Bonus point sheet 1

in spring 2018

for the course *Business Analytics*

(version: 9th May 2018)

1. The deadline for this sheets ends on **May 27, 2018** (end of day).
2. This sheet contains in total **3 questions**.
3. You receive **4.5 points** (i. e. 5 %) as bonus points for the exam if 50 % are answered correctly. However, these are only added if you pass the exam without bonus points.
4. Group work is not allowed.
5. Please label **your submission with your matriculation number, as well as first and last name.**
6. Submit your answer sheet through the designated submission system inside Moodle!
7. Write your answer sheet in **R Markdown** (`http://rmarkdown.rstudio.com/`). In addition, e offer an introductory guide as part of the course materials. Provide the answers to the questions also inside the file.
8. All questions cover in total **3 pages**.

Good luck!

Exercise 1: Data visualization in R

□

The file `survey.csv` is given containing various columns. This file contains answers that were submitted during a survey among students.

.5.5

0.5

a) Use the package `ggplot2` to create a point plot comparing the age and pulse of the participants. Color each point according to the gender of the person and additionally label the axis appropriately. Furthermore, change the color of the points such that males are highlighted with a dark blue color and females in dark red.

.5.5

0.5

b) Use `ggplot2` to create boxplot comparing the age distribution of males and females. Cut the y-axis at a maximum age of 30 for better readability. Interpret your results.

Exercise 2: Data analysis in R

□

We now study a dataset called `mtcars` that is shipped inside R. It contains different properties of cars, such as gas mileage and multiple variables. The dataset can be loaded via:

```r
data("mtcars")
head(mtcars)
```

You can find details on the dataset via `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html`. In the following, we want to investigate different properties that affect the overall fuel efficiency. For this purpose, we will draw upon a linear regression model and perform a corresponding analysis. The insights could later help policy-makers in designing schemes for incentivizing cars with low consumption.

.5.5

0.5

a) The first step in data analysis is to establish a model specification. In our example, we follow common practice and start with a linear model. The dependent variable is given by the fuel consumption (i. g. miles per gallon), as measured by variable mpg. How is the weight wt of a car affecting its consumption? For this purpose, estimate a model $mpg = \beta_0 + \beta_1 wt$. Afterwards, interpret the results with regard to our initial question. Does the coefficient β have the expected sign?

Now use `ggplot2` in order to visualize the trendline between the two variables, i. e. wt and mpg. In addition, add the observations to the plot and add labels to the axis.

.5.5

0.5

b) Now estimate a new model (called M in the following) which additionally contains the number of cylinders cyl and the number of forward gears $gear$:

$$mpg = \beta_0 + \beta_1 wt + \beta_2 cyl + \beta_3 gear. \tag{1}$$

.5.5 Explain the latter output briefly!

0.5

c) Split the data into training and test set. Use 30 % of the data for training and 70 % of the data for testing. Reestimate the model m by only using training data. Predict the mpg variable for the test data. What is the root mean squared error?

Exercise 3: Machine learning in R

□

A commonly used reference dataset in Machine Learning is the so-called *Iris flower data set* (for details, see `http://en.wikipedia.org/wiki/Iris_flower_data_set`). In this dataset, three species of the *iris* flower are estimated based on the size of various features. It can be loaded in R as follows:

```
library(e1071)
data(iris)
head(iris)
```

.5.5

0.5

a) Use the k-nearest neighbor method with $k = 3$ to predict the species with the properties `Sepal.Length=7`, `Sepal.Width=2`, `Petal.Length=1.5` and `Petal.Width=2`!

.5.5

0.5

b) Train a decision tree for the dataset! Use 30 % of the data for training and 70 % of the data for testing. Then display the confusion matrix and calculate the out-of-sample accuracy of the classifier?

.5.5

0.5

c) .5.5 In the next step, visualize the decision tree before and after tuning.

0.5

d) Now compare the prediction performance of a random forest. In addition, compute the variable importance based on the random forest. Which predictor is most relevant according to the variable importance metric?