Statistical Learning Statistics and Data Science: Assignment 1

Please make sure you answer each question clearly in full sentences and indicate the question number. You may upload your code in a separate file but your answers should contain all necessary information.

Part A. Supervised learning

Data

For this part of the assignment each student will download their own data. Please go to https://solo-fsw.shinyapps.io/GenerateDatasetSDS/, put in your student number (without the s), and download your data set. This will generate a data set of 10000 cases and 204 variables: one outcome Y and 203 predictors X_1-X_{203}. The first three predictors (X_1-X_3) are relevant for the prediction of Y, whereas the other predictor variables (X_4-X_{203}) are noise variables that are not related to Y (except for sampling fluctuations; check the generation code [*GenerateDataSetSDS.R* on Brightspace] to see how the data were generated). You still need to create a training set (train) and a test set (test). To this end, take the first 5000 cases as your training set and the next 5000 cases as your test set. The full training and test data set can be generated by means of the following code, which assumes that your student number is 123.

R:

```
library(readr)
data_set <- read_csv("Data123.csv")
train <- data_set[1:5000, ]
test <- data_set[5001:10000, ]
```

Python:

```
import pandas as pd
data_set = pd.read_csv("Data123.csv")
train = data_set.iloc[0:5000,]
test = data_set.iloc[5000:,]
```

Questions

1) Consider only Y and X_1-X_6 (3 relevant + 3 irrelevant predictors).

   Look carefully at the data generating function and, for these variables, indicate whether you expect a K-nearest neighbours analysis or a LASSO logistic regression to lead to a lower expected prediction error (assuming 0-1 loss). Justify your answer in terms of the bias-variance trade-off (max. 300 words).

2) Repeat question 1 but considering Y and X_1-X_{203} (all X-predictors). Compare your answer also with the answer of question 1. Are your expectations the same? Explain why or why not.

3) Consider only Y and X_1-X_6 (3 relevant + 3 irrelevant predictors).
   a) Select the optimal K for the KNN classifier using 10-fold cross-validation. Estimate the accuracy (as an estimate of the expected prediction error) of the optimal KNN classifier. Describe your results and the procedure followed to obtain these results.
   b) Select the optimal lambda for LASSO logistic regression using 10-fold cross-validation. Estimate the accuracy of LASSO logistic regression with the selected lambda. Describe your results and the procedure followed to obtain these results.
   c) Compare the accuracy estimates for both methods (from questions 3a and 3b) in light of the answer to question 1. Are the results as expected? Explain why or why not this is the case. (Note that this question is graded separately from question 1, so it may be needed to repeat some arguments given in question 1).

4) Consider Y and X_1-X_{203} (all X-predictors)
   a) Repeat 3a) using all predictors
   b) Repeat 3b) using all predictors
   c) Compare the accuracy estimates of both methods (from questions 4a and 4b) in light of the answer to question 2. Are the results as expected? Explain why or why not this is the case. (Note that this question is graded separately from question 2, so it may be needed to repeat some arguments given in question 2). Also, compare the results of questions 3c and 4c and explain the differences in results in light of the answers to questions 1 and 2. (Again, this question is graded separately from questions 1-2).

Part B. Unsupervised learning

For this part of the assignment, please use the attached dataset (*data.US.csv* or *data.US.txt*). The dataset consists of 1000 individuals who have been measured on 30 personality variables (facets). The variables are from V2 to V31: anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability, warmth, gregariousness, assertiveness, activity, excitement-seeking, positive emotions, fantasy, aesthetics, feelings, ideas, actions, values, trust, straightforwardness, altruism, compliance, modesty, tendermindedness, competence, order, dutifulness, achievement striving, self-discipline, deliberation. The variables already have been standardised.

Our aim is to identify groups of individuals with similar personalities. However, some of the variables are highly correlated so some dimension reduction may be needed. Analyse the dataset and answer the following questions:

B1. Do you think that reducing the variables to a smaller set of (new/derived) variables may be a good idea? Which statistical technique can be used to achieve such a dimension reduction? Explain why this technique is appropriate.

B2. How many new/derived variables should be computed to capture the most important part of the information in the original variables? Use at least four different methods to select the number of new/derived variables. Explain each method and thoroughly justify your final decision (e.g., by providing relevant figures).

B3. Can you give a meaning to these new/derived variables?

B4. Can you somehow quantify the degree to which the new/derived variables capture the information in the original variables?

Is it possible to group the study participants in terms of their personalities? To this end, use the new/derived variables (and not the original ones).

B5. Which statistical technique(s) can be used to group the participants? Explain why it is/they are appropriate.

B6. How many groups are there? How did you determine this? Thoroughly justify your answer (e.g., figures). If you identified more than one technique in the previous question, choose two and compare the techniques and the obtained results.

B7. How large is each group?

B8. If you identified more than one technique, what are the main differences between your chosen methods?

B9. What are the main differences between the groups in terms of personality?