Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering
IDP Institut für Datenanalyse
und Prozessdesign

HS 2016

Statistisches Data Mining (StDM)

Woche 11

Aufgabe 1

Krabben Geschlecht

Wir betrachten den Crabs Datensatz, der im MASS package enthalten ist. Machen Sie sich mit dem Datensatz und der rpart-Funktion vertraut.

a) Erstellen Sie mit der Funktion rpart aus der rpart Library einen Klassifikationsbaum für die Variable Sex. Stellen Sie den Klassifikationsbaum grafisch dar. Quantifizieren Sie die Performance auf dem Trainingsdatensatz durch eine Kofusionsmatrix. R-Hinweise:

```
t1 <- rpart(...)
windows(14,8)  # damit Raender gross genug fuer Beschriftung
plot(t1)
text(t1,use.n=TRUE)
```

b) Im Klassifikationsbaum wurden als Splitvariablen nur die Variablen RW und CL verwendet. Erstellen Sie ein Streudiagramm mit diesen beiden Variablen und färben Sie die Punkte gemäss dem Geschlecht der Krabben ein. Kennzeichnen Sie den ersten Split als Linie im Streudiagramm. Was ist der grosse Nachteil von Klassificationsbäumen?

c) Führen Sie mit den numerischen Variablen des Crabs Datensatz eine PCA durch und visualisieren Sie die Position der Datenpunkte im Streudiagramm der ersten beiden Hauptkomponenten. Färben Sie die Punkte gemäss dem Geschlecht der Krabben ein.

d) Fügen Sie die Hauptkomponenten als weitere Variablen an den Datensatz Crabs an. Trainieren Sie mit dem erweiterten Datensatz wieder einen Klassifikationsbaum und stellen Sie ihn dar. Welche Variablen wurden als Splitvariaben verwendet? Stellen Sie die Daten im Streudiagramm der beiden häufigsten benutzen Split-Variablen dar. Quantifizieren Sie die Performance auf dem Trainingsdatensatz durch eine Kofusionsmatrix. Was fällt im Vergleich zu den Ergebnissen aus den letzten Teilaufgaben auf? Können Sie sich die Unterschiede erklären?

Aufgabe 2

Manuelle Bagging Aufgabe

In this exercise we build classification trees from Boston dataset from MASS package and bag (bootstrap aggregate) them manually.

First, load the packages:

```
library(rpart)
library(MASS) # For the data set
```

a) Make a new binary variable from crim variable, which is `TRUE` if an observation of crim is greater than the median of crim, and `FALSE` otherwise.

b) Now split the data into training set (75%) and test set (25%) randomly. Hinweis: don't forget to set the seed for reproducible results.

```
set.seed(1)
index.train = sample(... , size = ...))
data.train = Boston[..., ]
data.test= Boston[..., ]
```

c) Fit a classification tree on the training data and predict its performance on the test data. Finally, produce a cross-table of correcly and incorreclty classified samples. Hinweis:

```
fit = rpart(...)
pred = predict(...)
table(...)
```

d) Now let's procede to bootstrap aggregation (bagging). As the name suggests, you would need to aggregate the predictions on various bootstrap samples from the data. You need to grow 100 trees and loop over them by taking a different bootstrap sample of observations every time, then fitting a classification tree to it, and finally predicting its performance on the test set.

Hinweise:

```
n.trees = 100
preds = rep(0, ...)
for (j in 1:n.trees ){
  # Doing the bootstrap
  index.train = sample( ... )

  # Fitting to the training data
  fit = rpart( ... )
  # Predict the test set
  tree.fit.test = predict( ... )
  preds = preds + ... #<--- Add +1 if predictions are correct
}
```

e) Now make a cross table of the prediction results.

f) Optional: Verify your results with random forest. How close are they? Hinweise:

```
library(randomForest)
fit = randomForest(... , data=..., ntree=..., mtry = ... )
fit
table(..., ...)
```