

# Multivariate Verfahren

## 6. Diskriminanzanalyse

Moritz Berger

Institut für Statistik, LMU München

Sommersemester 2019

# Ausgangssituation

Die Grundgesamtheit zerfällt in  $g \geq 2$  disjunkte Klassen mit Indikator  $Y \in \{1, \dots, g\}$ .

Man beobachtet Merkmalsvektoren  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , die Klassenzugehörigkeit  $Y_1, \dots, Y_n$  ist jedoch *unbekannt*.

## Problemstellung

Die Objekte  $a_1, \dots, a_n$  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

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

# Datengrundlage

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

- a priori-Wahrscheinlichkeiten:

$$p(r) = P(Y = r), \quad r = 1, \dots, g$$

- a posteriori-Wahrscheinlichkeiten:

$$P(r|\mathbf{x}) = P(Y = r|\mathbf{x}), \quad r = 1, \dots, g$$

- die Dichte von  $\mathbf{x}$  gegeben die Klasse:

$$f(\mathbf{x}|Y = 1), \dots, f(\mathbf{x}|Y = g)$$

- die Mischverteilung in der Gesamtpopulation:

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

# Fehlklassifikationswahrscheinlichkeiten

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

## Gesamt-Fehlerrate

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

## Fehlklassifikation, gegeben $\mathbf{x}$

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

# Fehlklassifikationswahrscheinlichkeiten

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

## Verwechslungswahrscheinlichkeit

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

## Fehlklassifikation, gegeben Klasse $r$

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

# Fehlklassifikationswahrscheinlichkeiten

Es gelten folgende Zusammenhänge:

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

und

$$\begin{aligned}\varepsilon &= P(\delta(\mathbf{x}) \neq Y) = \int P(\delta(\mathbf{x}) \neq Y | \mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\ &= \int \varepsilon(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}\end{aligned}$$



# Bayes-Zuordnung

Ordne das Objekt mit Merkmalsvektor  $\mathbf{x}$  in diejenige 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 ✓
- $f(\mathbf{x}|Y = r)$  bekannt  $\rightarrow$  Berechnung über den Satz von Bayes

$$P(r|\mathbf{x}) =$$

# Beispiel: Drogenkonsum

# Beispiel: Drogenkonsum - Fehlerraten

# Optimalität der Bayes-Zuordnung

Aus dem Zusammenhang auf Folie 8 kann man direkt sehen, dass die Gesamt-Fehlerrate  $\varepsilon$  minimal wird, falls  $\varepsilon(\mathbf{x})$  für alle  $\mathbf{x}$  minimal ist.

Somit ergibt sich die beste Regel im Sinne einer kleinstmöglichen Gesamt-Fehlerrate durch minimieren von  $\varepsilon(\mathbf{x}) = 1 - P(\delta(\mathbf{x})|\mathbf{x})$ .

⇒ Die Bayes-Zuordnung minimiert die Gesamtfehlerrate  $\varepsilon$ .

# Die Diskriminanzfunktion

Die Funktion  $d_r(\mathbf{x}) = P(r|\mathbf{x})$  heißt Diskriminanzfunktion.

Damit lässt sich die Bayes-Zuordnungsregel wie folgt formulieren:

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

Äquivalent dazu, sind die Diskriminanzfunktionen:

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

# Veranschaulichung der Bayes-Zuordnung

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

# Veranschaulichung der Bayes-Zuordnung

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

# Maximum-Likelihood-Zuordnung

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

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

Die ML-Zuordnung entspricht der Bayes-Zuordnung ohne Berücksichtigung der a priori-Wahrscheinlichkeiten bzw. mit gleichen a priori-Wahrscheinlichkeiten  $p(1) = \dots = p(r) = \frac{1}{r}$ .



# Kostenoptimale Bayes-Zuordnung

Betrachte die 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

Sei  $\delta$  eine bestimmte, feste Zuordnungsregel und  $(\mathbf{x}, Y)$  eine neue Beobachtung, so ist  $c_{Y,\delta(\mathbf{x})}$  eine Zufallsvariable.

Zur Bestimmung von  $\delta$  betrachtet man den zu erwartenden Schaden  $R := \mathbb{E}(c_{Y,\delta(\mathbf{x})})$ .

## Bedingtes Risiko, gegeben $\mathbf{x}$

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

Zu erwartender Schaden bei gegebenem  $\mathbf{x}$ .

# Kostenoptimale Bayes-Zuordnung

Das Gesamt-Risiko ergibt sich zu

$$R = \mathbb{E}_Y(c_{Y,\delta(\mathbf{x})}) =$$

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

# Kostenoptimale Bayes-Zuordnung

Ordne das Objekt mit Merkmalsvektor  $\mathbf{x}$  in diejenige 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 den Diskriminanzfunktionen

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

erhält man

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

# Spezialfälle

①  $c_{r\hat{r}} = c, r \neq \hat{r},$

d.h. jede Verwechslung hat den selben 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

# Klassifikation unter Normalverteilung

Im Folgenden wird angenommen, dass  $\mathbf{x}|r$  multivariat normalverteilt ist mit Dichte:

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

Betrachte die Bayes-Zuordnungsregel *ohne* Kosten und die zugehörige Diskriminanzfunktion

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

=

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

Die Diskriminanzfunktion ergibt sich zu:

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

Für den Vergleich zweier Klassen  $r$  und  $\tilde{r}$  folgt damit:

$$d_r(\mathbf{x}) \geq d_{\tilde{r}}(\mathbf{x}) \Leftrightarrow$$

# 1. Spezialfall: Skizze



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

Die Diskriminanzfunktion ergibt sich zu:

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

*Beachte:* Alle Terme, die nicht von der Klasse  $r$  abhängen, sind für den Vergleich der Diskriminanzfunktionen irrelevant und können daher vernachlässigt werden. Damit gilt:

$$d_r(\mathbf{x}) =$$

## 2. Spezialfall: Skizze

# Lineare vs. Quadratische Diskriminanzfunktion

Falls  $\mathbf{x}|r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$  ergibt sich die Diskriminanzfunktion:

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

→ Dies ist eine *lineare* Funktion in  $\mathbf{x}$ .

Im 3. Fall, wenn  $\mathbf{x}|r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ , kann kein Term der log-Dichte vernachlässigt werden und die Diskriminanzfunktion hat die Gestalt:

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

→ Dies ist eine *quadratische* Funktion in  $\mathbf{x}$ .

### 3. Fall: Skizze

# Geschätzte Zuordnungsregel

Bisher wurde davon ausgegangen, dass die wahre Verteilung zur Bestimmung der Zuordnung bekannt ist.

Sind Daten  $\mathbf{x}_{(1)}^r, \dots, \mathbf{x}_{(n_r)}^r$ ,  $r = 1, \dots, g$ , gegeben (Lernstichprobe), so erhält man eine geschätzte Diskriminanzfunktion, indem man die Schätzer  $\hat{\boldsymbol{\mu}}_r = \bar{\mathbf{x}}_r$  und  $\hat{\boldsymbol{\Sigma}}_r = \mathbf{S}_r$  einsetzt.

Somit gilt für eine neue Beobachtung  $\tilde{\mathbf{x}}$ :

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

# Eigenschaften

Für die Bayes-Zuordnung unter Normalverteilung gilt für den Gesamt-Fehler:

- $\varepsilon(\delta(\text{quadratisch})) \leq \varepsilon(\delta(\text{linear}))$
- $\varepsilon(\hat{\delta}) \geq \varepsilon(\delta)$

Betrachtet man den Erwartungswert  $\mathbb{E}_L(\varepsilon(\hat{\delta}))$  über mehrere Lernstichproben, so gilt auch  $\mathbb{E}_L(\varepsilon(\hat{\delta})) \geq \varepsilon(\delta)$ , jedoch 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

## (Zwei-Klassen-Fall)

Gegeben sind die Daten  $\mathbf{x}_{(1)}^1, \dots, \mathbf{x}_{(n_1)}^1$  und  $\mathbf{x}_{(1)}^2, \dots, \mathbf{x}_{(n_2)}^2$ .

Gesucht ist eine 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.

# Graphische Veranschaulichung



# Kriterium von Fisher

Um eine maximale Trennung der beiden Klassen zu erzielen, wird folgendes Kriterium maximiert:

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

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:

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

# Kriterium nach Fisher

Die Lösung ergibt sich durch Lösen der Gleichung:  $\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} = 0$ .

Also:

$$\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} =$$

# Vergleich zur linearen Diskriminanzanalyse

Unter Normalverteilungsannahme mit gleichen Kovarianzmatrizen gilt für den Vergleich zweier Diskriminanzfunktionen

$$d_1(\mathbf{x}) = d_2(\mathbf{x})$$

$$\Leftrightarrow \mathbf{a}_1^\top \mathbf{x} + a_{10} = \mathbf{a}_2^\top \mathbf{x} + a_{20}$$

$$\Leftrightarrow (\mathbf{a}_1 - \mathbf{a}_2)^\top \mathbf{x} = a_{20} - a_{10}$$

Da  $(\mathbf{a}_1 - \mathbf{a}_2) = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$  gilt, ergibt sich in diesem Fall das Kriterium nach Fisher.

# Diskriminanzanalyse nach Fisher

## (Mehr-Klassen-Fall)

Gegeben sind die Daten  $\mathbf{x}_{(1)}^r, \dots, \mathbf{x}_{(n_r)}^r$ ,  $r = 1, \dots, g$ .

Gesucht ist eine 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

Der Zähler von  $Q(\mathbf{a})$  ergibt:

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

Und damit (mit dem Resultat auf Folie 34):

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

# Kriterium nach Fisher

Die Lösung ergibt sich durch Lösen der Gleichung:  $\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} = 0$ .

Also:

$$\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} =$$

⇒ Es ergibt sich ein verallgemeinertes Eigenwertproblem!

# Lösung des Eigenwertproblems

Das verallgemeinerte Eigenwertproblem hat generell die Form

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

wobei  $\mathbf{W}$  und  $\mathbf{B}$  symmetrisch sind und  $\mathbf{W}$  außerdem positiv definit.

*Möglicher Lösungsansatz:*

Zerlege  $\mathbf{W}$  mittels Cholesky-Zerlegung:  $\mathbf{W} = \mathbf{L}\mathbf{L}^\top$ . Die Eigenwerte der Matrix  $\mathbf{H} := \mathbf{L}^{-1}\mathbf{B}(\mathbf{L}^{-1})^\top$  sind dann identisch zu den Eigenwerten von  $\mathbf{W}^{-1}\mathbf{B}$ .



# Lösung des Eigenwertproblems

Da  $\text{rg}(\mathbf{W}) = p$  und  $\text{rg}(\mathbf{B}) = q \leq \min\{p, g - 1\}$  ist, besitzt  $\mathbf{W}^{-1}\mathbf{B}$  höchstens  $q$  Eigenwerte  $\lambda_1, \dots, \lambda_q$  mit zugehörigen Eigenvektoren  $\mathbf{a}_1, \dots, \mathbf{a}_q$ .

Man erhält die Lösungen:

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

und die “kanonischen Variablen”

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

# Graphische Veranschaulichung

# 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 ist zur ML-Zuordnung unter Normalverteilungsannahme mit gleichen Kovarianzmatrizen.

# Klassifikation anhand der k-nächsten Nachbarn

Betrachte die Gesamtstichprobe  $(\mathbf{x}_i, Y_i)$ ,  $i = 1, \dots, n$ .

Bestimme zu jedem Merkmalsvektor  $\mathbf{x}_i$  diejenigen

Merkmalsvektoren, die *am nächsten* 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, bezeichnet durch  $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(k)}$  mit

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

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

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

# k-nächsten Nachbarn: Eigenschaften

- Die k-nächste Nachbarn Klassifikation ist ein verteilungsfreies bzw. nichtparametrisches Verfahren, d.h. es gibt keine Annahme an die Verteilung von  $\mathbf{x}|r$ .
- Stellschrauben des Verfahrens sind die Distanz  $d$  und die Anzahl der nächsten Nachbarn  $k$ .

# Logistische Regression (Zwei-Klassen-Fall)

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

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

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

Äquivalent kann postuliert werden:

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

# Logistische Regression: Zuordnungsregeln

Ordne das Objekt mit Merkmalsvektor  $\mathbf{x}$  gemäß

*Bayes-Zuordnungsregel* in Klasse 1 zu, falls

$$d(\mathbf{x}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta} \geq 0,$$

ansonsten zu Klasse 2.

Wird  $\beta_0$  durch  $\tilde{\beta}_0$  ersetzt erhält man eine Zuordnung gemäß *ML-Zuordnungsregel*.



# Logistische Regression: Eigenschaften

- Es wird keine Annahme an die Verteilung von  $\mathbf{x}|r$  gemacht.
- Einfache Verallgemeinerung für  $g$  Klassen ist möglich.
- Der Ansatz der (linearen) logistischen Regression ist für eine Reihe von Klassendichten erfüllt, z.B. für die Multinomialverteilung mit gleichen Kovarianzmatrizen.
- Die Parameter  $\beta_0$  und  $\beta$  sind in der Regel unbekannt und müssen geschätzt werden (analog zur Beschreibung auf Folie 29).