Bonus point sheet 2

in spring 2018

for the course *Business Analytics*

(version: 30th May 2018)

1. The deadline for this sheets ends on **July 31, 2018** (end of day).
2. This sheet contains in total **4 questions**.
3. You receive up to **4.5 points** (i. e. 5 %) as bonus points for the exam. 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, we 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 **4 pages**.

Good luck!

Exercise 1: Model tuning in R

.5.5

1.5

a) Load the `GermanCredit` dataset from the `caret` package. Use this package also to train a support vector machine with linear kernel (`method="svmLinear"`). What is the out-of-sample error? You can use the following split into training and test set:

```
library(caret)

## Loading required package:  lattice

## Loading required package:  ggplot2

data(GermanCredit)

set.seed(0) # this is added so that the random number generator is always giving the same split, whic

inTrain <- createDataPartition(GermanCredit$Class, p = 0.2, list = FALSE)

training_set <- GermanCredit[inTrain, ]
test_set <- GermanCredit[-inTrain, ]
```

Hint: make sure that the following package is loaded:

```
library(kernlab)

##
## Attaching package:  'kernlab'

## The following object is masked from 'package:ggplot2':
##
##    alpha
```

Exercise 2: Bootstrapping

1

a) Load the `Carseats` dataset from the library `ISLR`. Compute the mean of the `Sales` variable. In a next step, we want to understand the potential distribution of the mean. For this purpose, create 20 boostrapped replicates of the data. By using `boot()`, what are the confidence intervals and how does the distribution look like?

Exercise 3: Clustering

1

a) Perform a k-means clustering to determine the origin of wines. Use $k = 3$ means with $n = 10$ random initializations. What is the within-cluster sum of squares (WCSS) value? As the input data, we use the dataset `wines` that is included in the `kohonen` package.

```
library(kohonen)
data(wines)
```

The dataset contains 177 rows and thirteen columns; object *vintages* contains the class labels. For compatibility with older versions of the package, variable wine.classes is retained too. This data is the result of chemical analyses of wines grown in the same region of Italy (Piedmont) but derived from three different cultivars: Nebbiolo, Barberas and Grignolino grapes. The wine from the Nebbiolo grape is called Barolo. The data contains the quantities of several constituents found in each of the three types of wines, as well as some spectroscopic variables.

```
head(wines)
```

```
##      alcohol malic acid  ash ash alkalinity magnesium tot. phenols
## [1,]   13.20       1.78 2.14           11.2       100         2.65
## [2,]   13.16       2.36 2.67           18.6       101         2.80
## [3,]   14.37       1.95 2.50           16.8       113         3.85
## [4,]   13.24       2.59 2.87           21.0       118         2.80
## [5,]   14.20       1.76 2.45           15.2       112         3.27
## [6,]   14.39       1.87 2.45           14.6        96         2.50
##      flavonoids non-flav. phenols proanth col. int. col. hue OD ratio
## [1,]       2.76              0.26    1.28      4.38     1.05     3.40
## [2,]       3.24              0.30    2.81      5.68     1.03     3.17
## [3,]       3.49              0.24    2.18      7.80     0.86     3.45
## [4,]       2.69              0.39    1.82      4.32     1.04     2.93
## [5,]       3.39              0.34    1.97      6.75     1.05     2.85
## [6,]       2.52              0.30    1.98      5.25     1.02     3.58
##      proline
## [1,]    1050
## [2,]    1185
## [3,]    1480
## [4,]     735
## [5,]    1450
## [6,]    1290
```

Use the above result from the clustering procedure to plot data points and clusters in a 2-dimensional plot showing only the dimensions `alcohol` and `ash`.

Exercise 4: Principal component analysis

1

a) Given the following dataset and its corresponding PCA:

```
library(ISLR)
data(Carseats)
Carseats$ShelveLoc <- as.integer(Carseats$ShelveLoc)
Carseats$Urban <- as.integer(Carseats$Urban)
Carseats$US <- as.integer(Carseats$US)

pca <- prcomp(Carseats, scale = TRUE)
```

Calculate the proportion of variance explained for all principal components. How many principal components are needed to explain 80% of total variance? Plot the cumulative proportional variance explained and insert a vertical line on the number of components needed.