Project description for report 2

Objective: The objective of this second report is to apply the methods you have learned in the second section of the course on *"Supervised learning: Classification and regression"* in order to solve both a relevant classification and regression problem for your data.

Material: You can use the 02450Toolbox on Campusnet to see how the various methods learned in the course are used in Matlab, R or Python. In particular, you should review exercise 5 to 9 in order to see how the various tasks can be carried out.

Preparation: Exercise 1–9

Project report 2 should naturally follow project report 1 on *"Data: Feature extraction and visualization"* and cover what you have learned in the lectures and exercises of week 5 to 8 on *"Supervised learning: Classification and regression"*. The report should therefore include two sections. A section on regression and a section on classification. The material to be covered in each of these two sections is outlined below and the report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality.

Regression: In this section of the report you are to solve a relevant regression problem for your data. In particular, you should:

1. Explain which regression problem you have chosen to solve.
2. Apply linear regression with forward selection and consider if transforming or combining attributes potentially may be useful. For linear regression, plotting the residual error vs. the attributes can give some insight into whether including a transformation of a variable can improve the model, i.e. potentially describe parts of the residuals.
3. Explain how a new data observation is predicted according to the estimated model. I.e. what are the effects of the selected attributes in terms of predicting the data.
   (Notice, if you interpret the magnitude of the estimated coefficients this in general requires that each attribute be normalized prior to the analysis.).
4. Fit an artificial neural network (ANN) model to the data.
5. Statistically evaluate if there is a significant performance difference between the fitted ANN and linear regression models based on the same cross-validation splits (i.e., considering the credibility interval equivalent to the use of a paired t-test as described in lecture 6 and last exercise week 6). Compare in addition if the performance of your models are better than simply predicting the output to be the average of the training data output.

Classification: In this part of the report you are to solve a relevant classification problem for your data. In particular, you should:

1. Explain which classification problem you have chosen to solve.
2. Apply at least three of the following methods:
   Decision Trees, Logistic/Multinomial Regression, K-Nearest Neighbors (KNN), Naïve Bayes and Artificial Neural Networks (ANN).
   (Use cross-validation to select relevant parameters in an inner cross-validation loop and give in a table the performance results for the methods evaluated on the same cross-validation splits on the outer cross-validation loop, i.e. you should use two levels of cross-validation).
3. For the models you are able to interpret explain how a new data observation is classified.
   (If you have multiple models fitted, (i.e., one for each cross-validation split) either focus on one of these fitted models or consider fitting one model for the optimal setting of the parameters estimated by cross-validation to all the data.)
4. Statistically compare the performance of the two best performing models (i.e., considering the credibility interval equivalent to the use of a paired t-test as described in lecture 6 and last exercise week 6). Compare in addition if the performance of your models are better than simply predicting all outputs to be the largest class in the training data.

If your data has previously been analyzed by regression or classification in the literature, please report what methods have been used previously as well as their performance and relate your results to these previous results.

Notice, if the analysis of your data is too computationally demanding for choosing parameters in the inner cross-validation loop we suggest you use the hold-out method instead of K-fold cross-validation. Furthermore, if analyzing the data by ANN is too computationally demanding you can consider only analyzing a subset of your data by ANN.

The report should be 5-10 pages long including figures and tables and give a precise and coherent account of the results of the regression and classification methods applied to your data. **Each group member will be main responsible for a given part of the report. You therefore have to specify who have been responsible for each part of the report as well as outline how each member contributed to the report in an appendix to the report. All reports must contain this documentation in order to be accepted.** To ensure all group members get credit for the report, make sure also to put your names and study numbers on the front page and ensure you upload the report as a group hand in and put the name of your dataset on the front page. You cannot work in groups with more than 3 students.

Please hand in the report by uploading it as a single, uncompressed .pdf file to CampusNet no later than 14 November at 13:00.