**Statistical Learning Statistics and Data Science (2022-2023): Assignment 1**

**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/](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fsolo-fsw.shinyapps.io%2FGenerateDatasetSDS%2F&data=05%7C01%7Cj.d.karch%40fsw.leidenuniv.nl%7C769ea403f9b0410180e508dab0342d86%7Cca2a7f76dbd74ec091086b3d524fb7c8%7C0%7C0%7C638016035386349479%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=zUjHFXPs4x3dXx2ULR5yn6ALrTi7Bzw5hYKlNuuZF64%3D&reserved=0), put in your studentnumber (without the s), and download your data set. This will generate a data set of 10000 cases and 204 variables: one criterium and 203 predictors -. The first three predictor variables (-) are relevant for the prediction of , whereas the other predictor variables (-) are noise variables that are not related to (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 you 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 and - (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 logistic regression with a lasso penalty to give a better performance. Justify your answer in terms of bias, variance, and MSE (max. 300 words).

1. Consider and - (all -predictors)

For these variables, indicate whether you expect a K-nearest neighbours analysis or a logistic regression with a lasso penalty to give a better performance. Justify your answer in terms of bias, variance, and MSE. Compare your answer also with the answer on question 1. Are your expectations the same? Explain why or why not (max. 300 words).

1. Consider only and - (3 relevant + 3 irrelevant predictors)
2. Apply a K-nearest neighbour analysis to determine the model accuracy for this data of the (optimal) KNN classifier. Do not forget to also train the metaparameter K (i.e., model selection). For the learning of the metaparameter K, use 10-fold cross-validation. Describe your results and the procedure followed to obtain these results.
3. Apply logistic regression with a lasso penalty to determine the model accuracy for this data of the (optimal) logistic lasso-classifier. Do not forget to also train the metaparameter λ (i.e., model selection). For the learning of the metaparameter λ, use 10-fold cross-validation. Describe your results and the procedure followed to obtain these results.
4. For both methods, give the estimate of the test misclassification error rate (using an appropriate procedure) of the optimal model (as determined in 3a/3b). Also describe the procedure used to obtain this estimate. Compare this estimate of test error of both methods (from question 3a and 3b) in light of the answer on 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).
5. Consider and - (all -predictors)
6. Apply a K-nearest neighbour analysis to determine the model accuracy for this data of the (optimal) KNN classifier. Do not forget to also train the metaparameter K (i.e., model selection). For the learning of the metaparameter K, use 10-fold cross-validation. Describe your results and the procedure followed to obtain these results.
7. Apply logistic regression with a lasso penalty to determine the model accuracy for this data of the (optimal) logistic lasso-classifier. Do not forget to also train the metaparameter λ (i.e., model selection). For the learning of the metaparameter λ, use 10-fold cross-validation. Describe your results and the procedure followed to obtain these results.
8. For both methods, give the estimate of the test misclassification error rate (using an appropriate procedure) of the optimal model (as determined in 4a/4b). Also describe the procedure used to obtain this estimate. Compare this estimate of test error of both methods (from question 4a and 4b) in light of the answer on 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 of questions 1 and 2. (Again, this question is graded separately from questions 1-2).

**Part B. Unsupervised learning**

For this part of the assignment, you should use the Coping data set (*Coping.csv*). The data set consists of 300 cancer patients that have been measured on 9 coping styles (in the order of appearing): (*var1*) self-blame, (*var2*) accept, (*var3*) rumination, (*var4*) concentrate on other positive things, (*var5*) focus on planning, (*var6*) positive reinterpretation, (*var7*) put in perspective, (*var8*) dramatize, and (*var9*) blaming others. The variables already have been centered. A high (low) value implies that the patient uses the coping style in question to a large (small) extent. The main interest of the therapist who collected these data is to obtain a grouping of the patients in terms of their coping behavior. The therapist, however, noted that some of the variables are highly correlated and therefore thinks that prior to the grouping of the patients some dimension reduction may be needed. Analyze the data to help out this therapist and answer the following questions:

1. 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.
* how many new/derived variables should be computed in order to capture the most important part of the information in the original variables? Thoroughly justify your answer (e.g., figures).
* what do these new/derived variables represent from a substantive point of view? Give an interpretation of the content of these new variables and explain which part of the output is relevant to answer this question.
* can you somehow quantify the degree to which the new/derived variables capture the information in the original variables?

1. Is it possible to group the patients in terms of their coping behavior? To this end, use the **new/derived variables (and not the original ones)**

* which statistical technique can be used to group the patients? Explain why it is appropriate.
* how many patient groups are there? How did you determined this? Thoroughly justify your answer (e.g., figures).
* how large is each patient group?
* what are the main differences between the patient groups in terms of coping behavior (so give a substantive interpretation of the obtained groups and their main differences)?

Report

Guidelines for the report

* You do NOT need to make a report with introduction, methods and conclusions. Just make a document in which you answer to the questions (and make clear which answer belongs to which question).
* When using a procedure, like, for example, cross-validation, also explain the principles of the procedure. Also when presenting results (e.g., in tables and figures), clearly explain what the results mean (how I can see this in the table/figure) and what your conclusion is. You have to convince me that you master the subject!
* As an appendix you should add all R / Phyton-code used for the analyses (add comments to the syntax to make the code understandable for us). Do not refer to this appendix in your report (the answers must be self-contained).

The report (with separate R / Phyton-code) should be handed in

* Via Brightspace (word or pdf)

The deadline for the assignment is Thursday November 10th (2022) 23:59h.

Good luck !