Dr. Ralf Eggeling
Dana Petracek
Methods in Medical Informatics
University of Tuebingen

Assignment 7

Deadline: Monday, June 27, 2:00 p.m.

This problem set is worth 30 points. You can submit in groups of two people or alone. Submit your solutions by uploading them to moodle (none of the other students can see the files you upload). Name the files

`Assignment_7_[lastname].pdf` and `Assignment_7_[lastname].R`
for individual submissions and
`Assignment_7_[lastname1]_[lastname2].pdf` and `Assignment_7_[lastname1]_[lastname2].R`
for team submissions.

In the latter case, include names of both students at the top of both files. Both students must upload the identical files to moodle in time.

Moodle allows to upload **drafts**, which can then be further edited until the deadline. Use this for submitting finished tasks in case you might run out of time. Once you formally submit, the upload cannot be edited any longer and is ready for grading. If you never formally submit, the uploaded draft at the submission deadline will be graded.

Task 1 (30 Points)

Go through **8.3 Lab: Decision trees** (ISLR p.353–361).

Download and inspect the data set `Obesity.csv` from moodle. It includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The purpose of this assignment is to predict the categorical response *ObesityLevel* based on the 16 features and to study which of them are most relevant for that purpose. The first 1500 data points in the file are supposed to be used for training, the remaining 611 data points are the test data. To ensure reproducibility of the results, invoke `set.seed(42)` every time before you call a function that involves randomness.

(a) Learn a decision tree on the training data and display it graphically. What do you observe? (**3 Points**)

(b) Predict the *ObesityLevel* of the test data points using the learned decision tree, compute the accuracy and show the confusion matrix. Which classes are hard to predict? (**3 Points)**

(c) Learn a Naive Bayes classifier on the training data, inspect the parameters (no need to include all of them into the report), calculate test data accuracy, and compare the results to the decision tree. What do you observe? What is a likely explanation for your observation? *Useful function:* `naiveBayes()` from the `e1071` package. (**4 Points**)

(d) Learn a bagged tree ensemble and a random forest with default hyperparameters on the training data and calculate the test data accuracy for both models. What do you observe? (**4 Points**)

(e) Tune the *mtry*-parameter of the random forest by five-fold cross validation. For this purpose, split the training data in consecutive blocks (first block contains data points 1-300, second block 301-600, ...; the data is already shuffled). Plot mean cross validation accuracy (incl. standard errors) for each *mtry* value and mark the maximal value, $\widehat{mtry}$. (**4 Points**)

(f) Learn a random forest with the selected $\widehat{mtry}$ from the previous subtask on the entire training data set. Calculate test data accuracy, confusion matrix, and variable importance (mean decrease Gini). Compare with all previous results. (**5 Points**)

(g) Tune *mtry* according to the out-of-bag (OOB) error and compare to the CV-based tuning, both in terms of CV/OOB error for each *mtry*-value and performance of the optimal models on the final test data set. *Hint:* Remember the inverse relationship between (classification) error and accuracy. (**2 points**)

Note: 5 of the 30 points are awarded for presentation style, rewarding **clarity**.