Statistical Learning 2025

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.

Using generative AI tools is not allowed for this assignment.

Part A.

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.

The data for this part of the assignment are from a large epidemiological study on the course of depressive and anxiety disorders among adults living in the Netherlands. The data are from a subsample of 1500 subjects, who at the start of the study were suffering from an anxiety and/or depressive disorder, and were diagnosed as such. Twelve months after the start of the study, the severity of each subject's depressive symptoms was assessed again (variable dep_sev_fu; short for depression severity at follow-up). The goal of the analyses is to predict the severity of depressive symptoms after twelve months, using the characteristics that were assessed at the start of the study.

The dataset is available on Blackboard as 'MHpredict.csv'. It contains 20 potential predictor variables, which were assessed at the start of the study, and are described in Appendix I. You can read it into R as follows:

```
MH_data <- read.table("MHpredict.csv", sep = ",", header = TRUE)
```

Questions

1. Select two supervised learning methods from those that were discussed so far for analyzing this dataset. Justify why you would select each of these methods for this specific prediction problem.

2. Apply the two methods you selected to the dataset. Beforehand, randomly split the dataset into a training (n=1000) and test (n=500) dataset. Use your student number to set the seed of the random number generator. Motivate your choice of the main model-fitting parameters. Thus, make a well-informed choice and/or use cross-validation to set their values (even if you choose to use default settings).

3. Provide an interpretation of each of the resulting models:

- Describe which variables are most important in determining the value of the outcome variable.
- Describe the effect of the most important variables (e.g., describe the shape and direction of the effect on the outcome and/or provide and discuss plots of the variables' effects).
- Assess and compare the predictive accuracy of each of the models using the test set. Which model predicts best?