Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering
IDP Institut für Datenanalyse
und Prozessdesign

HS 2016

Statistisches Data Mining (StDM)

Praktikum Woche 13

Aufgabe 1 Boosting

In dieser Aufgabe werden verschieden Klassifikationsmethoden auf simulierte Daten angewandt und die Performance verglichen.

a) Simulieren Sie 12'000 mal 10 Standardnormalverteilte Beobachtungen $X_1, X_2, ..., X_{10}$. Generieren Sie dann die Variable Y_i aufgrund der simulierten Daten folgendermassen:

$$Y_i = \begin{cases} 0 & \text{falls } \sum_{i=1}^{10} X_i^2 <= 9.34 \\ 1 & \text{sonst} \end{cases}$$

Verwenden Sie 2000 Beispiele zum Trainieren und 10000 zum Testen.

b) Erstellen Sie einen Klassifikationsbaum auf dem Trainings-Datensatz der nur aus dem Stumpf besteht. Schätzen Sie die Fehlerrate mithilfe des Test-Datensatzes ab. Hinweis:

```
treeSingleStump<-rpart(y ~ .,data=train,
     control=list(maxdepth=1, cp=-1,minsplit=0,xval=0),method="class")
```

c) Erstellen Sie einen Klassifikationsbaum basierend auf den internen Stopp-Kriterien, berechnen Sie die Fehlerrate.

d) Wenden Sie einen Random Forest auf die Daten an.

e) Wenden Sie Adaboost auf die Daten an. Verwenden Sie dazu das packet `gbm` und spielen Sie mit den Parametern `n.trees` und `interaction.depth` bis Sie einen Error kleiner als 0.12 bekommen. Hinweis: Sie müssen die Zielvariable wieder in einen numerischen Wert (0,1) umwandeln, verwenden Sie die distribution='adaboost'.

f) Angenommen Sie kennen den Prozess der Datenerzeugung wie in a) beschrieben bis auf die Konstante 9.34. Überlegen Sie sich eine geschickte Transformation der Variablen $X_1, X_2 \ldots X_10$ in eine neue Variable X und wenden Sie diese Transformation an. Verwenden Sie einen Klassifikationsbaum wie in c) und bestimmen Sie die Fehlerrate.