uebung7.r

Janina

Sun Jun 18 19:38:31 2017

```
# Uebungsblatt 7
# Namen: Janina Schoenberger, Benjamin Weigner
# Tutorin: Gergana Stanilova
# Uebung: Mi 12-14 Uhr

# Aufgabe 17
# LDA: linear discriminant analysis
# Voraussage: Metastasenbildung ja/nein = 1/0
# Wahrscheinlichkeit fuer Auftreten von Metastasen p
p <- 0.5
# Mittelwert M0-Gruppe: 0 bzw 2
mu1 <- matrix(c(0,2), nrow=2, ncol=1)
mu1
```

```
##      [,1]
## [1,]    0
## [2,]    2
```

```
# Mittelwert M1-Gruppe: 2 bzw -2
mu2 <- matrix(c(2,2), nrow=2, ncol=1)
mu2
```

```
##      [,1]
## [1,]    2
## [2,]    2
```

```
# Kovarianzmatrix ueber beide Gruppen hat auf Hauptdiagonalen Eintraege 2, 0.5, auf Nebendiagonalen 0
epsilon = matrix( c(2,0,0,0.5), nrow=2, ncol=2, byrow=TRUE)
epsilon
```

```
##      [,1] [,2]
## [1,]    2  0.0
## [2,]    0  0.5
```

```
# A
# Was kann man aus der Kovarianzmatrix ueber die Korrelation zwischen den beiden Genen folgern?
# -> Auf den Nebendiagonalen der Kovarianzmatrix stehen Cov(M0,M1) und Cov(M1,M0)
# -> M0 und M1 haben keinen monotonen Zusammenhang, da die Kovarianz 0 ist d.h. sie korrelieren nicht
# Was kann man aus der Kovarianzmatrix ueber die Form der multivarianten Normalverteilung folgern?
# -> Die Varianz von M0 betraegt 2, die von M1 0.5
# -> Das bedeutet, dass die Normalverteilungskurve von M0 breiter ist, als die von M1

# B
# Entscheidungsformel der LDA im konkreten Beispiel als Funktion a+bx1^+cx2
lambda <- solve(epsilon)%*%(mu1-mu2)
lambda
```

```
##      [,1]
## [1,]   -1
## [2,]    0
```

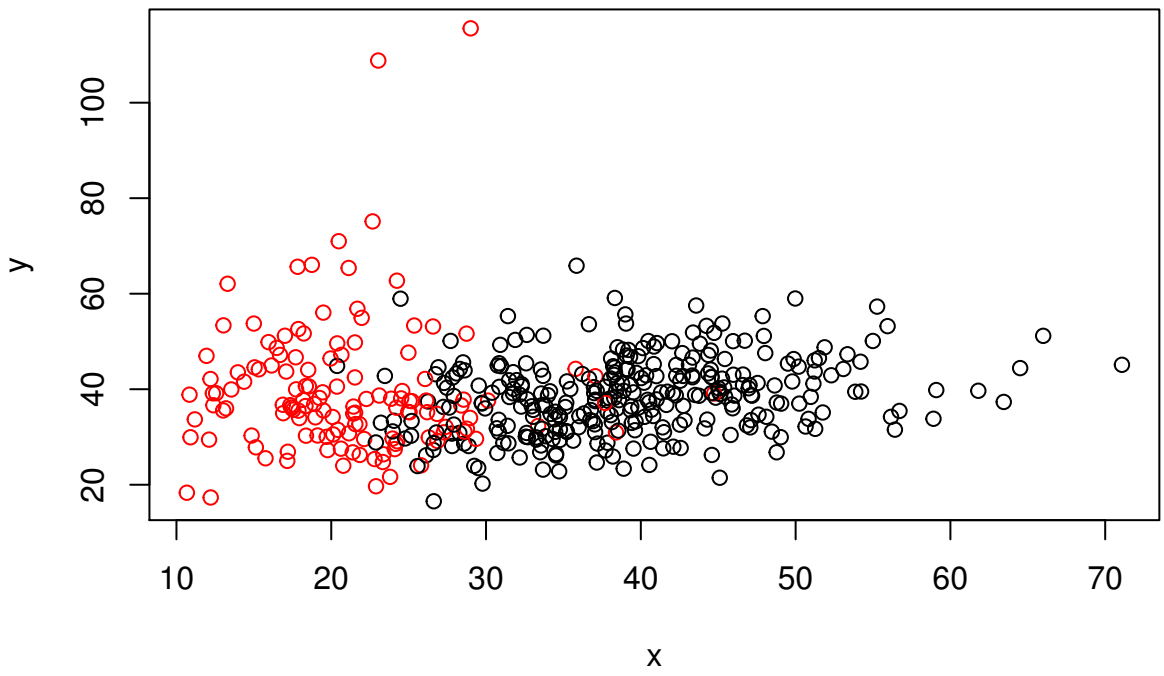
```r
# s. handschriftlich

# C
# Weitere Beobachtung p=(1,0) klassifizieren
# Klassifikationsergebnis basierend auf Entscheidungsformel:
# D(1,0) = -1
# -> mit Metastasen

# D
# Weitere Informationen benoetigt fuer Durchfuehrung einer QDA:
# Es muesste ueberprueft werden, ob die Daten (annaehernd) normalverteilt sind,
# denn das ist Voraussetzung fuer eine Diskriminanzanalyse
# (Eine gleiche Varianz ist nur bei LDA noetig)


# Aufgabe 18
# Datensatz einer Tumorprobe
# setwd(...)
Patient1 <- read.csv2("Patient1.txt", header=FALSE)
Patient1Label <- read.table("Patient1Label.txt", header=FALSE)
# Datensatz umfasst Messungen von 118 Massekanaelen und 440 raeumlichen Positionen
# Betrachtet werden Kaenaele 3 und 7
# Label-Daten = Response-Variable (0 fuer Bindegewebe, 1 fuer Tumorgewebe)

# A
# Massekanaele 3 und 7 und Labels visualisieren
x <- as.vector(Patient1[,3])
is.vector(x)
```

```
## [1] TRUE
```

```r
y <- as.vector(Patient1[,7])
is.vector(y)
```

```
## [1] TRUE
```

```r
label <- factor(Patient1Label[,1])
plot(x,y,col=label)
```



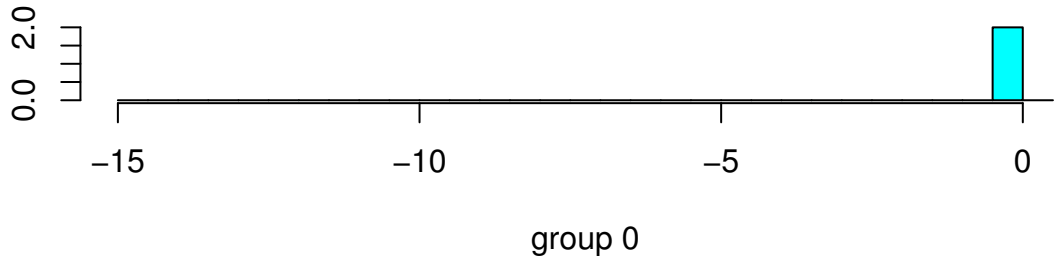

```
# Kommentar: Ist eine graphische Visualisieren leicht moeglich?
# Eine grafische Visualisierung ist moeglich, allerdings ueberschneiden sich die
# Punktewolken stark, dh es es kann nur eine sehr grobe Einschaetzung gewonnen werden

# B
# LDA Klassifikator (lda aus MASS Library)
library(MASS)
```

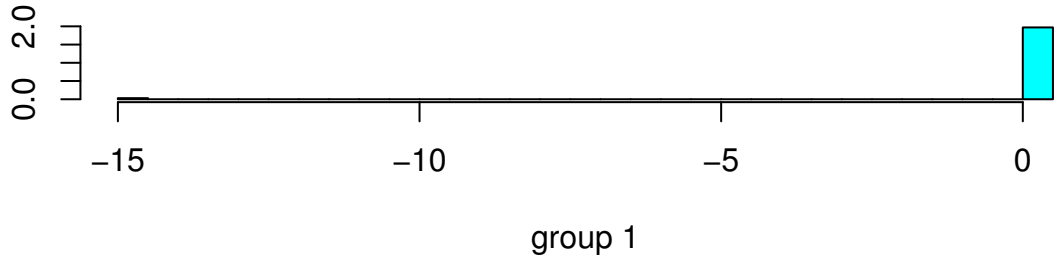
```
## Warning: package 'MASS' was built under R version 3.3.3
```

```
cl <- lda(Patient1Label[,1]~x+y)
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear
```

```
plot(cl)
```






```
# C
# Klassifikationsguete mit verschiedenen Massen
# predict benutzen
pl <- predict(cl,newdata=Patient1[,c(3,7)])$class

# confusion matrix
#t<- table(p,Patient1[,c(3,7)]) # ?
tl <- table(Patient1Label[,1], pl)
tl
```

```
##    pl
##       0   1
##   0 306   0
##   1 134   0
```

```
# accuracy
accl <-sum(diag(tl))/sum(tl)*100
accl
```

```
## [1] 69.54545
```

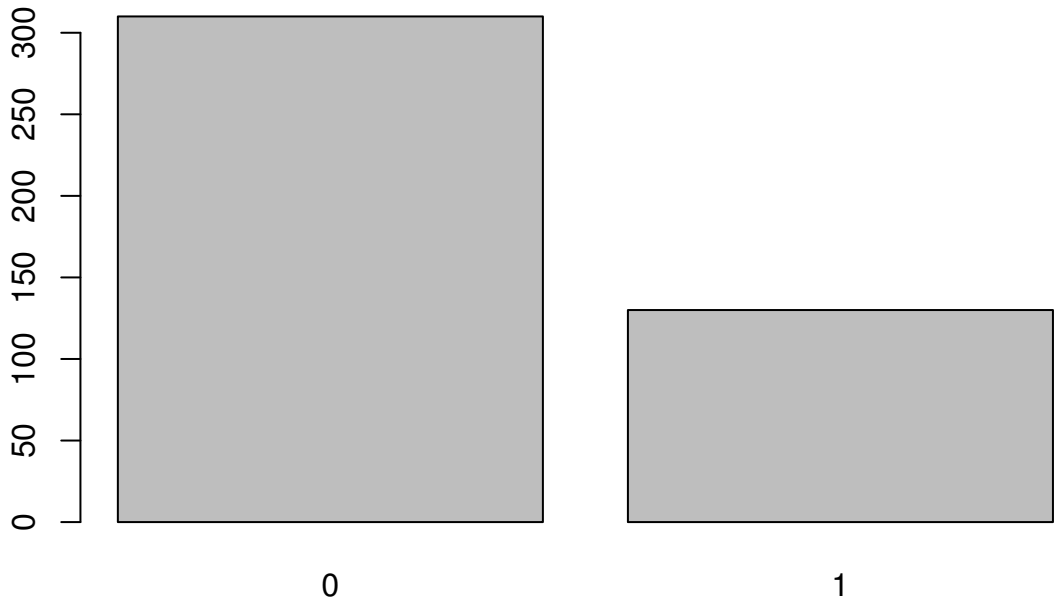
```
# missclassification error
# mean(vorhergesagte_werte != wirkliche werte)
misl <- tl[1,2]+tl[2,1]/(tl[1,1]+tl[1,2]+tl[2,1]+tl[2,2]) *100
misl
```

```
## [1] 30.45455
```

```
# D
# QDA Klassifikator (qda aus MASS Library)
x <- as.numeric(levels(Patient1[,3]))[Patient1[,3]]
y <- as.numeric(levels(Patient1[,7]))[Patient1[,7]]
cq <- qda(Patient1Label[,1]~x+y)
cq
```

```
## Call:
## qda(Patient1Label[, 1] ~ x + y)
##
## Prior probabilities of groups:
##         0         1
## 0.6954545 0.3045455
##
## Group means:
##          x        y
## 0 39.15822 38.54181
## 1 21.01077 39.59797
```

```
# E
# Klassifikationsguete mit verschiedenen Massen
pq <- predict(cq,newdata=Patient1[,c(3,7)])$class
plot(pq)
```



```r
# confusion matrix
tq <- table(Patient1Label[,1], pq)
tq
```

```
##    pq
##       0   1
##   0 290  16
##   1  20 114
```

```r
# accuracy
accq <- sum(diag(tq))/sum(tq)*100
accq
```

```
## [1] 91.81818
```

```r
# missclassification error
misq <- tq[1,2]+tq[2,1]/(tq[1,1]+tq[1,2]+tq[2,1]+tq[2,2]) *100
misq
```

```
## [1] 20.54545
```

```r
# Vergleich zu Guete von lda
# Die Accuracy von der quadratischen Diskriminanzanalyse ist deutlich hoeher
# als die Accuracy der linearen Diskrimanzanalyse. Ebenso ist der missclassification
# error niedriger.
# Daraus laesst sich schliessen, dass die QDA fuer unsere Daten besser geeignet ist
```

Entscheidungsformel:

$$\log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}(\mu_1+\mu_2)^T \Sigma^{-1}(\mu_1-\mu_2) + X^T \Sigma^{-1}(\mu_1-\mu_2)$$

$\pi_1 = \pi_2 = 0{,}5$

$\mu_1 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$

$\mu_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$

$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$

$$\Rightarrow \log(1) - \frac{1}{2}\left(\begin{pmatrix} 0 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)^T \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix} \left(\begin{pmatrix} 0 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right) + X^T \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix} \left(\begin{pmatrix} 0 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)$$

$$= 0 - \frac{1}{2}(2\ 4) \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} -2 \\ 0 \end{pmatrix} + X^T \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} -2 \\ 0 \end{pmatrix}$$

$$= 0 - (1\ 2) \begin{pmatrix} 1 \\ 0 \end{pmatrix} + X^T \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$= -1 + 1X1 + 0X2$$

$$= -1 + X_1$$