Machine Learning 2
Data Science
Winter Semester 2018/19

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 3
Boosting regression trees, introduction to support vector machines.

In this workshop you will need the packages `gbm` and `e1071` which you will probably need to install (`install.packages("gbm")`). Also used are `ISLR` and `MASS`, which you have already installed.

Exercise 1 Investigating a boosted tree model

You will use the R package `gbm` to fit a boosted regression tree. In this exercise you will halt the boosting algorithm after certain number of iterations to see how the *slow learning* in the boosting algorithm develops. The data set we will use is the `mtcars` data set to predict the variable `mpg` – fuel consumption in miles per gallon. This data set is small enough to be able to investigate in detail, but no so small that it has no interesting features.

In Moodle you can find the R code for this exercise in the file `workshop3.r`. A description of the steps in the source file are given below. Work through the source file slowly referring to the written details at each stage.

(a) Load the library and set the seed

(b) One iteration

(i) Fit a boosted tree to the `mtcars` Data with just one iteration (`n.trees=1`). The shrinkage parameter $\lambda = 0.1$ (large). Note that the default `interaction.depth` is 1, so each iteration fits a tree with just one split. The other parameters will be explained in Exercise 2.

(ii) Output summary information. Not so interesting after just one iteration.

(iii) Compute the MSE of the null model ($\widehat{y}_i = \overline{y}$) and compare it to the MSE after one iteration.

(iv) Plot the fitted values against the observed values, and add two reference lines $y = \overline{y}$ and $y = x$. A perfect fit puts all the points onto the diagonal line.

(v) This command gives the effect of the variable called `cyl`. The fitted split is `cyl>5`.

(c) Fit a boosted tree with two iterations. Obtain the MSE and again plot the fitted against the observed values. Note that exactly the same split has been fitted, but the separation between the two fitted value levels has got bigger.

(d) Fit a boosted tree with three iterations.

(i) Obtain the MSE and again plot the predicted against the observed values. Now there is a third level in the fitted values, the new level contains just two observations.

(ii) `summary(mtcars.boost)` gives the "relative influence" of each variable, which is a similar measure to the variable importance of a bagged-tree/random-forest.

(iii) The split at the 3rd iteration is on the variable `displ`. The marginal effect of the split can be inspected graphically using
`plot(mtcars.boost,i.var="displ")`

(e) Fit a boosted tree with 10 iterations,

(i) First inspect the model after 4 iterations. A new split occurs, what is this new split?

(ii) After 5 iterations the `cyl>5` is reinforced.

(iii) The `matplot` command plots the fitted values against iteration number. Many of the fitted values are the same, so what we see are the different levels in the fitted values.

(iv) Inspect the model after 10 iterations.

(f) Add another 90 iterations to the model and investigate.

(g) Add another 900 iterations to the model and investigate.

(h) Use another shrinkage rate λ and compare how quickly the model fits with different values of λ.

Exercise 2 Boosting with the Boston data

Load the packages `ISLR` and `MASS`. Define the training set, and the outcome variable for the test data using:

```
library(ISLR)
library(MASS)
set.seed(1)
train = sample(1:nrow(Boston), nrow(Boston)/2)
boston.test=Boston[-train,"medv"]
```

Work through subsection 8.3.4 in James et al on page 330.

Exercise 3 Introduction to support vector machines

A package to fit SVM models in R is called `e1071`, a helpful name which originates from a internal department code at the *Institut für Statistik und Wahrscheinlichkeitstheorie* in Vienna University!

Linear Support vector classifier

Work through subsection 9.6.1 in James et al on page 330.