Homework Assignment 3

Data Science and Machine Learning 1 - CEU 2018

Due date: 2018-02-12 8:00 a.m.

General information

The due date of this homework assignment is 2018-02-12 8:00 a.m. You are required to submit two files to Moodle: an `.Rmd` file and the rendered `.html` file with your solutions. Please also include code to the `.html` as well (use the default settings).

Please give short (2-3 sentences) interpretations, explanations to your answers, not only the program code and outputs.

Grading: 13 points can be earned in total. However, 10 points count as 100%. You can earn more than 100%, that is, giving a perfect answer to all questions means 130%.

```
library(data.table)
library(datasets)
library(MASS)
library(ISLR)
library(caret)
```

1. PCA for supervised learning (6 points)

In this problem you are going to analyze the `Boston` dataset from the `MASS` package (read more about the data here). The goal will be to predict the variable `crim` which is the crime rate.

```
data <- data.table(Boston)
```

a) Do a short exploration of data and find possible predictors of the target variable.
b) Create a training and a test set of 50%.
c) Use a linear regression to predict `crim` and use 10-fold cross validation to assess the predictive power.
d) Try to improve the model by using PCA for dimensionality reduction. Center and scale your variables and use `pcr` to conduct a search for the optimal number of principal components. Does PCA improve the fit over the simple linear model?
e) Use penalized linear models for the same task. Make sure to include lasso (`alpha = 0`) to your tune grid. How does the best model compare to that found in d)? Would pre-processing via PCA help this model? (add `pca` to `preProcess`). Why do you think the answer can be expected?
f) Evaluate your preferred model on the test set.

2. Clustering on the `USArrests` dataset (5 points)

In this problem use the `USArrests` dataset we used in class. Your task is to apply clustering then make sense of the clusters using the principal components.

a) Determine the optimal number of clusters as indicated by `NbClust` heuristics.
b) Use the k-means method to cluster states using the number of clusters found in a) and anything else that you think that makes sense. Plot observations colored by clusters in the space of urban population

and another (crime-related) variable. (See example code from class, use `factor(km$cluster)` to create a vector of class labels).

c) Perform PCA and get the first two principal component coordinates for all observations by

```
pca_result <- prcomp(data, scale. = TRUE)
first_two_pc <- data.table(pca_result$x[, 1:2])
```

Plot clusters in the coordinate system defined by the first two principal components. How do clusters relate to these?

3. PCA of high-dimensional data (2 points)

In this exercise you will perform PCA on 40 observations of 1000 variables. This is very different from what you are used to: there are much more variables than observations! These are measurments of genes of tissues of healthy and diseased patients: the first 20 observations are coming from healthy and the others from diseased patients.

```
data <- fread("../../data/gene_data_from_ISLR_ch_10/gene_data.csv")
data[, is_diseased := factor(is_diseased)]
dim(data)
tail(names(data))
```

a) Perform PCA on this data with scaling features.

```
data_features <- copy(data)
data_features[, is_diseased := NULL]
```

b) Visualize datapoints in the space of the first two principal components (look at the `fviz_pca_ind` function). What do you see in the figure?

c) Which individual features can matter the most in separating diseased from healthy? A strategy to answer this can be the following:

- we see that PC1 matters a lot
- so look at which features have high **loadings** for the first PC, that is, the largest coordinates (in absolute terms). (Hint: use the `$rotation`). Choose the two features with the largest coordinates and plot observations in the coordinate system defined by these two original features. What do you see?

PCA thus offers a way to summarize vast amounts of variables in a handful of dimensions. This can serve as a tool to pick interesting variables where, for example, visual inspection would be hopeless.