Multivariate Verfahren

Diskriminanzanalyse

Annika Hoyer

Sommersemester 2020

Diskriminanzanalyse - Inhalt

Ausgangssituation

Ausgangssituation

- Grundgesamtheit zerfällt in $g \geq 2$ disjunkte Klassen mit Indikator $Y \in \{1, \ldots, g\}$
- Beobachtung von Merkmalsvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- Aber: Klassenzugehörigkeit $Y_1, \ldots, Y_n$ unbekannt

Problemstellung

Die Objekte $a_1, \ldots, a_n$ sollen mithilfe der an ihnen beobachteten Merkmalsvektoren in eindeutiger Weise jeweils genau einer Klasse zugeordnet werden.

Beispiele

- Unterscheidung von Kreditnehmern (vertrauenswürdig / nicht vertrauenswürdig)
- Klassifizierung von Krankheiten (Bronchitis / Lungenentzündung)
- Bestimmung des Krankheitsstatus (krank / gesund)
- Identifizierung von Drogenkonsumenten (User / Non-User)
- Mustererkennung in Texten (Buchstaben)

Merkmale der Diskriminanzanalyse

- Klassen, in die Objekte eingeteilt werden sollen, sind vorab bekannt
- Diskriminanzanalyse ist Verfahren des ”supervised learning”
- Synonyme: ”Klassifikation” und ”Pattern Recognition”
- Grundlage: Fehlerraten bzw. Fehlklassifikationswahrscheinlichkeiten
- Auswahl der Merkmale, die für die Klassifikation herangezogen werden, ist von zentraler Bedeutung

Datengrundlage

Der Merkmalsvektor $\mathbf{x}$ und die Klasse Y sind charakterisiert durch:

- a priori-Wahrscheinlichkeiten:
  $p(r) = P(Y = r),\ r = 1, \ldots, g$
- a posteriori-Wahrscheinlichkeiten:
  $P(r|\mathbf{x}) = P(Y = r|\mathbf{x}),\ r = 1, \ldots, g$
- die Dichte von $\mathbf{x}$ gegeben der Klasse:
  $f(\mathbf{x}|Y = 1), \ldots, f(\mathbf{x}|Y = g)$
- die Mischverteilung in der Gesamtpopulation:
  $f(\mathbf{x}) = \sum_{j=1}^{g} f(\mathbf{x}|j)p(j)$

Diskriminanzanalyse

Beispiel: Studienanfänger Wirtschaftswissenschaften

- Test zur Mittelstufenalgebra (26 Fragen) im Wintersemester 1988/89 bei Studienanfängern der Wirtschaftswissenschaften an der FU Berlin
- Variablen:
  - Geschlecht: w/m (1/0)
  - Besuch Leistungskurs Mathe: j/n (1/0)
  - Abitur im Jahr 1988: j/n (1/0)
  - Abinote Mathematik
  - Anzahl der im Test richtig gelösten Aufgaben
  - Gruppe 1: mindestens 14 Punkte erreicht $\rightarrow$ Test bestanden
  - Gruppe 2: weniger als 14 Punkte $\rightarrow$ Test nicht bestanden

Beispiel: Studienanfänger Wirtschaftswissenschaften

Ergebnisse der Studienanfänger bei dem Mathetest

Geschlecht	MatheLK	MatheNote	Abitur88	Gruppe
0	0	3	0	2
0	0	4	0	2
0	0	4	0	2
0	0	4	0	2
1	0	3	0	2
...	...	...	...	...

Beispiel: Studienanfänger Wirtschaftswissenschaften

- 20 Studierende: 9 in Gruppe 1, 11 in Gruppe 2
- Ziel: Ordne Studierenden der Gruppe zu, zu der er gehört, ohne zu wissen, um welche Gruppe es sich handelt

$\rightarrow$ basierend auf Merkmalen

Kontingenztabelle der Merkmale Geschlecht und Gruppe

Gruppe	1	2
Geschlecht		
0	4	6
1	5	5

Beispiel: Studienanfänger Wirtschaftswissenschaften

Verteilung des Merkmals Geschlecht in den Gruppen

Gruppe	1	2
Geschlecht		
0	0.44	0.55
1	0.56	0.45

- Liegt Gruppe 1 vor, beträgt Wahrscheinlichkeit, eine weibliche Person auszuwählen 5/9 = 0.56
- Liegt Gruppe 2 vor, beträgt Wahrscheinlichkeit, eine weibliche Person auszuwählen 5/11 = 0.45

$\rightarrow$ Wahrscheinlicher, aus Gruppe 1 eine weibliche Person auszuwählen, als aus Gruppe 2

$\rightarrow$ Entscheiden uns für Gruppe, bei der Merkmalsausprägung ”weiblich” wahrscheinlicher ist

Maximum-Likelihood-Zuordnung

- Zuordnung beruht auf Likelihood-Prinzip
- Sei X das Merkmal ”Geschlecht”

$$f_i(x) = P(X = x \mid \text{Person kommt aus Gruppe } i)$$

- Es gilt:

$$f_1(0) = \frac{4}{9} = 0.44, \quad f_1(1) = \frac{7}{9} = 0.56$$

$$f_2(0) = \frac{6}{11} = 0.55 \quad f_2(1) = \frac{5}{11} = 0.45$$

Maximum-Likelihood-Zuordnung für 2 Gruppen

Definition: Maximum-Likelihood-Zuordnung

Ein Objekt mit Merkmalsausprägung x wird nach der Maximum-Likelihood-Zuordnung der Gruppe 1 zugeordnet, wenn gilt

$$\frac{f_1(x)}{f_2(x)} > 1.$$

Es wird der Gruppe 2 zugeordnet, wenn gilt

$$\frac{f_1(x)}{f_2(x)} < 1.$$

Gilt

$$\frac{f_1(x)}{f_2(x)} = 1,$$

so kann man es willkürlich einer der beiden Gruppen zuordnen.

Maximum-Likelihood-Zuordnung für r-Klassen

Definition: Maximum-Likelihood-Zuordnung

Ordne das Objekt mit Merkmalsvektor $\mathbf{x}$ derjenigen Klasse zu, für welche die Dichte maximal ist, d.h.:

$$\delta_{\mathsf{ML}}(\mathbf{x}) = r \;\Leftrightarrow\; f(\mathbf{x}|r) = \max_j f(\mathbf{x}|j)$$

Beispiel: Studienanfänger Wirtschaftswissenschaften

- Betrachten Merkmale Geschlecht und MatheLK gleichzeitig
- Merkmalspaar (x_1, x_2) mit $x_1 =$ Geschlecht und $x_2 =$ MatheLK

Gruppe (x_1, x_2)	1	2
(0,0)	0.00	0.45
(0,1)	0.44	0.09
(1,0)	0.11	0.27
(1,1)	0.44	0.18

$\rightarrow$ Ordne Studierenden Gruppe 1 zu, wenn er die Merkmalsausprägungen (0,1) oder (1,1) besitzt

Bayes-Zuordnung

- Maximum-Likelihood-Zuordnung berücksichtigt nicht, dass Populationen unterschiedlich groß sein können
- a priori-Wahrscheinlichkeiten:
  $p(r) = P(Y = r)$, $r = 1, \ldots, g$

$\rightarrow$ Multipliziere Likelihood-Funktionen mit priori-Wahrscheinlichkeiten, um Populationsgröße zu berücksichtigen

Bayes-Zuordnung, 2-Gruppen-Fall

Definition: Bayes-Zuordnung

Eine Objekt mit Merkmalsausprägung **x** wird nach der Bayes-Zuordnung der Gruppe 1 zugeordnet, wenn gilt

$$p(1)f_1(\mathbf{x}) > p(2)f_2(\mathbf{x}).$$

Es wird Gruppe 2 zugeordnet, wenn gilt

$$p(1)f_1(\mathbf{x}) < p(2)f_2(\mathbf{x}).$$

Gilt

$$p(1)f_1(\mathbf{x}) = p(2)f_2(\mathbf{x}).$$

so kann man es willkürlich einer der beiden Gruppen zuordnen.

Beispiel: Studienanfänger Wirtschaftswissenschaften

Geschlecht / Gruppe	1	2
0	4	6
1	5	5

- Es gilt: $p(1) = 0.45$ und $p(2) = 0.55$

$$
\begin{aligned}
p(1)f_1(0) &= 0.45 \cdot \frac{4}{9} = 0.2 \\
p(2)f_2(0) &= 0.55 \cdot \frac{6}{11} = 0.3 \\
p(1)f_1(1) &= 0.45 \cdot \frac{5}{9} = 0.25 \\
p(2)f_2(1) &= 0.55 \cdot \frac{5}{11} = 0.25
\end{aligned}
$$

- Person wird Gruppe 2 zugeordnet, wenn sie männlich ist
- Weibliche Person wird willkürlich einer Gruppe zugeordnet

Bayes-Zuordnung

- a posteriori-Wahrscheinlichkeiten: $P(Y=1 \mid \mathbf{x}), P(Y=0 \mid \mathbf{x})$
- Satz von Bayes:

$$\begin{aligned} P(Y=1 \mid \mathbf{x}) &= \frac{p(1)f_1(\mathbf{x})}{f(\mathbf{x})} \\ P(Y=0 \mid \mathbf{x}) &= \frac{p(2)f_2(\mathbf{x})}{f(\mathbf{x})} \end{aligned}$$

mit

$$f(\mathbf{x}) = p(1)f_1(\mathbf{x}) + p(2)f_2(\mathbf{x}).$$

- Es gilt:

$$\begin{aligned} p(1)f_1(\mathbf{x}) &= P(Y=1 \mid \mathbf{x})f(\mathbf{x}) \\ p(2)f_2(\mathbf{x}) &= P(Y=0 \mid \mathbf{x})f(\mathbf{x}) \end{aligned}$$

→ Ein Objekt mit Merkmalsausprägung **x** wird nach der Bayes-Zuordnung der Gruppe 1 zugeordnet, wenn gilt

$$P(Y=1 \mid \mathbf{x}) > P(Y=0 \mid \mathbf{x})$$

Beispiel: Studienanfänger Wirtschaftswissenschaften

	Gruppe	1	2
Geschlecht			
0		4	6
1		5	5

$$
\begin{aligned}
P(Y = 0 \mid 0) &= 0.6 \\
P(Y = 1 \mid 0) &= 0.4 \\
P(Y = 0 \mid 1) &= 0.5 \\
P(Y = 1 \mid 1) &= 0.5
\end{aligned}
$$

Bayes-Zuordnung

Definition: Bayes-Zuordnung

Ordne das Objekt mit Merkmalsvektor **x** derjenigen Klasse zu, für welche die a posteriori-Wahrscheinlichkeit maximal ist, d.h.:

$$\delta(\mathbf{x}) = r \;\Leftrightarrow\; P(r|\mathbf{x}) = \max_j P(j|\mathbf{x})\,.$$

Wie erhält man die Zuordnung?

- $P(Y = r|\mathbf{x})$ bekannt $\checkmark$
- $f(\mathbf{x}|Y = r)$ bekannt $\rightarrow$ Berechnung über den Satz von Bayes

$$P(r|\mathbf{x}) = \frac{f(\mathbf{x} \mid r)p(r)}{\sum_{j=1}^{g} f(\mathbf{x} \mid j)p(j)}$$

Fehlklassifikationswahrscheinlichkeiten

Zuordnung	wahre Gruppe			
	$Y = 1$	$Y = 2$	...	$Y = g$
$Y = 1$	✓	$P(\delta(\mathbf{x}) = 1 \mid Y = 2)$	...	$P(\delta(\mathbf{x}) = 1 \mid Y = g)$
$Y = 2$	$P(\delta(\mathbf{x}) = 2 \mid Y = 1)$	✓	...	$P(\delta(\mathbf{x}) = 2 \mid Y = g)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$Y = g$	$P(\delta(\mathbf{x}) = g \mid Y = 1)$	$P(\delta(\mathbf{x}) = g \mid Y = 2)$	...	✓

- Verwechslungswahrscheinlichkeiten: Wahrscheinlichkeit, ein Objekt der Klasse s zuzuordnen, obwohl es aus Klasse r stammt ($s \neq r$)
- Gesamtfehlerrate: Wahrscheinlichkeit mit einer Entscheidungsregel eine falsche Zuordnung zu erhalten

Fehlklassifikationswahrscheinlichkeiten

Sei δ eine bestimmte, feste Zuordnungsregel und $(\mathbf{x}, Y)$ ein Zufallsvektor.

Gesamtfehlerrate

$$\varepsilon = P\left(\delta(\mathbf{x}) \neq Y\right), \quad \delta(\mathbf{x}) \in \{1, \ldots, g\}$$

Fehlklassifikation, gegeben $\mathbf{x}$

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

Fehlklassifikationswahrscheinlichkeiten

Sei δ eine bestimmte, feste Zuordnungsregel und $(\mathbf{x}, Y)$ ein Zufallsvektor.

Verwechslungswahrscheinlichkeit

$$\varepsilon_{rs} = P\left(\delta(\mathbf{x}) = s | Y = r\right) = \int_{\mathbf{x}:\delta(\mathbf{x})=s} f(\mathbf{x}|r)dx$$

Fehlklassifikation, gegeben Klasse r

$$\varepsilon_{r} = P\left(\delta(\mathbf{x}) \neq r | Y = r\right) = \sum_{r \neq s} \varepsilon_{rs}$$

Fehlklassifikationswahrscheinlichkeiten

Es gelten folgende Zusammenhänge:

$$\varepsilon = P\left(\delta(\mathbf{x}) \neq Y\right) = \sum_{r=1}^{g} P\left(\delta(\mathbf{x}) \neq r | Y = r\right) p(r)$$

$$= \sum_{r=1}^{g} \varepsilon_r p(r) = \sum_{r=1}^{g} \sum_{s \neq r} \varepsilon_{rs} p(r)$$

und

$$\varepsilon = P\left(\delta(\mathbf{x}) \neq Y\right) = \int P\left(\delta(\mathbf{x}) \neq Y | \mathbf{x}\right) f(\mathbf{x}) d\mathbf{x}$$

$$= \int \varepsilon(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

Beispiel: Studienanfänger Wirtschaftswissenschaften

Geschlecht \ Gruppe	1	2
0	4	6
1	5	5

- Ordnen Person Gruppe 2 zu, wenn sie männlich ist
- Verwechslungswahrscheinlichkeiten:
  1. Ordne Mann aus Gruppe 1 fälschlicherweise Gruppe 2 zu:

$$\varepsilon_{12} = P\left(\delta(\mathbf{x}) = 2 | Y = 1\right) = \frac{4}{9} = 0.44$$

  2. Ordne Frau aus Gruppe 2 fälschlicherweise Gruppe 1 zu:

$$\varepsilon_{21} = P\left(\delta(\mathbf{x}) = 1 | Y = 2\right) = \frac{5}{11} = 0.45$$

- Gesamtfehlerrate:

$$\begin{aligned}\varepsilon &= P\left(\delta(\mathbf{x}) \neq Y\right) = \sum_{r=1}^{2} P\left(\delta(\mathbf{x}) \neq r | Y = r\right) p(r) \\ &= 0.44 \cdot 0.45 + 0.45 \cdot 0.55 = 0.4455\end{aligned}$$

Optimalität der Bayes-Zuordnung

Optimalität Bayes-Zuordnung

Unter allen Entscheidungsregeln besitzt die Bayes-Zuordnung dür alle $\mathbf{x}$ die kleinste bedingte Fehlerrate, wenn $\mathbf{x}$ beobachtet wird, und damit auch die kleinste Gesamtfehlerrate. Sie ist im Sinne der Gesamtfehlerrate also optimal.

Es gilt

$$\varepsilon = P\left(\delta(\mathbf{x}) \neq Y\right) = \int \varepsilon(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

$\rightarrow$ ε minimal, falls $\varepsilon(\mathbf{x})$ für alle $\mathbf{x}$ minimal

$\rightarrow$ minimiere $\varepsilon(\mathbf{x}) = 1 - P(\delta(\mathbf{x})|\mathbf{x})$

Diskriminanzfunktion

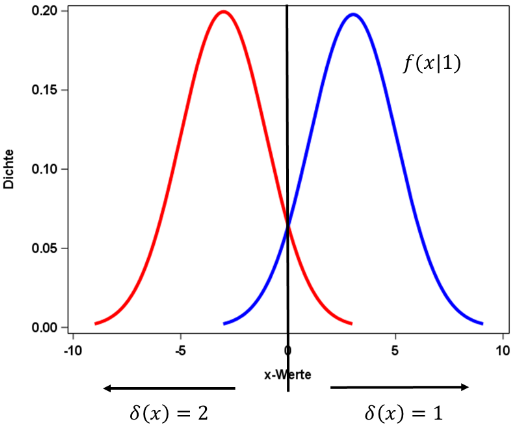
- $d_r(\mathbf{x}) = P(r|\mathbf{x})$ heißt Diskriminanzfunktion
- Formulierung der Bayes-Zuordnung:

$$\delta(\mathbf{x}) = r \Leftrightarrow d_r(\mathbf{x}) = \max_j d_j(\mathbf{x}).$$

- Äquivalent:
  - $d_r(\mathbf{x}) = f(\mathbf{x}|r)p(r)$
  - $d_r(\mathbf{x}) = \log(f(\mathbf{x}|r)) + \log(p(r))$

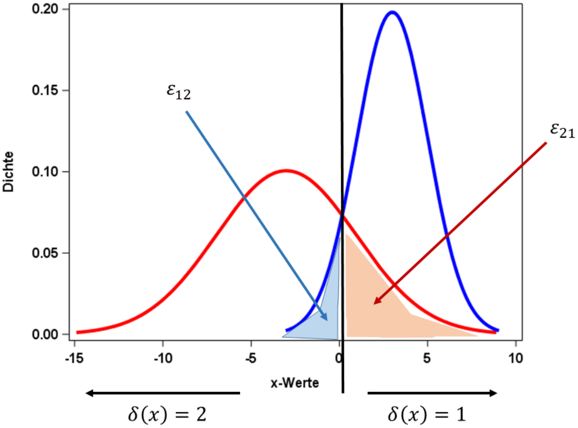
Veranschaulichung der Bayes-Zuordnung

Betrachte zwei Klassen, wobei gilt: $p(1) = p(2)$



Veranschaulichung der Bayes-Zuordnung

Betrachte zwei Klassen, wobei gilt: $p(1) > p(2)$



Kostenoptimale Bayes-Zuordnung

- Kostenfunktion:

$$c(r, \hat{r}) = c_{r\hat{r}}$$

- Kosten der Zuordnung eines Objekts der Klasse r in die Klasse $\hat{r}$ ($\hat{=}$ Risiko oder Schaden)
- Annahmen:
  - $c_{r\hat{r}} \geq 0$
  - $c_{rr} = 0$

Kostenoptimale Bayes-Zuordnung

- δ: bestimmte, feste Zuordnungsregel
- $(\mathbf{x}, Y)$: neue Beobachtung

$\rightarrow$ $c_{Y,\delta(\mathbf{x})}$ ist Zufallsvariable

- Bestimmung von δ über zu erwartenden Schaden
  $R := \mathbb{E}_{Y,\mathbf{x}}(c_{Y,\delta(\mathbf{x})})$

Bedingtes Risiko, gegeben x

$$r(\mathbf{x}) = \sum_{r=1}^{g} c_{r,\delta(\mathbf{x})} P(r|\mathbf{x}) = \mathbb{E}_{Y|\mathbf{x}}(c_{Y,\delta(\mathbf{x})})$$

$\rightarrow$ Zu erwartender Schaden bei gegebenem **x**

Kostenoptimale Bayes-Zuordnung

Als Gesamt-Risiko ergibt sich:

$$
\begin{aligned}
R = \mathbb{E}_{Y,\mathbf{x}}(c_{Y,\delta(\mathbf{x})}) &= \mathbb{E}_{\mathbf{x}}\left(\mathbb{E}_{Y|\mathbf{x}}(c_{Y,\delta(\mathbf{x})})\right) \\
&= \mathbb{E}_{\mathbf{x}}\left(\sum_{r=1}^{g} c_{r,\delta(\mathbf{x})} P(r|\mathbf{x})\right) \\
&= \int \sum_{r=1}^{g} c_{r,\delta(\mathbf{x})} P(r|\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int r(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

$\rightarrow$ Die Minimierung von $r(\mathbf{x})$ für jedes $\mathbf{x}$ ergibt eine Minimierung des Gesamt-Risikos R.

Kostenoptimale Bayes-Zuordnung

- Ordne Objekt mit Merkmalsvektor **x** derjenigen Klasse zu, für welche der zu erwartende Schaden minimal ist, d.h.:

$$\delta_K(\mathbf{x}) = r \;\Leftrightarrow\; \sum_{k=1}^{g} c_{kr} P(k|\mathbf{x}) = \min_j \sum_{k=1}^{g} c_{kj} P(k|\mathbf{x})$$

- Mit Diskriminanzfunktionen

$$d_r(\mathbf{x}) = -\sum_{k=1}^{g} c_{kr} P(k|\mathbf{x}),$$

erhält man

$$\delta_K(\mathbf{x}) = r \;\Leftrightarrow\; d_r(\mathbf{x}) = \max_j d_j(\mathbf{x})$$

Kostenoptimale Bayes-Zuordnung für 2 Gruppen

Kostenoptimale Bayes-Zuordnung, 2 Gruppen

Seien c_{12} die Kosten, die entstehen, wenn man ein Objekt, das in Gruppe 1 gehört, irrtümlich Gruppe 2 zuordnet, und c_{21} die Kosten, die entstehen, wenn man ein Objekt, das in Gruppe 2 gehört, irrtümlich Gruppe 1 zuordnet. Ein Objekt mit Merkmalsvektor $\mathbf{x}$ wird nach der kostenoptimalen Entscheidungsregel der Gruppe 1 zugeordnet, wenn gilt

$$c_{21} p(2) f_2(\mathbf{x}) < c_{12} p(1) f_1(\mathbf{x}).$$

Es wird Gruppe 2 zugeordnet, wenn gilt

$$c_{21} p(2) f_2(\mathbf{x}) > c_{12} p(1) f_1(\mathbf{x}).$$

Gilt

$$c_{21} p(2) f_2(\mathbf{x}) = c_{12} p(1) f_1(\mathbf{x}),$$

so kann man es willkürlich einer der beiden Gruppen zuordnen.

Spezialfälle

1. $c_{r\hat{r}} = c,\ r \neq \hat{r}$, d.h. jede Verwechslung hat denselben Schaden

$\rightarrow$ Bayes-Zuordnung

2. $c_{r\hat{r}} = \frac{c}{p(r)}$, d.h. Schaden ist proportional zur Größe der Klassen

$\rightarrow$ ML-Zuordnung

Diskriminanzanalyse bei normalverteilten Grundgesamtheiten

Klassifikation unter Normalverteilung

- Annahme: $\mathbf{x}|r$ multivariat normalverteilt mit Dichte:

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

- Betrachte Bayes-Zuordnungsregel ohne Kosten und zugehörige Diskriminanzfunktion

$$\begin{aligned} d_r(\mathbf{x}) &= \log(f(\mathbf{x}|r)) + \log(p(r)) \\ &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_r)^\top \boldsymbol{\Sigma}_r^{-1}(\mathbf{x} - \boldsymbol{\mu}_r) - \frac{1}{2}\log(\mid \boldsymbol{\Sigma}_r \mid) \\ &\quad -\frac{p}{2}\log(2\pi) + \log(p(r)) \end{aligned}$$

1. Spezialfall: $\mathbf{x}|r \sim N_p(\boldsymbol{\mu}_r, \sigma^2\mathbf{I})$

Diskriminanzfunktion:

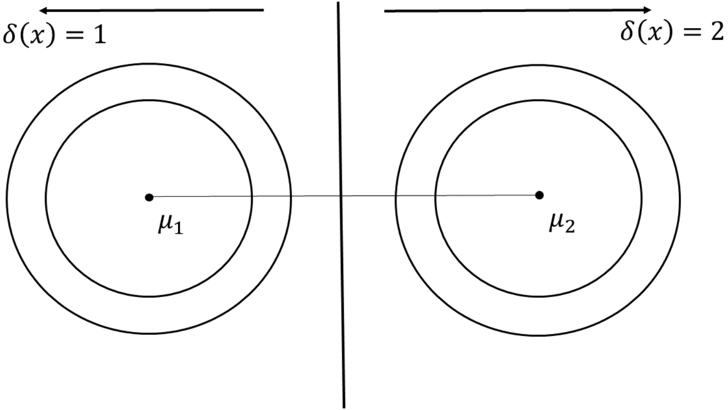
$$d_r(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_r)^\top(\mathbf{x}-\boldsymbol{\mu}_r) - \frac{1}{2}\log(|\sigma^2\mathbf{I}|) + \log(p(r))$$

$\rightarrow$ Vergleich zweier Klassen r und $\tilde{r}$ mit $p(r) = p(\tilde{r})$:

$$\begin{aligned}
d_r(\mathbf{x}) &\geq d_{\tilde{r}}(\mathbf{x}) \Leftrightarrow \\
-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_r)^\top(\mathbf{x}-\boldsymbol{\mu}_r) &\geq -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_{\tilde{r}})^\top(\mathbf{x}-\boldsymbol{\mu}_{\tilde{r}}) \\
-\frac{1}{2\sigma^2}||\mathbf{x}-\boldsymbol{\mu}_r||^2 &\geq -\frac{1}{2\sigma^2}||\mathbf{x}-\boldsymbol{\mu}_{\tilde{\mathbf{r}}}||^2 \\
||\mathbf{x}-\boldsymbol{\mu}_r||^2 &\leq ||\mathbf{x}-\boldsymbol{\mu}_{\tilde{\mathbf{r}}}||^2
\end{aligned}$$

$\rightarrow$ Objekt wird Klasse mit geringstem Abstand zum Erwartungswert zugeordnet (Minimum-Distanz-Regel)

1. Spezialfall: Skizze



2. Spezialfall: $\mathbf{x}|r \sim N_p(\boldsymbol{\mu}_r, \Sigma)$

Diskriminanzfunktion:

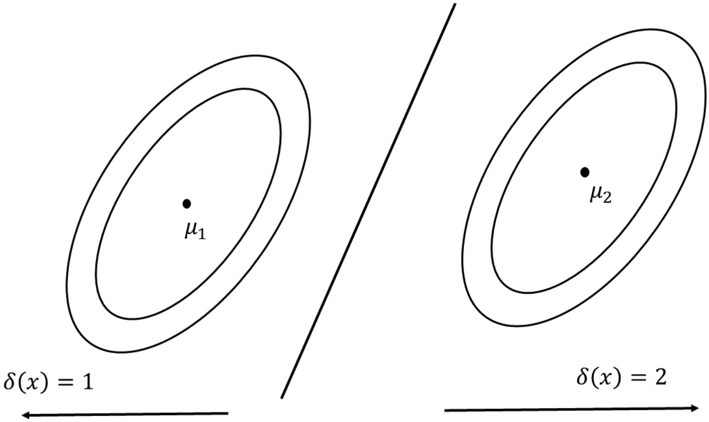
$$
\begin{aligned}
d_r(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_r) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \log(p(r)) \\
&= -\frac{1}{2}||\mathbf{x}-\boldsymbol{\mu}_r||^2_{\boldsymbol{\Sigma}^{-1}} - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \log(p(r)) \\
&= -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_r - \frac{1}{2}\boldsymbol{\mu}_r^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_r - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) + \log(p(r
\end{aligned}
$$

$\rightarrow$ Ordne Objekt der Klasse zu, zu der **x** die kleinste Mahalanobis-Distanz hat

Beachte: Terme, die nicht von r abhängen, sind irrelevant:

$$d_r(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_r - \frac{1}{2}\boldsymbol{\mu}_r^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_r + \log(p(r))$$

2. Spezialfall: Skizze



2. Spezialfall: Beispiel mit 2 Gruppen

- Annahme: $p(1) = p(2)$ (ML-Zuordnung)
- Diskriminanzfunktion für die 1. Gruppe:

$$d_1(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \frac{1}{2} \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1$$

- Diskriminanzfunktion für die 2. Gruppe:

$$d_2(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 - \frac{1}{2} \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2$$

$\rightarrow$ Ordne ein Objekt Gruppe 1 zu, wenn gilt $d_1(\mathbf{x}) > d_2(\mathbf{x})$ bzw. $d_1(\mathbf{x}) - d_2(\mathbf{x}) > 0$

$$\begin{aligned}
\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \frac{1}{2} \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \frac{1}{2} \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 &> 0 \Leftrightarrow \\
\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^\top \mathbf{x} - \frac{1}{2} \left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) &> 0 \Leftrightarrow \\
\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^\top \mathbf{x} > \frac{1}{2} \left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)
\end{aligned}$$

Lineare vs. Quadratische Diskriminanzfunktion

- $\mathbf{x}|r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$ mit Diskriminanzfunktion:

$$\begin{aligned} d_r(\mathbf{x}) &= \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_r - \frac{1}{2} \boldsymbol{\mu}_r^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_r + \log(p(r)) \\ &= \mathbf{a}_r^\top \mathbf{x} + a_{r0} \end{aligned}$$

$\rightarrow$ lineare Funktion in $\mathbf{x}$

- 3. Fall: $\mathbf{x}|r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$

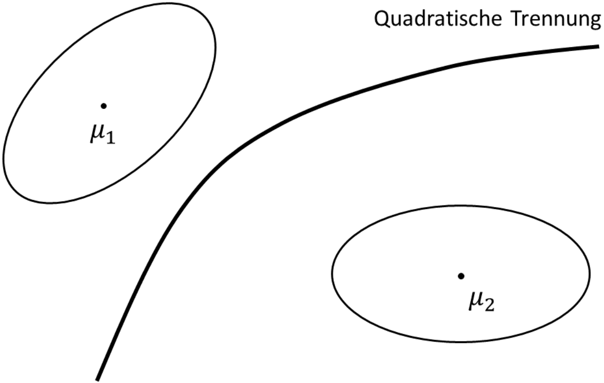
$\rightarrow$ kein Term der log-Dichte kann vernachlässigt werden

$\rightarrow$ Diskriminanzfunktion:

$$d_r(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}_r \mathbf{x} + \mathbf{a}_r^\top \mathbf{x} + a_{r0}$$

$\rightarrow$ quadratische Funktion in $\mathbf{x}$

3. Fall: Skizze



Geschätzte Zuordnungregel

- Ausgangspunkt bisher: wahre Verteilung zur Bestimmung der Zuordnung bekannt
- Jetzt: Daten $\mathbf{x}^r_{(1)}, \ldots, \mathbf{x}^r_{(n_r)},\ r = 1, \ldots, g$ gegeben (Lernstichprobe)

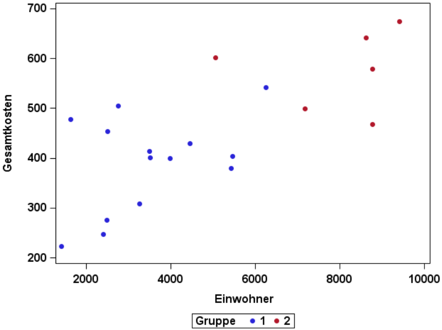
$\rightarrow$ geschätzte Diskriminanzfunktion durch Einsetzen der Schätzer $\hat{\boldsymbol{\mu}}_r = \bar{\mathbf{x}}_r$ und $\hat{\boldsymbol{\Sigma}}_r = \mathbf{S}_r$

$\rightarrow$ Für neue Beobachtung $\tilde{\mathbf{x}}$ gilt:

$$\hat{\delta}(\tilde{\mathbf{x}}) = r \ \Leftrightarrow\ d_r(\tilde{\mathbf{x}} \,|\, \bar{\mathbf{x}}_r, \mathbf{S}_r) = \max_j d_j(\tilde{\mathbf{x}} \,|\, \bar{\mathbf{x}}_j, \mathbf{S}_j)$$

Beispiel: Kreditinstitute

- Betrachtung von 20 Zweigstellen, die sich in 2 Gruppen einteilen lassen
- 14 haben hohen Marktanteil und ein überdurchschnittliches Darlehens- und Kreditgeschäft
- 6 sind technisch gut ausgestattet, besitzen überdurchschnittliches Einlage- und Kreditgeschäft und hohe Mitarbeiterzahl



Beispiel: Kreditinstitute

- Ordne Objekt der ersten Gruppe zu, wenn gilt:

$$\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^\top \mathbf{x} > \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

- Geschätzte Erwartungswerte:

$$\hat{\boldsymbol{\mu}}_1 = \bar{\mathbf{x}}_1 = \begin{pmatrix} 3510.4 \\ 390.2 \end{pmatrix} \quad \hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{x}}_2 = \begin{pmatrix} 7975.2 \\ 577.5 \end{pmatrix}$$

- Geschätzte Kovarianzmatrix:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \begin{pmatrix} 2267168.12 & 49088.07 \\ 49088.07 & 8381.77 \end{pmatrix}$$

Beispiel: Kreditinstitute

- Ordne Beobachtung **x** der ersten Gruppe zu, falls gilt:

$$\left(\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right)^\top \mathbf{x} > \frac{1}{2}\left(\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right)^\top (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

- Inverse geschätzte Kovarianzmatrix:

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.00000051 & -0.00000296 \\ -0.00000296 & 0.00013663 \end{pmatrix}$$

- Mittelwertsdifferenz:

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \begin{pmatrix} -4464.8 \\ -187.2 \end{pmatrix}$$

- Daraus ergibt sich:

$$\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \begin{pmatrix} -0.00170 \\ -0.01237 \end{pmatrix}$$

Beispiel: Kreditinstitute

Klassifizieren eine Zweigstelle mit dem Merkmalsvektor

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

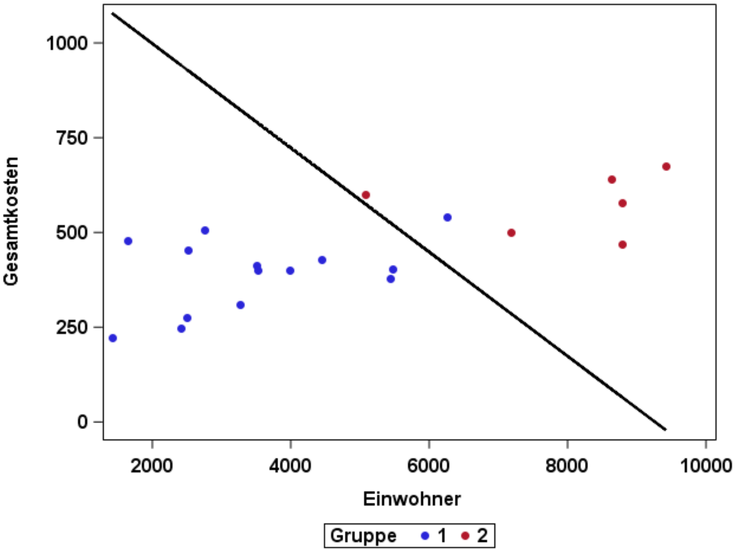
zur Gruppe 1, falls gilt:

$$\begin{pmatrix} -0.0017 & -0.01237 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} > \begin{pmatrix} -0.0017 & -0.01237 \end{pmatrix} \begin{pmatrix} 5742.8 \\ 483.8 \end{pmatrix}$$

$\rightarrow$ Vereinfachen zu:

$$0.0017 x_1 + 0.01237 x_2 < 15.747$$

Beispiel: Kreditinstitute



Eigenschaften

- Gesamtfehler für Bayes-Zuordnung unter Normalverteilung:
  - $\varepsilon(\delta(\text{quadratisch})) \;\le\; \varepsilon(\delta(\text{linear}))$
  - $\varepsilon(\hat{\delta}) \;\ge\; \varepsilon(\delta)$
- Erwartungswert $\mathbb{E}_L(\varepsilon(\hat{\delta}))$ über mehrere Lernstichproben
  - $\mathbb{E}_L(\varepsilon(\hat{\delta})) \;\ge\; \varepsilon(\delta)$
  - keine Dominanz der quadratischen Diskriminanzfunktion, d.h.:
    $$\mathbb{E}_L(\varepsilon(\hat{\delta}(\text{quadratisch}))) \;\not\leq\; \mathbb{E}_L(\varepsilon(\hat{\delta}(\text{linear})))$$

Diskriminanzanalyse nach Fisher

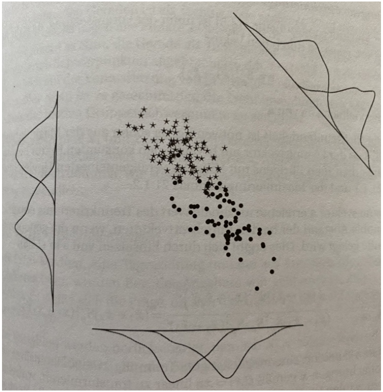
Diskriminanzanalyse nach Fisher (Zwei-Klassen-Fall)

- Gegeben: Daten $\mathbf{x}^1_{(1)}, \ldots, \mathbf{x}^1_{(n_1)}$ und $\mathbf{x}^2_{(1)}, \ldots, \mathbf{x}^2_{(n_2)}$
- Ziel: Finde Projektion, d.h. eine Linearkombination

$$y = \mathbf{a}^\top \mathbf{x}, \quad \text{mit} \quad \|\mathbf{a}\| = 1,$$

  sodass die beiden Klassen bestmöglich getrennt werden
- Anforderungen:
  - Streuung zwischen Gruppen möglichst groß
  - Streuung innerhalb der Gruppen möglichst klein

Grafische Veranschaulichung



Quelle: R. Schlittgen: Multivariate Statistik. 2009, Oldenbourg

Kriterium von Fisher

Maximiere folgendes Kriterium, um eine maximale Trennung der beiden Klassen zu erzielen:

$$Q(\mathbf{a}) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{w_1^2 + w_2^2}\,,$$

wobei

$$\bar{y}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} \mathbf{a}^\top \mathbf{x}_{(i)}^r = \mathbf{a}^\top \bar{\mathbf{x}}_r, \quad r = 1, 2\,,$$

$$w_r^2 = \sum_{i=1}^{n_r} (\mathbf{a}^\top \mathbf{x}_{(i)}^r - \mathbf{a}^\top \bar{\mathbf{x}}_r)^2, \quad r = 1, 2\,.$$

Kriterium von Fisher

Es gilt:

$$\begin{aligned} w_1^2 + w_2^2 &= \sum_{r=1}^{2} \sum_{i=1}^{n_r} (\mathbf{a}^\top \mathbf{x}_{(i)}^r - \mathbf{a}^\top \bar{\mathbf{x}}_r)^2 \\ &= \mathbf{a}^\top \sum_{r=1}^{2} \sum_{i=1}^{n_r} (\mathbf{x}_{(i)}^r - \bar{\mathbf{x}}_r)(\mathbf{x}_{(i)}^r - \bar{\mathbf{x}}_r)^\top \mathbf{a} \\ &= \mathbf{a}^\top \mathbf{W}\, \mathbf{a} \end{aligned}$$

Und damit:

$$Q(\mathbf{a}) = \frac{(\mathbf{a}^\top (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{\mathbf{a}^\top \mathbf{W}\, \mathbf{a}} \;\to\; \max_{\mathbf{a} \neq \mathbf{0}}$$

Kriterium nach Fisher

Maximierung durch Bilden der Ableitung und lösen der Gleichung: $\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} = 0$.

Damit gilt:

$$\begin{aligned}\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} &= \frac{2(\mathbf{a}^\top\bar{\mathbf{x}}_1 - \mathbf{a}^\top\bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{a}^\top\mathbf{W}\mathbf{a} - 2\mathbf{W}\mathbf{a}(\mathbf{a}^\top\bar{\mathbf{x}}_1 - \mathbf{a}^\top\bar{\mathbf{x}}_2)^2}{(\mathbf{a}^\top\mathbf{W}\mathbf{a})^2} \\ &= 0 \Leftrightarrow\end{aligned}$$

$$\begin{aligned}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{a}^\top\mathbf{W}\mathbf{a} - \mathbf{W}\mathbf{a}(\mathbf{a}^\top\bar{\mathbf{x}}_1 - \mathbf{a}^\top\bar{\mathbf{x}}_2) &= 0 \Leftrightarrow \\ \mathbf{W}\mathbf{a}\underbrace{\left(\frac{\mathbf{a}^\top\bar{\mathbf{x}}_1 - \mathbf{a}^\top\bar{\mathbf{x}}_2}{\mathbf{a}^\top\mathbf{W}\mathbf{a}}\right)}_{konst} &= \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\end{aligned}$$

$\rightarrow$ Wähle $\mathbf{a} = \mathbf{W}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

Vergleich zur linearen Diskriminanzanalyse

Unter Normalverteilungsannahme mit gleichen Kovarianzmatrizen gilt für den Vergleich zweier Diskriminanzfunktionen

$$\begin{aligned} & d_1(\mathbf{x}) = d_2(\mathbf{x}) \\ \Leftrightarrow\ & \mathbf{a}_1^\top \mathbf{x} + a_{10} = \mathbf{a}_2^\top \mathbf{x} + a_{20} \\ \Leftrightarrow\ & (\mathbf{a}_1 - \mathbf{a}_2)^\top \mathbf{x} = a_{20} - a_{10} \end{aligned}$$

mit $a_1 = \Sigma^{-1}\mu_1$ und $a_2 = \Sigma^{-1}\mu_2$.
$\rightarrow$ Schätzung aus Datensatz: $(\mathbf{a}_1 - \mathbf{a}_2) = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \rightarrow$
Kriterium nach Fisher mit $\mathbf{S} = \frac{1}{n_1+n_2-2}\mathbf{W}$

Beispiel: Studienanfänger

Geschlecht	MatheLK	MatheNote	Abitur88	Gruppe
0	0	3	0	2
0	0	4	0	2
0	0	4	0	2
0	0	4	0	2
1	0	3	0	2
...	...	...	...	...

- Mittelwerte der Gruppen:

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 0.556 \\ 0.889 \\ 2.667 \\ 0.333 \end{pmatrix} \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 0.455 \\ 0.273 \\ 3.182 \\ 0.364 \end{pmatrix}$$

Beispiel: Studienanfänger

$$\mathbf{S}_1 = \begin{pmatrix} 0.278 & -0.056 & 0.208 & 0.042 \\ -0.056 & 0.111 & -0.167 & -0.083 \\ 0.208 & -0.167 & 1.000 & 0.125 \\ 0.042 & -0.083 & 0.125 & 0.250 \end{pmatrix}$$

$$\mathbf{S}_2 = \begin{pmatrix} 0.273 & 0.064 & -0.191 & 0.218 \\ 0.064 & 0.218 & -0.255 & 0.091 \\ -0.191 & -0.255 & 0.564 & -0.173 \\ 0.218 & 0.091 & -0.173 & 0.255 \end{pmatrix}$$

$$\mathbf{W} = \begin{pmatrix} 4.954 & 0.192 & -0.246 & 2.516 \\ 0.192 & 3.068 & -3.886 & 0.246 \\ -0.246 & -3.886 & 13.640 & -0.730 \\ 2.516 & 0.246 & -0.730 & 4.550 \end{pmatrix}$$

$$\mathbf{W}^{-1} = \begin{pmatrix} 0.2812 & -0.0145 & -0.0074 & -0.1559 \\ -0.0145 & 0.5108 & 0.1455 & 0.0037 \\ -0.0074 & 0.1455 & 0.1154 & 0.0147 \\ -0.1559 & 0.0037 & 0.0147 & 0.3082 \end{pmatrix}$$

Beispiel: Studienanfänger

Es ergibt sich

$$\mathbf{a} = \mathbf{W}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \begin{pmatrix} 0.028 \\ 0.238 \\ 0.029 \\ -0.030 \end{pmatrix}$$

$\rightarrow \mathbf{a}^\top \bar{\mathbf{x}}_1 = \bar{y}_1 = 0.295, \mathbf{a}^\top \bar{\mathbf{x}}_2 = \bar{y}_2 = 0.159$
$\rightarrow$ Ordne Objekt Gruppe 1 zu, wenn gilt:

$$|\mathbf{a}^\top \mathbf{x} - \bar{y}_1| < |\mathbf{a}^\top \mathbf{x} - \bar{y}_2|.$$

Erster Student:

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 4 \\ 1 \end{pmatrix}$$

Es gilt $\mathbf{a}^\top \mathbf{x} = 0.114$. $\rightarrow$ Gruppe 2

Diskriminanzanalyse nach Fisher (Mehr-Klassen-Fall)

- Gegeben: Daten $\mathbf{x}_{(1)}^r, \ldots, \mathbf{x}_{(n_r)}^r$, $r = 1, \ldots, g$
- Gesucht: Projektion $y = \mathbf{a}^\top \mathbf{x}$, welche durch Maximierung von

$$Q(\mathbf{a}) = \frac{\sum_{r=1}^{g} n_r (\bar{y}_r - \bar{y})^2}{\sum_{r=1}^{g} w_r^2}$$

bestimmt ist, wobei

$$\bar{y} = \frac{1}{n} \sum_{r=1}^{g} n_r \bar{y}_r$$

Kriterium nach Fisher

- Zähler von $Q(\mathbf{a})$:

$$\sum_{r=1}^{g} n_r(\bar{y}_r - \bar{y})^2 = \sum_{r=1}^{g} n_r(\mathbf{a}^\top \bar{\mathbf{x}}_r - \mathbf{a}^\top \bar{\mathbf{x}})^2$$

$$= \mathbf{a}^\top \sum_{r=1}^{g} n_r(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^\top \mathbf{a} = \mathbf{a}^\top \mathbf{B}\mathbf{a}$$

- Mit dem Resultat von Folie 54:

$$Q(\mathbf{a}) = \frac{\mathbf{a}^\top \mathbf{B}\,\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\,\mathbf{a}} \rightarrow \max_{\mathbf{a} \neq \mathbf{0}}$$

Kriterium nach Fisher

Lösung mithilfe der Gleichung: $\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} = 0$

$$
\begin{aligned}
\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} &= \frac{2\mathbf{Ba}(\mathbf{a}^\top \mathbf{Wa}) - 2(\mathbf{a}^\top \mathbf{Ba})\mathbf{Wa}}{(\mathbf{a}^\top \mathbf{Wa})^2} \\
&= \frac{2(\mathbf{a}^\top \mathbf{Wa})[\mathbf{Ba} - \frac{\mathbf{a}^\top \mathbf{Ba}}{\mathbf{a}^\top \mathbf{Wa}}\mathbf{Wa}]}{(\mathbf{a}^\top \mathbf{Wa})^2} \\
&= \frac{2(\mathbf{Ba} - Q(\mathbf{a})\mathbf{Wa})}{\mathbf{a}^\top \mathbf{Wa}} = 0
\end{aligned}
$$

$$
\begin{aligned}
\Leftrightarrow \mathbf{Ba} &= Q(\mathbf{a})\mathbf{Wa} \\
\Leftrightarrow \mathbf{W}^{-1}\mathbf{Ba} &= \underbrace{Q(\mathbf{a})}_{\text{Skalar } \lambda}\, \mathbf{a} = \lambda \mathbf{a}
\end{aligned}
$$

$\rightarrow$ Es ergibt sich ein verallgemeinertes Eigenwertproblem!

Lösung des Eigenwertproblems

- Generelle Form verallgemeinertes Eigenwertproblem:

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}\,,$$

  wobei **W** und **B** symmetrisch und **W** außerdem positiv definit
- Möglicher Lösungsansatz:
- Zerlege **W** mittels Cholesky-Zerlegung: $\mathbf{W} = \mathbf{L}\mathbf{L}^\top$
- Eigenwerte der Matrix $\mathbf{H} := \mathbf{L}^{-1}\mathbf{B}(\mathbf{L}^{-1})^\top$ sind identisch zu Eigenwerten von $\mathbf{W}^{-1}\mathbf{B}$

Lösung des Eigenwertproblems

- $\mathrm{rg}(\mathbf{W}) = p$ und $\mathrm{rg}(\mathbf{B}) = q \leq \min\{p, g-1\}$

$\rightarrow$ $\mathbf{W}^{-1}\mathbf{B}$ hat höchstens q Eigenwerte $\lambda_1, \ldots, \lambda_q$ mit zugehörigen Eigenvektoren $\mathbf{a}_1, \ldots, \mathbf{a}_q$

- Lösungen:

$$\lambda_r = \frac{\mathbf{a}_r^\top \mathbf{B}\, \mathbf{a}_r}{\mathbf{a}_r^\top \mathbf{W}\, \mathbf{a}_r}, \quad r = 1, \ldots, q\,,$$

und die ”kanonischen Variablen”

$$y_r = \mathbf{a}_r^\top \mathbf{x}, \quad r = 1, \ldots, q$$

Praktisches Vorgehen

- Ordne die Eigenwerte $\lambda_1, \dots, \lambda_q$ der Größe nach und verwende alle oder nur $m \leq q$ Komponenten (Projektionsrichtungen)
- Betrachte die Entscheidungsregel:

$$\delta(\mathbf{x}) = r \;\Leftrightarrow\; \sum_{r=1}^{m} (\mathbf{a}_r\mathbf{x} - \mathbf{a}_r^\top \bar{\mathbf{x}}_r)^2 = \min_j \sum_{r=1}^{m} (\mathbf{a}_r\mathbf{x} - \mathbf{a}_r^\top \bar{\mathbf{x}}_j)^2$$

- Beachte: Es kann (wieder) gezeigt werden, dass dieses Kriterium äquivalent zur ML-Zuordnung unter Normalverteilungsannahme mit gleichen Kovarianzmatrizen ist

k-nächste Nachbarn

Klassifikation anhand der k-nächsten Nachbarn

- Betrachte Gesamtstichprobe $(\mathbf{x}_i, Y_i)$, $i = 1, \ldots, n$
- Bestimme zu jedem Merkmalsvektor $\mathbf{x}_i$ diejenigen Merkmalsvektoren, die am nähesten an $\mathbf{x}_i$ liegen
- Berechne dafür die Distanzen

$$d(\mathbf{x}_i, \mathbf{x}_s), \quad i \neq s\,,$$

über ein geeignetes Distanzmaß d, z.B. die quadrierte euklidische Distanz

Klassifikation anhand der k-nächsten Nachbarn

- Bestimme zu $\mathbf{x}_i$ die k nächsten Nachbarn
- Bezeichnung: $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$ mit

$$d(\mathbf{x}_i, \mathbf{x}_{(1)}) \leq d(\mathbf{x}_i, \mathbf{x}_{(2)}) \leq \ldots \leq d(\mathbf{x}_i, \mathbf{x}_{(k)}) \,.$$

- Bezeichnet $Y_{(1)}, \ldots, Y_{(k)}$ die Klasse zu $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$, so ergibt sich die Zuordnungsregel:

$$\delta(\mathbf{x}_i) = r \;\Leftrightarrow\; r \text{ ist die häufigste Klasse in } \{Y_{(1)}, \ldots, Y_{(k)}\}$$

k-nächsten Nachbarn: Eigenschaften

- Verteilungsfreies bzw. nichtparametrisches Verfahren, d.h. es gibt keine Annahme zur Verteilung von $\mathbf{x}|r$
- Stellschrauben des Verfahrens sind die Distanz d und die Anzahl der nächsten Nachbarn k

Logistische Diskriminanzanalyse

Logistische Regression (Zwei-Klassen-Fall)

- Ausgehend von Zufallsvektoren $(\mathbf{x}, Y)$ postuliert man für die a posteriori-Wahrscheinlichkeiten:

$$P(Y = 1|\mathbf{x}) = \frac{\exp(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})}{1 + \exp(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})} \quad \text{und}$$

$$P(Y = 2|\mathbf{x}) = 1 - P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})}$$

- Äquivalent kann postuliert werden:

$$\log\left(\frac{f(\mathbf{x}|1)}{f(\mathbf{x}|2)}\right) = \tilde{\beta}_0 + \mathbf{x}^\top \boldsymbol{\beta}, \quad \text{wobei} \quad \tilde{\beta}_0 = \beta_0 + \log\left(\frac{p(1)}{p(2)}\right)$$

Logistische Regression: Zuordnungsregeln

- Diskriminanzfunktion:

$$d(\mathbf{x}) = \ln\left(\frac{P(Y=1|\mathbf{x})}{P(Y=2|\mathbf{x})}\right) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$$

→ Ordne Objekt mit Merkmalsvektor **x** gemäß Bayes-Zuordnungsregel Klasse 1 zu, falls

$$d(\mathbf{x}) \geq 0$$

- Ordne das Objekt ansonsten Klasse 2 zu
- Ersetzen von β_0 durch $\tilde{\beta}_0$ ergibt Zuordnung gemäß ML-Zuordnungsregel

Logistische Regression: Eigenschaften

- Keine Annahme bzgl. Verteilung von $\mathbf{x}|r$
- Einfache Verallgemeinerung für g Klassen möglich
- Ansatz der logistischen Regression ist für Reihe von Klassendichten erfüllt, z.B. für Multinomialverteilung mit gleichen Kovarianzmatrizen
- Parameter β_0 und $\boldsymbol{\beta}$ sind in der Regel unbekannt und müssen geschätzt werden (analog zur Beschreibung auf Folie 44).