der damit verbundene Schaden, wenn $\hat{\boldsymbol{\pi}}$ als Schätzung von $\boldsymbol{y}$ aufgefaSSt wird und das zugehörige Risiko für eine zukünftige Beobachtung wiedergegeben. Die zugrundeliegenden Schadensfunktionen, wenn $\hat{\boldsymbol{\pi}}$ als Schätzung von $\boldsymbol{y}$ verstanden wird, sind in Abb. 1.4 dargestellt. Dabei wird für den dichotomen Fall $k = 2$ von einer Beobachtung $Y = 1$, $((y_1, y_2) = (1, 0))$ ausgegeangen. Wie man sieht, hängt die Bayes-orientierte Schätzung nur davon ab, ob die Schwelle $\hat{\pi} = 0.5$ überschritten wird, die quadratische Funktion bestraft starke Abweichungen weniger als die logarithmische Funktion. Welche Schadensfunktion gewählt wird, hängt davon ab, wie stark man die Abweichung zwischen $\hat{\pi}$ und $Y = 1$ bestrafen will.

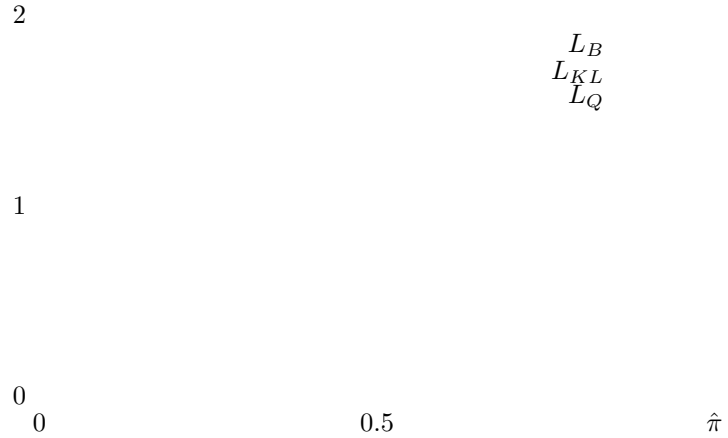

Abbildung 1.4: Schadensfunktionen $L((1,0),(\hat{\pi},1-\hat{\pi}))$ für Bayes-, Kullback-Leibler- und quadratischen Schaden.

Der zu erwartende Schaden wird auch als tatsächlicher Irrtum bzw. tatsächliches Risiko bezeichnet (vgl. Van Houwelingen & le Cessie 1990). Für Beobachtungen $(\boldsymbol{y}_i, \boldsymbol{x}_i),\ i = 1,\ldots,n$, und zugehörige Wahrscheinlichkeiten bzw. Schätzungen $\boldsymbol{\pi}_i' = (\pi_{i1},\ldots,\pi_{ik})$, $\hat{\boldsymbol{\pi}}_i' = (\hat{\pi}_{i1},\ldots,\hat{\pi}_{ik})$ erhält man die plug-in-Schätzungen

$$
\begin{aligned}
L_{B,PI} &= \sum_{i=1}^{n} L_B(\boldsymbol{\pi}_i, \hat{\boldsymbol{\pi}}_i), \\
L_{KL,PI} &= \sum_{i=1}^{n} KL(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{\pi}}_i) + \sum_{r=1}^{k}\{-\hat{\pi}_{ir}\log(\hat{\pi}_{ir})\} = \sum_{i=1}^{n}\sum_{r=1}^{k}(-\hat{\pi}_{ir}\log(\hat{\pi}_{ir})), \\
L_{Q,PI} &= \sum_{i=1}^{n} L_Q(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{\pi}}_i) + \sum_{r=1}^{k}\hat{\pi}_{ir}(1-\hat{\pi}_{ir}) = \sum_{i=1}^{n}\sum_{r=1}^{k}\hat{\pi}_{ir}(1-\hat{\pi}_{ir}),
\end{aligned}
$$

die eine erhebliche Unterschätzung der tatsächlichen Risiken aufweisen. Davon zu unterscheiden sind die ebenfalls verzerrten Resubstitutionsfehler und das weniger verzerrte, aber mit groSSer Variabilität behaftete Verfahren der Kreuzvalidierungs- oder leaving-one-out-Methode. Für eine Gesamtstichprobe erhält man die Kreuzva-

$L_B(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k} \pi_r(1 - \text{Ind}_r(\hat{\boldsymbol{\pi}})) = \varepsilon(x)$	Bayes-Schaden
$L_B(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = 1 - \text{Ind}_Y(\hat{\boldsymbol{\pi}}) = \frac{1}{2}\sum_{r=1}^{k} \lvert y_r - \text{Ind}_r(\hat{\boldsymbol{\pi}})\rvert$	0–1-Schadensfunktion
$E(L_B(\boldsymbol{y}, \hat{\boldsymbol{\pi}})) = L_B(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}})$	Tatsächliche Fehlerrate
$L_{KL}(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k} \pi_r \log\left(\frac{\pi_r}{\hat{\pi}_r}\right)$	Kullback-Leibler-Distanz
$L_{KL}(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = -\log(\hat{\pi}_Y) = -\sum_{r=1}^{k} y_r \log(\hat{\pi}_r)$	Logarithmierter Score
$E(L_{KL}(\boldsymbol{y}, \hat{\boldsymbol{\pi}})) = KL(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) + \sum_{r=1}^{k} -\pi_r \log(\pi_r)$	Tatsächliches KL-Risiko
$L_Q(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \sum_{r=1}^{k}(\pi_r - \hat{\pi}_r)^2$	Quadratischer Schaden
$L_Q(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = (1 - \hat{\pi}_Y)^2 + \sum_{i \neq Y} \hat{\pi}_i^2$	Quadratischer Score
$E(L_Q(\boldsymbol{y}, \hat{\boldsymbol{\pi}})) = L_Q(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) + \sum_{r=1}^{k} \pi_r(1 - \pi_r)$	Tatsächliches quadratisches Risiko

Tabelle 1.3: Schadensfunktionen für $\hat{\boldsymbol{\pi}}$ als Schätzung von $\boldsymbol{\pi}$ und für $\hat{\boldsymbol{\pi}}$ als Schätzung von $\boldsymbol{y}$ sowie die zugehörigen Erwartungswerte

lidierungsraten

$$L_{B,CV} = \sum_{i=1}^{n} L_B(\boldsymbol{y}_i, \hat{\boldsymbol{\pi}}_i^{-i}),$$

$$L_{KL,CV} = \frac{1}{n}\sum_{i=1}^{n} L_{KL}(\boldsymbol{y}_i, \hat{\boldsymbol{\pi}}^{-i}) = -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{\pi}_{Y_i}),$$

$$L_{Q,CV} = \frac{1}{n}\sum_{i=1}^{n} L_Q(\boldsymbol{y}_i, \hat{\boldsymbol{\pi}}^{-i}),$$

wobei $\hat{\boldsymbol{\pi}}^{-i}$ die Schätzung ohne die Beobachtung i bezeichnet. $L_{KL,CV}$ läSSt sich auch als kreuzvalidierte Devianz bzw. Likelihood auffassen.

Alle diese Schätzungen zielen allerdings eher auf das zu erwartende tatsächliche Risiko ab als auf das tatsächliche Risiko, das von der Stichprobe abhängig und damit

eine Zufallsvariable ist. Das zu erwartende Risiko hat die Form

$$E_S\left(E_{y,\boldsymbol{x}}(L(y|\boldsymbol{x}),\hat{\pi}(\boldsymbol{x}))\right),$$

wobei $E_{y,\boldsymbol{x}}$ den Erwartungswert bzgl. einer zukünftigen Beobachtung $(y,\boldsymbol{x})$ bezeichnet und E_S den Erwartungswert über die Stichprobenverteilung darstellt, die zur Schätzung führt. Für quadratischen und Kullback-Leibler-Schaden zerfällt das Kriterium in einen Term, der nur von der Schätzgüte abhängt und einen konstanten Term. Für festes $\boldsymbol{x}$ erhält man

$$\begin{aligned} E_S E_y(L_Q(\boldsymbol{y},\hat{\boldsymbol{\pi}})) &= E_S\{L_Q(\boldsymbol{\pi},\hat{\boldsymbol{\pi}})\} + \sum_{r=1}^{k}\pi_r(1-\pi_r) \\ &= \sum_{r=1}^{k}(E(\hat{\pi}_r)-\pi_r)^2 + var(\hat{\pi}_r)) + \sum_{r=1}^{k}\pi_r(1-\pi_r), \end{aligned}$$

wobei $E(\hat{\pi}_r)-\pi_r$ die Verzerrung enthält. Für den Kullback-Leibler-Schaden ergibt sich

$$\begin{aligned} E_S E_y(L_{KL}(\boldsymbol{y},\hat{\boldsymbol{\pi}})) =& E_S\{KL(\boldsymbol{\pi},\hat{\boldsymbol{\pi}})\} - \sum_{r=1}^{k}\pi_r\log(\pi_r) \\ =& KL(\boldsymbol{\pi},E_S(\hat{\boldsymbol{\pi}})) + \sum_{r=1}^{k}\pi_r\{\log(E_S(\hat{\pi}_r)) - E_S(\log(\hat{\pi}_r))\} \\ & - \sum_{r=1}^{k}\pi_r\log(\pi_r), \end{aligned}$$

wobei die ersten beiden Terme wiederum eine Zerlegung in Verzerrung und Variabilität darstellen.

Implizit wurde bisher immer davon ausgegangen, daSS $\hat{\pi}=\hat{\pi}(\boldsymbol{x})$ eine Schätzung für den zugrundeliegenden Vektor der bedingten Wahrscheinlichkeiten $\boldsymbol{\pi}'(\boldsymbol{x}) = (P(Y=1|\boldsymbol{x}),\ldots,P(Y=k|\boldsymbol{x}))$ darstellt, der selbst eine gute Prognose für eine zukünftige Beobachtung y darstellt. Die Optimalität von $\boldsymbol{\pi}(\boldsymbol{x})$ als Voraussage ergibt sich, wenn man $\pi(\boldsymbol{x})$ mit einer alternativen Prognosefunktion $\tilde{\boldsymbol{\pi}}(\boldsymbol{x})$ vergleicht. Wie man einfach ableitet, gilt

$$E_{y,\boldsymbol{x}}\{KL(\boldsymbol{y},\tilde{\boldsymbol{\pi}}(\boldsymbol{x}))\} = E_{y,\boldsymbol{x}}\{KL(\boldsymbol{y},\boldsymbol{\pi}(\boldsymbol{x}))\} + E_{\boldsymbol{x}}\{KL(\boldsymbol{\pi}(\boldsymbol{x}),\tilde{\boldsymbol{\pi}}(\boldsymbol{x}))\}.$$

Diese Funktion wird minimal für $\tilde{\boldsymbol{\pi}}(\boldsymbol{x})=\boldsymbol{\pi}(\boldsymbol{x})$. Die bedingte Wahrscheinlichkeit ist somit optimal im Sinne der minimalen zu erwartenden Kullback-Leibler-Distanz.

Beispiel 1.7 : Unternehmensgründungen
Das in Beispiel 1.1 betrachtete Scheitern von neugegründeten Unternehmen innerhalb von drei Jahren wurde untersucht mit den Variablen Startkapital, Eigenkapital, Rechtsform,

Fremdkapital und Alter des Unternehmensführers (vgl. Anhang **??***). Bezeichne* $Y = 1$ *das Scheitern und* $Y = 2$ *das Überleben des Unternehmens. Für die lineare Diskriminanzanalyse ergab sich nach der Reklassifikationsmethode folgende Verwechslungstabelle in der Lernstichprobe.*

		Zugeordnete Klasse		
		1	*2*	
Klasse	*1*	*69*	*231*	*300*
	2	*73*	*851*	*924*
		142	*1082*	*1224*

Die Resubstitutions-Verwechlungswahrscheinlichkeiten ergeben sich durch $\hat{\varepsilon}_{12} = 0.77$, $\hat{\varepsilon}_{21} = 0.08$ *und die Gesamtfehlerrate beträgt* $\hat{\varepsilon}_R = (231 + 73)/1224 = 0.248$. *Für die Kreuzvalidierung, bei der jeweils eine Beobachtung der Lernstichprobe prognostiziert wird, ergibt sich*

		Zugeordnete Klasse		
		1	*2*	
Klasse	*1*	*67*	*233*	*300*
	2	*79*	*845*	*924*
		146	*1078*	*1224*

mit den Fehlerraten $\hat{\varepsilon}_{12} = 0.77$, $\hat{\varepsilon}_{21} = 0.08$ *und* $\hat{\varepsilon}_{CV} = 0.254$, *was im Vergleich zur Resubstitutionsfehlerrate eine unerhebliche Abweichung darstellt.* □

Literaturverzeichnis

BISHOP, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

BRÜDERL, J., PREISENDÖRFER, P., UND ZIEGLER, R. (1992). Survival chances of newly founded business organizations. *American Sociological Review 57*, 227–242.

FAHRMEIR, L., HAMERLE, A., UND TUTZ, G. (1996). *Multivariate statistische Verfahren* (2. Aufl.). Berlin, New York: de Gruyter.

FAHRMEIR, L., HÄUSSLER, W., UND TUTZ, G. (1996). Diskriminanzanalyse. In L. Fahrmeir, A. Hamerle, & G. Tutz (Hrsg.), *Multivariate statistische Verfahren*. De Gruyter.

RIPLEY, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: University Press.

VAN HOUWELINGEN, J. UND LE CESSIE, S. (1990). Predictive value of statistical models. *Statistics in Medicine 9*, 1303–1325.