Multivariate Verfahren

MSc Klinische Psychologie und Psychotherapie

MSc Umweltpsychologie/Mensch-Technik-Interaktion

MSc Psychologie

Joram Soch | Wintersemester 2025/2026

(10) Bayes-Klassifikation

Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

Prävalenz der Alzheimer-Krankheit liegt bei 2% für 65-70-Jährige.

Prävalenz steigt dann exponentiell bis Mitte 80 Lebensjahre.

Durch höhere Lebenserwartung steigt die Prävalenz in der Gesamtbevölkerung.

Standardmäßig erfolgt die Diagnose durch den Nachweis von Beta-Amyloid ("Plaques") im Gehirn sowie den Nachweis des Tau-Proteins im Liquor (CSF).

Kostengünstige Idenfitikation von Alzheimer-Erkrankten wäre wünschenswert.

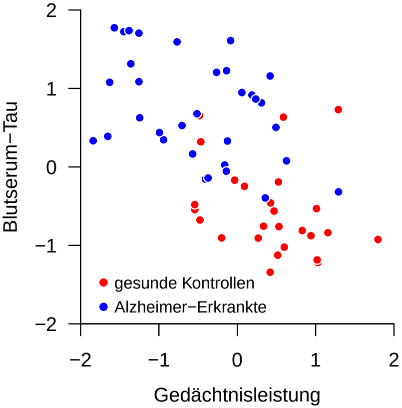
- kognitive Ergebnismaße, z.B. Gedächtnistests
- Biomarker-Variablen, z.B. Blutabnahme
- Lifestyle-Faktoren, z.B. Bewegung/Rauchen

Anwendungsbeispiele

- Gedächtnisleistung und Blutserum-Tau als Indikatoren für Alzheimer-Erkrankung
- Lineare Diskriminanzanalyse, Logistische Regression, Neuronale Netze, SVMs
- Zhao et al. (2024) geben eine aktuelle Übersicht zur Linearen Diskriminanzanalyse.

Beispieldatensatz

Gedächtnisleistung und Blutserum-Tau als Indikatoren für Alzheimer-Erkrankung



Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Definition (Bernoulli-Zufallsvariable)

Es sei y eine Zufallsvariable mit Ergebnisraum $\mathcal{Y} := \{0, 1\}$ und WMF

$$p : \mathcal{Y} \to [0, 1], y \mapsto p(y) := \mu^y \, (1 - \mu)^{1-y} \quad \text{mit} \quad \mu \in [0, 1]. \tag{1}$$

Dann sagen wir, dass y einer *Bernoulli-Verteilung mit Parameter* $\mu \in [0, 1]$ unterliegt und nennen y eine *Bernoulli-Zufallsvariable*. Wir kürzen dies mit $y \sim \text{Bern}(\mu)$ ab. Die WMF einer Bernoulli-Zufallsvariable bezeichnen wir mit

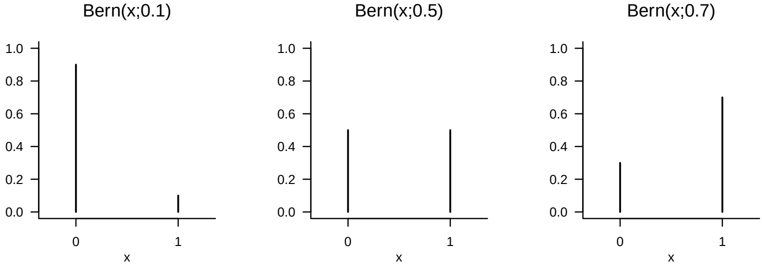
$$\text{Bern}(y; \mu) := \mu^y \, (1 - \mu)^{1-y}. \tag{2}$$

Bemerkungen

- Eine Bernoulli-Zufallsvariable kann als Modell eines Münzwurfs dienen.
- μ ist die Wahrscheinlichkeit dafür, dass y den Wert 1 annimmt:

$$\mathbb{P}(y = 1) = \mu^1 (1 - \mu)^{1-1} = \mu. \tag{3}$$

Beispiele für Bernoulli-Zufallsvariablen



Definition (Multivariate Normalverteilung)

x sei ein n-dimensionaler Zufallsvektor mit Ergebnisraum $\mathbb{R}^n$ und WDF

$$p : \mathbb{R}^n \to \mathbb{R}_{>0},\, x \mapsto p(x) := \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right). \tag{4}$$

Dann sagen wird, dass x einer *multivariaten (oder n-dimensionalen) Normalverteilung* mit *Erwartungswertparameter* $\mu \in \mathbb{R}^n$ und positiv-definitem *Kovarianzmatrixparameter* $\Sigma \in \mathbb{R}^{n\times n}$ unterliegt und nennen x einen *(multivariat) normalverteilten Zufallsvektor*. Wir kürzen dies mit $x \sim N(\mu, \Sigma)$ ab. Die WDF eines multivariat normalverteilten Zufallsvektors bezeichnen wir mit

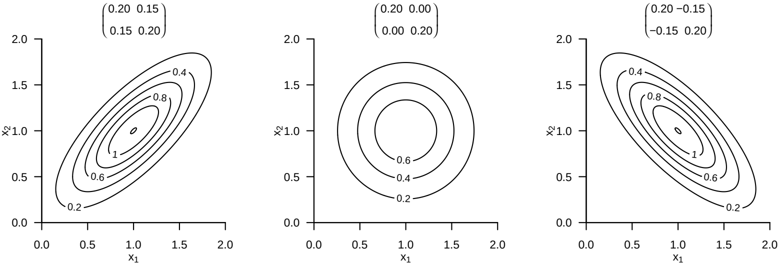
$$N(x; \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right). \tag{5}$$

Bemerkungen

- Der Parameter $\mu \in \mathbb{R}^n$ entspricht dem Wert höchster Wahrscheinlichkeitsdichte.
- Die Diagonalelemente von Σ spezifizieren die Breite der WDF bezüglich $x_1, ..., x_n$.
- Das (i, j)te Element von Σ spezifiziert die Kovarianz on x_i und x_j.
- Der Term $1/\sqrt{(2\pi)^n|\Sigma|}$ ist die Normalisierungskonstante für den Exponentialfunktionsterm.

Beispiele für multivariate Normalverteilungen

$\mu = (1,1)^{\mathrm{T}}, \Sigma \in \mathbb{R}^{2\times 2}$ wie im Abbildungstitel vermerkt.



Definition (Modell der Linearen Diskriminanzanalyse)

Gegeben seien ein m-dimensionaler Zufallsvektor $x \in \mathbb{R}^m$ und eine Zufallsvariable $y \in \{0,1\}$. Dann ist das *Modell der Linearen Diskriminanzanalyse* (LDA) die gemeinsame Verteilung

$$\mathbb{P}(x,y) = \mathbb{P}(x|y)\,\mathbb{P}(y), \tag{6}$$

wobei die diskrete marginale Verteilung von y durch die WMF

$$p(y) = \text{Bern}(y;\mu) \tag{7}$$

mit $\mu \in [0,1]$ und die kontinuierliche bedingte Verteilung von x gegeben y durch die WDF

$$p(x|y) = \begin{cases} N(x;\mu_0,\Sigma)\ , & \text{wenn } y = 0 \\ N(x;\mu_1,\Sigma)\ , & \text{wenn } y = 1\ . \end{cases} \tag{8}$$

mit $\mu_0, \mu_1 \in \mathbb{R}^m$ und $\Sigma \in \mathbb{R}^{m \times m}$ p.d. definiert ist.

Bemerkungen

- Zum Zwecke späterer Beweisführung bietet es sich an, die bedingte WDF $p(x|y)$ unter Ausnutzung der diskreten Natur von y wie folgt zu schreiben:

$$p(x|y) = N(x;\mu_0,\Sigma)^{1-y}\,N(x;\mu_1,\Sigma)^y. \tag{9}$$

- Das Modell der Linearen Diskriminanzanalyse kann zur Grundlage eines *Gauß-Klassifikators* gemacht werden.

Modellformulierung

Bemerkungen

Aus generativer Sicht wird ein Trainingsdatensatz

$$\{(x_1, y_1), ..., (x_n, y_n)\} \quad \text{mit} \quad x_i \in \mathbb{R}^m \quad \text{und} \quad y_i \in \{0, 1\} \tag{10}$$

eines Modells zur Linearen Diskriminanzanalyse wie folgt erzeugt:

(1) y_i wird zunächst durch Ziehen aus einer Bernoulli-Verteilung mit Parameter μ erzeugt.

(2) In Abhängigkeit vom Wert von y_i wird x_i dann durch Ziehen aus einer multivariaten Normalverteilung mit Kovarianzmatrixparameter Σ und Erwartungswertparameter μ_0 für $y_i = 0$ oder μ_1 für $y_i = 1$ erzeugt.

(3) Dies wird für alle $i = 1, ..., n$ wiederholt.

Datengenerierung

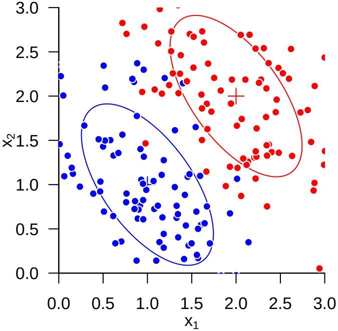
```r
# Modellformulierung
library(mvtnorm)                                    # multivariate Normalverteilung
set.seed(0)                                         # reproduzierbare Ergebnisse
m     = 2                                           # Anzahl Features
n     = 2e2                                         # Anzahl Trainingsdatenpunkte
mu    = 0.5                                         # wahrer, aber unbekannter Bernoulliparameter \mu
mu_0  = c(1,1)                                      # wahrer, aber unbekannter Erwartungswertparameter \mu_0
mu_1  = c(2,2)                                      # wahrer, aber unbekannter Erwartungswertparameter \mu_1
Sigma = matrix(c( 0.40,-0.30,                       # wahrer, aber unbekannter Kovarianzmatrixparameter
                 -0.30, 0.60),
               byrow = TRUE,
               nrow  = m)

# Datengenerierung
y     = matrix(rep(NaN,n)  , nrow = 1)              # Labeldatenarray
x     = matrix(rep(NaN,n*m), nrow = m)              # Featurevektorarray
for (i in 1:n) {                                    # Iteration über Datenpunkte
    y[i]  = rbinom(1,1,mu)                          # y_i \sim Bern(\mu)
    x[,i] = ((rmvnorm(1, mu_0, Sigma)**(1-y[i]))    # x_i \sim N(\mu_0,\Sigma)^{1-y} N(\mu_1,\Sigma)^y
            *(rmvnorm(1, mu_1, Sigma)**(y[i])))
}

# Datensatzerzeugung
D = rbind(x,y)
```

Modellformulierung

Datengenerierung

$$m = 2, \quad n = 200, \quad \mu = 0.5, \quad \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.40 & -0.30 \\ -0.30 & 0.60 \end{pmatrix}$$

$+\ \mu_0$ $\bullet\ x_i$ mit $y_i = 0$, $\quad +\ \mu_1$ $\bullet\ x_i$ mit $y_i = 1$, $\quad i = 1, ..., n$



Definition (Modell der LDA mit unabhängigen Features)

Gegeben seien ein m-dimensionaler Zufallsvektor $x \in \mathbb{R}^m$ und eine Zufallsvariable $y \in \{0, 1\}$. Dann ist das *Modell der LDA mit unabhängigen Features* die gemeinsame Verteilung

$$\mathbb{P}(x, y) = \mathbb{P}(x|y)\,\mathbb{P}(y), \tag{11}$$

wobei die diskrete marginale Verteilung von y durch die WMF

$$p(y) = \text{Bern}(y; \mu) \tag{12}$$

mit $\mu \in [0, 1]$ und die kontinuierliche bedingte Verteilung von x gegeben y durch die WDF

$$p(x|y) = \begin{cases} N(x; \mu_0, \Sigma)\ , & \text{wenn } y = 0 \\ N(x; \mu_1, \Sigma)\ , & \text{wenn } y = 1\ . \end{cases} \tag{13}$$

mit $\mu_0, \mu_1 \in \mathbb{R}^m$ und $\Sigma = \text{diag}(\sigma_1^2, ..., \sigma_m^2) \in \mathbb{R}^{m \times m}$ definiert ist.

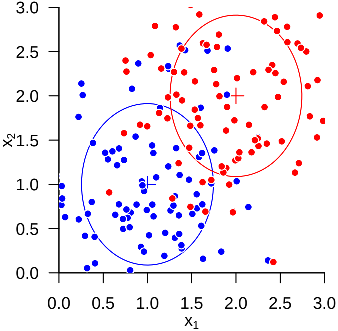
Bemerkungen

- Ein diagonaler Kovarianzmatrixparameter $\Sigma = \text{diag}(\sigma_1^2, ..., \sigma_m^2)$ induziert eine multivariate Normalverteilung, die dadurch charakterisiert ist, dass die Einträge des normalverteilten Zufallsvektors unabhängige Zufallsvariablen sind (siehe Einheit (4) in *Allgemeines Lineares Modell*).
- Das Modell kann auch als *Modell der LDA mit diagonaler Kovarianzmatrix* bezeichnet werden.
- Die Zahlen $\sigma_1^2, ..., \sigma_m^2$ werden auch *Varianzen der Features* oder *Featurevarianzen* genannt.
- Das Modell der LDA mit unabhängigen Features bildet die Grundlage eines *"naiven" Gauß-Klassifikators*.

Datengenerierung

$$m = 2, \quad n = 200, \quad \mu = 0.5, \quad \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.40 & 0 \\ 0 & 0.60 \end{pmatrix}$$

$+\ \mu_0$ $\bullet$ x_i mit $y_i = 0$, $\quad +\ \mu_1$ $\bullet$ x_i mit $y_i = 1$, $\quad i = 1, ..., n$



Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Theorem (Theorem von Bayes)

$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $A_1, ..., A_k$ sei eine Partition von Ω mit $\mathbb{P}(A_i) > 0$ für alle $i = 1, ..., k$. Wenn $\mathbb{P}(B) > 0$ gilt, dann gilt für jedes $i = 1, ..., k$, dass

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_{i=1}^k \mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}. \quad (14)$$

Bemerkungen

- Das Theorem gehört zu den Grundlagen elementarer Wahrscheinlicheitstheorie (siehe Einheit (2) in *Wahrscheinlichkeitstheorie und Frequentistische Inferenz*).
- Im Rahmen der Bayesianischen Inferenz wird $\mathbb{P}(A_i)$ oft *Prior-Wahrscheinlichkeit* und $\mathbb{P}(A_i|B)$ oft *Posterior-Wahrscheinlichkeit* des Ereignisses A_i genannt.
- Für uns ist das Theorem dahingehend von Bedeutung, dass es uns erlaubt, die Posterior-Wahrscheinlichkeit der Klassenvariable, gegeben den Featurevektor, zu berechnen:

$$p(y=1|x) = \frac{p(x|y=1)\,p(y=1)}{p(x|y=0)\,p(y=0) + p(x|y=1)\,p(y=1)}. \quad (15)$$

Theorem (Posterior-Wahrscheinlichkeit bei Linearer Diskriminanzanalyse)

Gegeben sei das Modell der Linearen Diskriminanzanalyse mit dem Featurevektor $x \in \mathbb{R}^m$ und der Klassenvariable $y \in \{0, 1\}$ sowie der Prior-Wahrscheinlichkeit $p(y = 1) = \mu$. Dann ergibt sich die Posterior-Wahrscheinlichkeit als

$$p(y = 1|x) = \frac{1}{1 + \exp(-\tilde{x}^{\mathrm{T}}\beta)} \quad \text{und} \quad p(y = 0|x) = 1 - p(y = 1|x), \tag{16}$$

wobei

$$\tilde{x} := \begin{pmatrix} 1 \\ x \end{pmatrix} \in \mathbb{R}^{m+1} \tag{17}$$

der *erweiterten Featurevektor* und

$$\beta := \begin{pmatrix} \frac{1}{2}\left(\mu_0^{\mathrm{T}}\Sigma^{-1}\mu_0 - \mu_1^{\mathrm{T}}\Sigma^{-1}\mu_1\right) + \ln\left(\frac{\mu}{1-\mu}\right) \\ -\Sigma^{-1}(\mu_0 - \mu_1) \end{pmatrix} \in \mathbb{R}^{m+1}. \tag{18}$$

der *Inferenzparametervektor* sind.

Bemerkungen

- $p(y = 1|x)$ bzw. $p(y = 0|x)$ kann zur Prädiktion der Klasse eines $x \in \mathbb{R}^m$ genutzt werden.
- Diese Prädiktion hängt von den Parametern μ, μ_0, μ_1, Σ des Modells der Linearen Diskriminanzanalyse ab.

Inferenz und Klassifikation

Beweis

Wir halten zunächst fest, dass

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)\,p(y=1)}{p(x|y=0)\,p(y=0)+p(x|y=1)\,p(y=1)} \\
&= \frac{p(x,y=1)}{p(x,y=0)+p(x,y=1)} \\
&= \frac{\frac{p(x,y=1)}{p(x,y=1)}}{\frac{p(x,y=0)}{p(x,y=1)}+\frac{p(x,y=1)}{p(x,y=1)}} \\
&= \frac{1}{1+\frac{p(x,y=0)}{p(x,y=1)}} \\
&= \frac{1}{1+\exp\left(\ln\left(\frac{p(x,y=0)}{p(x,y=1)}\right)\right)} \\
&= \frac{1}{1+\exp\left(-\ln\left(\frac{p(x,y=1)}{p(x,y=0)}\right)\right)}
\end{aligned}
\tag{19}
$$

Mit der Definition des Modells der Linearen Diskriminanzanalyse gilt dann

$$
\begin{aligned}
p(x,y=1) &= p(x|y=1)\,p(y=1) = N(x;\mu_1,\Sigma)\,\mu \\
p(x,y=0) &= p(x|y=0)\,p(y=0) = N(x;\mu_0,\Sigma)\,(1-\mu).
\end{aligned}
\tag{20}
$$

Inferenz und Klassifikation

Beweis (fortgeführt)

Wir erhalten also

$$
\begin{aligned}
\ln\left(\frac{p(x,y=1)}{p(x,y=0)}\right) &= \ln\left(\frac{N(x;\mu_1,\Sigma)\,\mu}{N(x;\mu_0,\Sigma)\,(1-\mu)}\right) \\
&= \ln\left(N(x;\mu_1,\Sigma)\mu\right) - \ln\left(N(x;\mu_0,\Sigma)(1-\mu)\right) \\
&= \ln\mu + \ln N(x;\mu_1,\Sigma) - \ln(1-\mu) - \ln N(x;\mu_0,\Sigma) \\
&= \ln\mu - \ln(1-\mu) - \frac{m}{2}\ln 2\pi - \ln|\Sigma| - \frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) \\
&\qquad\qquad + \frac{m}{2}\ln 2\pi + \ln|\Sigma| + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0) \\
&= -\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0) + \ln\mu - \ln(1-\mu) \\
&= -\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x + x^{\mathrm{T}}\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^{\mathrm{T}}\Sigma^{-1}\mu_1 + \frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x - x^{\mathrm{T}}\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_0^{\mathrm{T}}\Sigma^{-1}\mu_0 + \ln\left(\frac{\mu}{1-\mu}\right) \\
&= \frac{1}{2}\mu_0^{\mathrm{T}}\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^{\mathrm{T}}\Sigma^{-1}\mu_1 - x^{\mathrm{T}}\Sigma^{-1}(\mu_0-\mu_1) + \ln\left(\frac{\mu}{1-\mu}\right) \\
&= \begin{pmatrix} 1 & x^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \frac{1}{2}\left(\mu_0^{\mathrm{T}}\Sigma^{-1}\mu_0 - \mu_1^{\mathrm{T}}\Sigma^{-1}\mu_1\right) + \ln\left(\frac{\mu}{1-\mu}\right) \\ -\Sigma^{-1}(\mu_0-\mu_1) \end{pmatrix} \\
&=: \tilde{x}^{\mathrm{T}}\beta
\end{aligned}
$$

$\square$

Definition (Bayes-Klassifikator)

Gegeben seien die marginale Verteilung $\mathbb{P}(y)$ einer diskreten Zufallsvariable $y \in C$ und die bedingte Verteilung $\mathbb{P}(x|y)$ eines kontinuierlichen Zufallsvektors $x \in \mathbb{R}^m$. Dann ist ein Bayes-Klassifikator definiert als Funktion

$$c_{\theta,\pi} : \mathbb{R}^m \to C,\ x \mapsto c_{\theta,\pi}(x) := \arg\min_{j \in C} p(y = j|x), \tag{21}$$

wobei θ die Menge der Parameter der bedingten Verteilung $\mathbb{P}(x|y)$ ist und π die Prior-Wahrscheinlichkeiten von y angibt:

$$p(y = j) = \pi_j \quad \text{für alle} \quad j \in C. \tag{22}$$

Bemerkungen

- C ist eine Menge von Klassen, ihre Kardinalität wird als Anzahl der Klassen bezeichnet.
- Ein Bayes-Klassifikator ordnet einem Featurevektor $x \in \mathbb{R}^m$ diejenige Klasse $y \in C$ zu, die die Posterior-Wahrscheinlichkeit $p(y|x)$ maximiert. Man sagt auch, der Bayes-Klassifikator sei ein *maximum-a-posteriori*-Schätzer (MAP).

Definition (Bayes-Klassifikator für die Lineare Diskriminanzanalyse)

Gegeben sei das Modell der Linearen Diskriminanzanalyse mit dem Featurevektor $x \in \mathbb{R}^m$ und der Klassenvariable $y \in \{0,1\}$ sowie der Prior-Wahrscheinlichkeit $p(y=1) = \mu$. Dann ist ein Bayes-Klassifikator definiert als Funktion

$$c_{\theta,\mu} : \mathbb{R}^m \to \{0,1\},\ x \mapsto c_{\theta,\mu}(x) := \begin{cases} 0 & \text{für } p(y=0|x) \geq p(y=1|x) \\ 1 & \text{für } p(y=0|x) < p(y=1|x), \end{cases} \tag{23}$$

wobei $\theta = \{\mu_0, \mu_1, \Sigma\}$ die Parameter der multivariaten Normalverteilungen von x gegeben y sind.

Bemerkungen

- Für zwei Klassen vereinfacht sich die Klassifikationsregel der Linearen Diskriminanzanalyse zu

$$c_{\theta,\mu}(x) = 1 \quad \Leftrightarrow \quad p(y=1|x) > p(y=0|x) \quad \Leftrightarrow \quad p(y=1|x) > 0.5. \tag{24}$$

- Der LDA-Bayes-Klassifikator wird auch *Gauß-Klassifikator* genannt.
- Im Modell der LDA mit unabhängigen Features, d.h. mit diagonalem Kovarianzmatrixparameter $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$, wird er als *"naiver" Gauß-Klassifikator* bezeichnet.

Inferenz und Klassifikation bei bekannten Modellparametern

$$\mu = 0.5, \quad \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.40 & -0.30 \\ -0.30 & 0.60 \end{pmatrix}$$

```
# Inferenz und Klassifikation für die ersten k Datenpunkte
k       = 13                                                # Anzahl Datenpunkte
x_tilde = rbind(rep(1,k), x[,1:k])                          # erweiterte Featurevektoren
beta    = matrix(c((1/2)*(t(mu_0) %*% solve(Sigma) %*% mu_0     # Inferenzparametervektor
                       -t(mu_1) %*% solve(Sigma) %*% mu_1)
                  +log(mu/(1-mu)),
                  -solve(Sigma) %*% (mu_0-mu_1)), nrow = 3)
p_y_x   = 1/(1+exp(-t(x_tilde) %*% beta))                   # p(y = 1|x)
c_x     = as.numeric(p_y_x > 0.5 )                          # Klassifikationsregel
```

	1	2	3	4	5	6	7	8	9	10	11	12	13
x_1	1.15	1.83	2.37	0.92	2.37	2.05	0.24	0.89	2.48	1.43	0.50	3.62	1.40
x_2	3.37	2.06	1.45	0.76	2.37	2.69	0.98	0.62	2.32	0.54	1.43	0.86	2.15
$p(y = 1 \vert x)$	1.00	0.99	0.99	0.00	1.00	1.00	0.00	0.00	1.00	0.01	0.00	1.00	0.92
c(x)	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	1.00
y	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00

Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Theorem (ML-Schätzer der Linearen Diskriminanzanalyse)

Gegeben seien das Modell der Linearen Diskriminanzanalyse mit den Parametern $\{\mu, \mu_0, \mu_1, \Sigma\}$ sowie ein LDA-Trainingsdatensatz $\{(x_1, y_1), ..., (x_n, y_n)\}$ sei ein LDA Trainingsdatensatz. Dann sind die Maximum-Likelihood-Schätzer für μ, μ_0, μ_1 und Σ gegeben durch

$$
\begin{aligned}
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\hat{\mu}_0 &= \frac{1}{\sum_{i=1}^{n}(1 - y_i)} \sum_{i=1}^{n} (1 - y_i)\, x_i \\
\hat{\mu}_1 &= \frac{1}{\sum_{i=1}^{n} y_i} \sum_{i=1}^{n} y_i\, x_i \\
\hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \hat{\mu}_{y_i}\right)\left(x_i - \hat{\mu}_{y_i}\right)^{\mathrm{T}} \ .
\end{aligned}
\tag{25}
$$

Bemerkungen

- Wir machen uns hier zunutze, dass $(1 - y_i) = 1$ wenn $y_i = 0$ und $(1 - y_i) = 0$ wenn $y_i = 1$.
- μ wird als die relative Häufigkeit der $y_i = 1$ im Trainingsdatensatz geschätzt.
- μ_0 und μ_1 werden als Stichprobenmittel aller x_i mit $y_i = 0$ bzw. $y_i = 1$ geschätzt.
- Σ wird durch die empirische Kovarianzmatrix aller x_i für $i = 1, ..., n$ geschätzt.
- Substitution ergibt den Schätzer $\hat{\beta}$.

Beweis

(1) Formulierung der Log-Likelihood-Funktion

$$
\begin{aligned}
\ell(\mu, \mu_0, \mu_1, \Sigma) &:= \ln \prod_{i=1}^n p\left(x_i, y_i\right) \\
&= \ln \prod_{i=1}^n p\left(x_i | y_i\right) \, p\left(y_i\right) \\
&= \sum_{i=1}^n \ln p\left(x_i | y_i\right) + \ln p\left(y_i\right) \\
&= \sum_{i=1}^n \ln \left(N(x_i; \mu_0, \Sigma)\right)^{1-y_i} \left(N(x_i; \mu_1, \Sigma)\right)^{y_i} + \ln \left(\mu^{y_i} (1-\mu)^{1-\mu^{y_i}}\right) \\
&= \sum_{i=1}^n (1-y_i) \ln N(x_i; \mu_0, \Sigma) + y_i \ln N(x_i; \mu_1, \Sigma) + y_i \ln \mu + (1-y_i) \ln(1-\mu) \\
&= \quad \sum_{i=1}^n (1-y_i) \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x_i - \mu_0\right)^{\mathrm{T}} \Sigma^{-1} \left(x_i - \mu_0\right)\right) \\
&\quad + \sum_{i=1}^n y_i \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x_i - \mu_1\right)^{\mathrm{T}} \Sigma^{-1} \left(x_i - \mu_1\right)\right) \\
&\quad + \sum_{i=1}^n y_i \ln \mu + \sum_{i=1}^n (1-y_i) \ln(1-\mu).
\end{aligned}
$$

Beweis (fortgeführt)

(2) Gradient der Log-Likelihood-Funktion

Der Gradient der Log-Likelihood-Funktion des Modells der Linearen Diskriminanzanalyse besteht aus den partiellen Ableitungen von ℓ hinsichtlich von μ, μ_0, μ_1 und Σ. Wie unten gezeigt wird, ergibt er sich als

$$\nabla \ell(\mu, \mu_0, \mu_1, \Sigma) = \begin{pmatrix} \frac{\partial}{\partial \mu} \ell(\mu, \mu_0, \mu_1, \Sigma) \\ \frac{\partial}{\partial \mu_0} \ell(\mu, \mu_0, \mu_1, \Sigma) \\ \frac{\partial}{\partial \mu_1} \ell(\mu, \mu_0, \mu_1, \Sigma) \\ \frac{\partial}{\partial \Sigma} \ell(\mu, \mu_0, \mu_1, \Sigma) \end{pmatrix} = \begin{pmatrix} \frac{1}{\mu} \sum_{i=1}^n y_i - \frac{1}{1-\mu} \sum_{i=1}^n (1 - y_i) \\ -\frac{1}{2} \sum_{i=1}^n (1 - y_i) \left((x_i - \mu_0)^{\mathrm{T}} \Sigma^{-1} \right) \\ -\frac{1}{2} \sum_{i=1}^n y_i \left((x_i - \mu_1)^{\mathrm{T}} \Sigma^{-1} \right) \\ \frac{n}{2} \Sigma - \frac{1}{2} \sum_{i=1}^n \left(x_i - \mu_{y_i} \right) \left(x_i - \mu_{y_i} \right)^{\mathrm{T}} \end{pmatrix}.$$

Beweis (fortgeführt)

Für die partielle Ableitung hinsichtlich μ_0 und ähnlich für μ_1 ergibt sich

$$
\begin{aligned}
\frac{\partial}{\partial \mu_0} \ell(\mu, \mu_0, \mu_1, \Sigma) &= \frac{\partial}{\partial \mu_0} \sum_{i=1}^{n} (1 - y_i) \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x_i - \mu_0)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu_0) \right) \\
&= \sum_{i=1}^{n} (1 - y_i) \frac{\partial}{\partial \mu_0} \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x_i - \mu_0)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu_0) \right) \\
&= \sum_{i=1}^{n} (1 - y_i) \left(-\frac{1}{2} \frac{\partial}{\partial \mu_0} (x_i - \mu_0)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu_0) \right) \\
&= \sum_{i=1}^{n} (1 - y_i) \left(-\frac{1}{2} (x_i - \mu_0)^{\mathrm{T}} \Sigma^{-1} \right) \\
&= -\frac{1}{2} \sum_{i=1}^{n} (1 - y_i) \left((x_i - \mu_0)^{\mathrm{T}} \Sigma^{-1} \right) \\
&= -\frac{1}{2} \sum_{i=1}^{n} (1 - y_i) \left((x_i - \mu_0)^{\mathrm{T}} \Sigma^{-1} \right).
\end{aligned}
$$

Beweis (fortgeführt)

Für die partielle Ableitung hinsichtlich Σ ergibt sich

$$
\begin{aligned}
\frac{\partial}{\partial \Sigma}\ell(\mu, \mu_1, \mu_0, \Sigma) = \quad & \frac{\partial}{\partial \Sigma}\sum_{i=1}^{n}(1-y_i)\left(-\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x_i-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu_0)\right) \\
& + \frac{\partial}{\partial \Sigma}\sum_{i=1}^{n} y_i\left(-\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x_i-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu_1)\right) \\
= \quad & \sum_{i=1}^{n}(1-y_i)\left(-\frac{1}{2}\frac{\partial}{\partial \Sigma}\ln|\Sigma| - \frac{1}{2}\frac{\partial}{\partial \Sigma}(x_i-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu_0)\right) \\
& + \sum_{i=1}^{n} y_i\left(-\frac{1}{2}\frac{\partial}{\partial \Sigma}\ln|\Sigma| - \frac{1}{2}\frac{\partial}{\partial \Sigma}(x_i-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu_1)\right) \\
= \quad & \sum_{i=1}^{n}(1-y_i)\left(-\frac{1}{2}\Sigma - \frac{1}{2}(x_i-\mu_0)(x_i-\mu_0)^{\mathrm{T}}\right) \\
& + \sum_{i=1}^{n} y_i\left(-\frac{1}{2}\Sigma - \frac{1}{2}(x_i-\mu_1)(x_i-\mu_1)^{\mathrm{T}}\right) \\
= \; & \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}\left(x_i-\mu_{y_i}\right)\left(x_i-\mu_{y_i}\right)^{\mathrm{T}}.
\end{aligned} \tag{26}
$$

Beweis (fortgeführt)

Für die partielle Ableitung hinsichtlich μ ergibt sich

$$
\begin{aligned}
\frac{\partial}{\partial \mu}\ell(\mu,\mu_1,\mu_0,\Sigma) &= \frac{\partial}{\partial \mu}\left(\sum_{i=1}^n y_i \ln \mu + \sum_{i=1}^n (1-y_i)\ln(1-\mu)\right) \\
&= \sum_{i=1}^n y_i \frac{\partial}{\partial \mu} \ln \mu + \sum_{i=1}^n (1-y_i)\frac{\partial}{\partial \mu}\ln(1-\mu) \\
&= \frac{1}{\mu}\sum_{i=1}^n y_i - \frac{1}{1-\mu}\sum_{i=1}^n (1-y_i).
\end{aligned}
$$

(4) Auflösen der Maximum-Likelihood-Gleichungen

Nullsetzen der partiellen Ableitungen des Gradienten der Log-Likelihood-Funktion und Auflösen der resultierenden Log-Likelihood-Gleichungen ergibt dann die Maximum-Likelihood-Schätzer des Modells der Linearen Diskriminanzanalyse.

Beweis (fortgeführt)

Nullsetzen der ersten Gradientenkomponente ergibt

$$
\begin{aligned}
\frac{1}{\hat{\mu}}\sum_{i=1}^n y_i - \frac{1}{1-\hat{\mu}}\sum_{i=1}^n (1-y_i) = 0 \\
\Leftrightarrow \quad \frac{1-\hat{\mu}}{\hat{\mu}}\sum_{i=1}^n y_i - \sum_{i=1}^n (1-y_i) = 0 \\
\Leftrightarrow \quad \frac{1-\hat{\mu}}{\hat{\mu}}\sum_{i=1}^n y_i - n + \sum_{i=1}^n y_i = 0 \\
\Leftrightarrow \quad (1-\hat{\mu})\sum_{i=1}^n y_i - \hat{\mu} n + \hat{\mu}\sum_{i=1}^n y_i = 0 \\
\Leftrightarrow \quad (1-\hat{\mu}+\hat{\mu})\sum_{i=1}^n y_i = \hat{\mu} n \\
\Leftrightarrow \quad \hat{\mu} n = \sum_{i=1}^n y_i \\
\Leftrightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^n y_i.
\end{aligned}
$$

Beweis (fortgeführt)

Nullsetzen der zweiten Gradientenkomponente ergibt

$$
\begin{aligned}
& \sum_{i=1}^{n}(1-y_i)\left((x_i-\hat{\mu}_0)^{\mathrm{T}}\,\Sigma^{-1}\right) = 0 \\
\Leftrightarrow\quad & \sum_{i=1}^{n}(1-y_i)\,(x_i-\hat{\mu}_0)^{\mathrm{T}} = 0 \\
\Leftrightarrow\quad & \sum_{i=1}^{n}(1-y_i)x_i - \sum_{i=1}^{n}(1-y_i)\hat{\mu}_0 = 0 \\
\Leftrightarrow\quad & \sum_{i=1}^{n}(1-y_i)\hat{\mu}_0 = \sum_{i=1}^{n}(1-y_i)x_i \\
\Leftrightarrow\quad & \hat{\mu}_0 = \frac{1}{\sum_{i=1}^{n}(1-y_i)}\sum_{i=1}^{n}(1-y_i)\,x_i.
\end{aligned}
$$

Nullsetzen der dritten Gradientenkomponente ergibt in ähnlicher Weise

$$
\hat{\mu}_1 = \frac{1}{\sum_{i=1}^{n} y_i}\sum_{i=1}^{n} y_i\,x_i. \tag{27}
$$

Beweis (fortgeführt)

Nullsetzen der vierten Gradientenkomponente ergibt dann schließlich

$$
\begin{aligned}
0 &= \frac{n}{2}\hat{\Sigma} - \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu_{y_i}\right)\left(x_i - \mu_{y_i}\right)^{\mathrm{T}} \\
\Leftrightarrow \quad n\hat{\Sigma} &= \sum_{i=1}^{n}\left(x_i - \mu_{y_i}\right)\left(x_i - \mu_{y_i}\right)^{\mathrm{T}} \\
\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_{y_i}\right)\left(x_i - \mu_{y_i}\right)^{\mathrm{T}}.
\end{aligned}
$$

□

Anwendung

$$m = 2, \quad n = 200, \quad \mu = 0.5, \quad \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.40 & -0.30 \\ -0.30 & 0.60 \end{pmatrix}$$

```
# Parameterlernen
n             = ncol(x)                                                          # n
m             = nrow(x)                                                          # m
mu_hat        = mean(y)                                                          # \hat{\mu}
mu_0_hat      = rowMeans(x[,y == 0])                                             # \hat{\mu}_0
mu_1_hat      = rowMeans(x[,y == 1])                                             # \hat{\mu}_1
Sigma_hat     = matrix(rep(0,m^2), nrow = m)                                     # \hat{\Sigma}
for (i in 1:n) {
    Sigma_hat = (Sigma_hat + (1/n)*
                ((y[i] == 0)*(x[,i]-mu_0_hat) %*% t((x[,i]-mu_0_hat))
                +(y[i] == 1)*(x[,i]-mu_1_hat) %*% t((x[,i]-mu_1_hat))))
}
beta_hat      = matrix(c((1/2)*( t(mu_0_hat) %*% solve(Sigma_hat) %*% mu_0_hat   # \hat{beta}
                               - t(mu_1_hat) %*% solve(Sigma_hat) %*% mu_1_hat)
                               + log(mu_hat/(1-mu_hat)),
                         -solve(Sigma_hat) %*% (mu_0_hat-mu_1_hat)),
                       nrow = m+1)
```

Definition (Bayes-Klassifikator mit Maximum-Likelihood-Schätzern)

Gegeben seien das Modell der Linearen Diskriminanzanalyse mit dem Featurevektor $x \in \mathbb{R}^m$ und der Klassenvariable $y \in \{0,1\}$ sowie die Maximum-Likelihood-Schätzer $\hat{\mu}$, $\hat{\mu}_0$, $\hat{\mu}_1$ und $\hat{\Sigma}$. Dann ist der geschätzte Bayes-Klassifikator definiert als

$$c_{\hat{\theta},\hat{\mu}}(x) := \begin{cases} 0 & \text{für } \; p(y=0|x) \leq 0.5 \\ 1 & \text{für } \; p(y=1|x) > 0.5, \end{cases} \tag{28}$$

wobei $\hat{\theta} = \{\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}\}$ die aus dem Trainingsdatensatz geschätzten Parameter der multivariaten Normalverteilungen von x gegeben y sind.

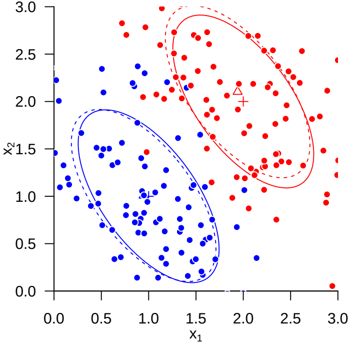
Bemerkungen

- Die ML-Schätzer der LDA-Modellparameter werden in den Bayes-Klassifikator eingesetzt.

Anwendung

$$m = 2, \quad n = 200, \quad \mu = 0.5, \quad \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.40 & -0.30 \\ -0.30 & 0.60 \end{pmatrix}$$

$+$ μ_0, $\bullet$ x_i mit $y_i = 0$, $\triangle$ $\hat{\mu}_0$, $+$ μ_1, $\bullet$ x_i mit $y_i = 1$, $\triangle$ $\hat{\mu}_1$



Definition (Bayes-Klassifikator mit Prior-Wahrscheinlichkeit)

Gegeben seien das Modell der Linearen Diskriminanzanalyse mit dem Featurevektor $x \in \mathbb{R}^m$ und der Klassenvariable $y \in \{0, 1\}$ sowie die Maximum-Likelihood-Schätzer $\hat{\mu}_0$, $\hat{\mu}_1$, $\hat{\Sigma}$ und die Prior-Wahrscheinlichkeit $p(y = 1) = \mu$. Dann ist der geschätzte Bayes-Klassifikator definiert als

$$c_{\hat{\theta},\mu}(x) := \begin{cases} 0 & \text{für } \; p(y = 0|x) \leq 0.5 \\ 1 & \text{für } \; p(y = 1|x) > 0.5, \end{cases} \tag{29}$$

wobei $\hat{\theta} = \{\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}\}$ die aus dem Trainingsdatensatz geschätzten Parameter der multivariaten Normalverteilungen von x gegeben y sind.

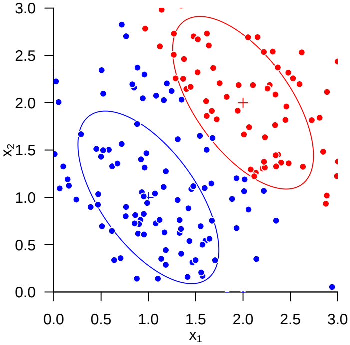
Bemerkungen

- Die Parameter der bedingten Verteilung $\mathbb{P}(x|y)$ werden als Schätzer aus einem Trainingsdatensatz bestimmt, der Parameter der marginalen Verteilung $\mathbb{P}(y)$ wird als Prior-Wahrscheinlichkeit vorgegeben.
- Der Wert von μ modelliert *a priori* bekannte Klassenhäufigkeiten: Kleinere Werte von μ führen bevorzugt zu "negativen" Klassifikationen $\hat{y}_i = 0$, größere Werte von μ führen bevorzugt zu "positiven" Klassifikationen $\hat{y}_i = 1$.

Anwendung

$$m = 2, \quad n = 200, \quad \mu = 0.5, \quad \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.40 & -0.30 \\ -0.30 & 0.60 \end{pmatrix}$$

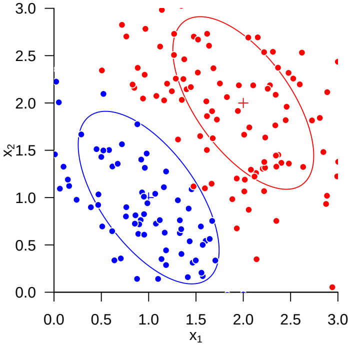
$+$ μ_0, $\bullet$ x_i mit $\hat{y}_i = 0$, $+$ μ_1, $\bullet$ x_i mit $\hat{y}_i = 1$, mit $p(y = 1) = 0.1$



Anwendung

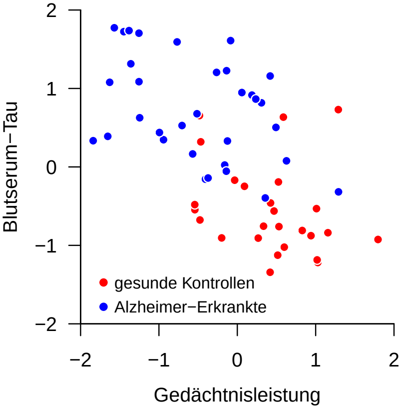
$$m = 2, \quad n = 200, \quad \mu = 0.5, \quad \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.40 & -0.30 \\ -0.30 & 0.60 \end{pmatrix}$$

$+\ \mu_0$, $\bullet\ x_i$ mit $\hat{y}_i = 0$, $\quad +\ \mu_1$, $\bullet\ x_i$ mit $\hat{y}_i = 1$, $\quad$ mit $p(y = 1) = 0.9$



Beispieldatensatz

Gedächtnisleistung und Blutserum-Tau als Indikatoren für Alzheimer-Erkrankung



Anwendungsbeispiel

Beispieldatensatz

Gedächtnisleistung und Blutserum-Tau als Indikatoren für Alzheimer-Erkrankung

MEM	TAU	ALZ
-1.26	1.70	1
-0.71	0.53	1
-0.16	0.02	1
0.34	-0.76	0
0.06	0.95	1
-0.26	1.20	1
-0.48	-0.68	0
0.27	-0.91	0
0.19	0.92	1
0.94	-0.88	0
-0.03	-0.17	0
1.29	-0.32	1
1.29	0.73	0
-0.41	-0.16	1
0.60	-1.02	0
-0.51	0.68	1
1.03	-1.22	0
0.52	-1.12	0
-1.63	1.08	1
0.31	0.82	1

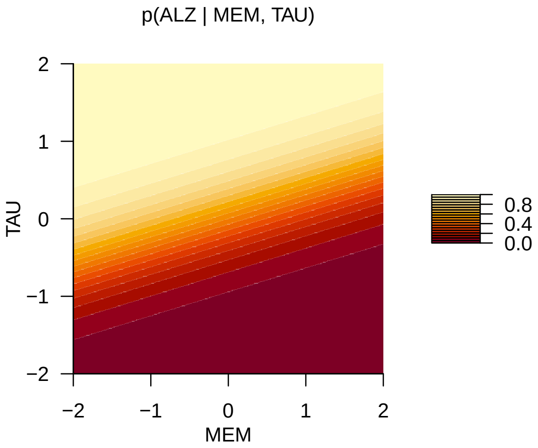
Anwendungsbeispiel

Evaluation der Wahrscheinlichkeit für Positiv-Fälle

```
# Daten extrahieren
D         = read.csv("Daten/Bayes-Klassifikation.csv", header = T)            # Datensatz
x         = t(D[,1:2])                                                        # Featurevektoren
y         = t(D[,3])                                                          # Labelvariable

# Daten analysieren
n         = ncol(x)                                                           # n
m         = nrow(x)                                                           # m
mu_hat    = mean(y)                                                           # \hat{\mu}
mu_0_hat  = rowMeans(x[,y == 0])                                              # \hat{\mu}_0
mu_1_hat  = rowMeans(x[,y == 1])                                              # \hat{\mu}_1
Sigma_hat = matrix(rep(0,m^2), nrow = m)                                      # \hat{\Sigma}
for (i in 1:n) {
    Sigma_hat = (Sigma_hat + (1/n)*                                           # \hat{\Sigma}
                 ((y[i] == 0)*(x[,i]-mu_0_hat) %*% t((x[,i]-mu_0_hat))
                 +(y[i] == 1)*(x[,i]-mu_1_hat) %*% t((x[,i]-mu_1_hat))))
}
beta_hat  = matrix(c((1/2)*( t(mu_0_hat) %*% solve(Sigma_hat) %*% mu_0_hat      # \hat{beta}
                           - t(mu_1_hat) %*% solve(Sigma_hat) %*% mu_1_hat)
                           + log(mu_hat/(1-mu_hat)),
                        -solve(Sigma_hat) %*% (mu_0_hat-mu_1_hat)), nrow = m+1)

# Ergebnisraumdefinition
x_min     = -2                                                                # MEM/TAU Minimum
x_max     = +2                                                                # MEM/TAU Maximum
x_res     = 4e2                                                               # MEM/TAU Auflösung
mem       = seq(x_min, x_max, length.out = x_res)                             # MEM
tau       = seq(x_min, x_max, length.out = x_res)                             # TAU
p_y       = matrix(rep(NaN, x_res*x_res), nrow = x_res)                       # p(y=1|(MEM,TAU))
for (i in 1:x_res) {                                                          # MEM-Iterationen
    for (j in 1:x_res) {                                                      # TAU-Iterationen
        x_tilde  = rbind(1, mem[i], tau[j])                                   # \tilde{x}
        p_y[i,j] = 1/(1+exp(-t(x_tilde) %*% beta_hat))                        # p(y=1|(MEM,TAU))
    }
}
```

Anwendungsbeispiel

Evaluation der Wahrscheinlichkeit für Positiv-Fälle



Leave-one-out cross-validation zur Bestimmung der Featureprädiktivität

```
# Daten analysieren
n       = nrow(D)                                                                        # Anzahl Datenpunkte
y_pred  = matrix(rep(NaN, n*2), nrow = n)                                                # wahre und prädizierte Label
for (i in 1:n) {                                                                         # leave-one-out cross-validation
    x_train      = t(D[-i,1:2])                                                          # Trainingsdatensatzfeatures
    y_train      = t(D[-i,3  ])                                                          # Trainingsdatensatzlabels
    x_test       = t(D[ i,1:2])                                                          # Testdatensatzfeaturevektor
    y_pred[i,1]  = t(D[ i,3])                                                            # Testdatensatzfeaturevektorlabel
    n_train      = ncol(x_train)                                                         # Anzahl Trainingsdatensatz
    m            = nrow(x_train)                                                         # Anzahl Features
    mu_hat       = mean(y_train)                                                         # \hat{\mu}
    mu_0_hat     = rowMeans(x_train[,y_train == 0])                                      # \hat{\mu}_0
    mu_1_hat     = rowMeans(x_train[,y_train == 1])                                      # \hat{\mu}_1
    Sigma_hat    = matrix(rep(0,m^2), nrow = m)                                          # \hat{\Sigma}
    for (j in 1:n_train) {
        Sigma_hat= (Sigma_hat + (1/n_train)*
                    ((y_train[j] == 0)*(x_train[,j]-mu_0_hat) %*% t((x_train[,j]-mu_0_hat))
                    +(y_train[j] == 1)*(x_train[,j]-mu_1_hat) %*% t((x_train[,j]-mu_1_hat))))
    }                                                                                    # \hat{beta}
    beta_hat     = matrix(c((1/2)*( t(mu_0_hat) %*% solve(Sigma_hat) %*% mu_0_hat
                                  - t(mu_1_hat) %*% solve(Sigma_hat) %*% mu_1_hat)
                                  + log(mu_hat/(1-mu_hat)),
                             -solve(Sigma_hat) %*% (mu_0_hat-mu_1_hat)), nrow = m+1)
    x_test_tilde = rbind(1, x_test)                                                      # \tilde{x}
    p_y          = 1/(1+exp(-t(x_test_tilde) %*% beta_hat))                              # p(y = 1|x)
    y_pred[i,2]  = as.numeric(p_y >= 0.5)                                                # Prädiktion
}
# Klassifikation evaluieren
TN      = sum(y_pred[y_pred[,1] == 0, 2] == 0)                                           # true negatives
FP      = sum(y_pred[y_pred[,1] == 0, 2] == 1)                                           # false positives
FN      = sum(y_pred[y_pred[,1] == 1, 2] == 0)                                           # false negatives
TP      = sum(y_pred[y_pred[,1] == 1, 2] == 1)                                           # true positives
TPR     = (TP)/(TP+FN)                                                                   # true positive rate (sensitivity)
TNR     = (TN)/(TN+FP)                                                                   # true negative rate (specificity)
Acc     = (TN+TP)/(TN+FP+FN+TP)                                                          # Genauigkeit (accuracy)
cat("Accuracy    : ", Acc, "\nSensitivity : ", TPR, "\nSpecificity : ", TNR)             # Ergebnisausgabe
```

```
Accuracy    :  0.8333333
Sensitivity :  0.8235294
Specificity :  0.8461538
```

Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition einer Bernoullizufallsvariable wieder.
2. Geben Sie die Definition des Modells der Linearen Diskriminanzanalyse (LDA) wieder.
3. Worin besteht der Unterschied zwischen der LDA und der LDA mit unabhängigen Features?
4. Erläutern Sie die Generierung von Daten unter dem Modell der Linearen Diskriminanzanalyse.
5. Geben Sie das Theorem von Bayes wieder.
6. Geben Sie das Theorem zur Posterior-Wahrscheinlichkeit bei Linearer Diskriminanzanalyse wieder.
7. Geben Sie die Definition eines Bayes-Klassifikators wieder.
8. Erläutern Sie den Bayes-Klassifikator für die Lineare Diskriminanzanalyse.
9. Geben Sie das Theorem zur Maximum-Likelihood-Schätzung der Linearen Diskriminanzanalyse wieder.
10. Worin besteht der Unterschied zwischen dem LDA-Bayes-Klassifikator mit ML-Schätzern und dem LDA-Bayes-Klassifikator mit Prior-Wahrscheinlichkeit?
11. Welche Rolle spielt die Prior-Wahrscheinlichkeit beim Bayes-Klassifikator für die Lineare Diskriminanzanalyse?
12. Erläutern Sie, wie mithilfe einer Linearen Diskriminanzanalyse die Wahrscheinlichkeit ermittelt werden kann, eine bestimmte psychische Erkrankung zu haben.

Zhao, Shuping, Bob Zhang, Jian Yang, Jianhang Zhou, and Yong Xu. 2024. "Linear Discriminant Analysis." *Nature Reviews Methods Primers* 4 (1): 70. https://doi.org/10.1038/s43586-024-00346-y.