Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering
IDP Institut für Datenanalyse
und Prozessdesign

HS 2016

Statistisches Data Mining (StDM)

Praktikum Woche 12

Aufgabe 1 Random Forest

In this excercise, we try to detect spam given some features of the email. Source: A part of the exercise is from Dr. Markus Kalisch.

a) Have a look at the data set “spam“in the package “ElemStatLearn“. Fit a classification tree based on the gini and the information-criteria. Calculate the naive error rate.

b) Calculate the error rate based upon 10-fold cross-validation for gini method (default)

c) Fit a random Forest with the default settings. (Use seed 123 in order to reproduce the solution). Be patient: this may take several seconds.

d) Plot the error rate vs. the number of fitted trees. How many trees are necessary? Refit the model with the chosen number of trees. How long does it take now? Have a look at the output. What error rate do you expect for new predictions (OOB error rate)? What is the error rate in the ’spam’-class?

e) Suppose, we get a new email and want to predict the spam label. For simplicity, we refit the Random Forest on 2601 randomly chosen emails and save the remaining 2000 emails as test set. How does the OOB error compare with the error on the test set? (use ntree = 100, and set.seed = 123)

f) Suppose we don’t want to compute all variables for each new incoming mail, but only use the best 5. Which 5 variables should we choose? Compare the OOB error using all variables, the best 5 and the worst 5 (according to decrease in accuracy; use ntree = 100 and seed = 123).