Prof. Dr. Katharina Morik,
JProf. Dr. Uwe Ligges
Dipl.-Inform. Felix Jungermann,
Dipl.-Stat. Julia Schiffner

Dortmund, 08.06.10
Abgabe: bis Di, 15.06., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2010

Blatt 8

Aufgabe 8.1 (4 Punkte)
Erzeugen Sie mit Hilfe des folgenden R-Codes zwei Datensätze:

```
library(mvtnorm)
set.seed(2)

datena <- rmvnorm(n=100, mean = c(0,-3), sigma = matrix(c(0.5,0,0,5),2,2))
datenb <- rmvnorm(n=100, mean = c(0,3), sigma = matrix(c(0.5,0,0,5),2,2))
daten1 <- rbind(datena, datenb)
y <- factor(c(rep(1,100), rep(2,100)))
daten1 <- data.frame(daten1, y = y)

datena <- rmvnorm(n=100, mean = c(0,-1), sigma = matrix(c(0.5,0.5,0.5,5),2,2))
datenb <- rmvnorm(n=100, mean = c(0,1), sigma = matrix(c(0.5,0.5,0.5,5),2,2))
daten2 <- rbind(datena, datenb)
daten2 <- data.frame(daten2, y = y)
```

a) Sagen Sie für beide Datensätze `y` mit Hilfe von Klassifikationsbäumen (Paket `rpart`) vorher. Plotten und beschriften Sie jeweils die Bäume. Visualisieren Sie außerdem jeweils die Entscheidungsgrenze. Was fällt Ihnen auf?

b) Verwenden Sie nun jeweils Zufallswälder für die Vorhersage (Paket `randomForest`). Visualisieren Sie jeweils die Entscheidungsgrenze und interpretieren Sie die Plots.

Aufgabe 8.2 (6 Punkte)
Auf der Homepage steht der Datensatz `spam.txt` sowie die Datei `info.txt`, die einige Informationen zu den Daten enthält.

Wenden Sie die folgenden Verfahren auf den Datensatz an und berechnen Sie die Wiedereinsetzungsfehlerrate, die 10-fach kreuzvalidierte Fehlerrate, den e0-bootstrap-Schätzer sowie den .632-bootstrap-Schätzer und vergleichen Sie die Ergebnisse:

a) lineare Diskriminanzanalyse (R-Funktion `lda` im Paket `MASS`)

b) Klassifikationsbaum

c) Zufallswald.