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

Handin Checklist

- Specify **names *and* study numbers** of each group member on the front page
- According to the DTU regulations, each students contribution to the report must be clearly specified. Therefore, for each section, specify which student was responsible for it (use a list or table). **A report must contain this documentation to be accepted**
- Your handin should consist of a `.pdf` file containing the report, and the code you have used as one (or more) files with the extension `.py`, `.R` or `.m`. The reports are not evaluated based on the quality of the code (comments, etc.), however we ask the code is included to avoid any potential issues of illegal collaboration between groups. Please do not compress or convert these files.
- Reports are evaluated based on how well they address the questions below. Therefore, to get the best evaluation, address all questions
- Use the group handin feature on campusnet. **Do not upload separate reports for each team member as this will lead to duplicate work and unhappy instructors**
- **Deadline for handin is no later than 13 November at 13:00**. Late handins will not be accepted under normal circumstances

Description

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 and statistically evaluate the result. In particular, you should:

1. Explain which regression problem you have chosen to solve. The explanation of the problem should mention what variable is predicted based on which other variables and a brief discussion of appropriate feature transformation choices such as one-of-K coding.[1].

2. Apply linear regression with forward selection to your dataset and report the selected parameters.

3. Explain how a new data observation is predicted according to the linear model estimated by forward selection. 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. Select a relevant complexity-controlling parameter (such as number of hidden neurons) and apply two-level cross-validation to both optimize the parameter and estimate the generalization error. Recall in two-level cross-validation (see Algorithm 5 of the lecture notes), each inner loop selects an "optimal" value of the parameter and estimate the generalization error for that selected value (this is then repeated in each fold of the outer loop). We want you to produce a table (or graph) that shows, for each iteration of the outer loop, the selected value of the parameter (s^* in the notation of the lecture notes, algorithm 5) and the corresponding value of the generalization error (E_i^{test}). Finally compute the two-level cross-validation estimate of the generalization error and also include this in the report. [2]

5. Statistically evaluate if there is a significant performance difference between the fitted ANN and linear regression models using the credibility-interval method discussed in the lecture notes (Example 2 in section 9.4.3) as well as a baseline.

[1] We treat feature transformations and linear regression in a very condensed manner in this course. Note for real-life applications, it may be a good idea to consider interaction terms and the last category in a one-of-K coding is redundant (you can perhaps convince yourself why). We consider this out of the scope for this report

[2] Note for this, and the subsequent cross-validation questions, it is vital you compute the error as the *average* errors on your test set (and not the sum) as is done in the lecture notes. This is important because you want to compute their absolute magnitudes.

Recall that to accomplish this, the linear regression model and the ANN needs to be evaluated on the same cross-validation splits. Therefore, re-use the outermost cross-validation splits used in the previous question and then compute the test-errors of the linear regression model on the same splits.

In addition to this, compare if the performance of your models are better than a simple baseline obtained by predicting the output to be the average of the training data. In other words complete three test: Linear regression vs. ANN, linear regression vs. average output, ANN vs. average output and discuss your findings.

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

1. Explain which classification problem you have chosen to solve.

2. Apply three of the following methods:
Decision Trees, Logistic/Multinomial Regression, K-Nearest Neighbors (KNN), Naïve Bayes and Artificial Neural Networks (ANN).

   For the three selected methods, select three relevant complexity-controlling parameters and re-do the two-level cross-validation exercise above. In other words, produce a table that shows for each outer split the estimated generalization error E_i^{test} for each of your three methods. Finally, combine these to obtain the two-level cross-validation estimate of the generalization error for each of your three methods and include the result in the report.

3. Select one of your three methods and explain how a new data observation is classified.
(As you have multiple models fitted corresponding to different cross-validation splits, either focus on one of these fitted models or re-estimate the model on all the data and a plausible value of your complexity-controlling parameter)

4. Select the two best-performing models (as estimated by two-level cross-validation) as well as a baseline model that predicts everything as belonging to the largest class in the training data. Compare these three models (model 1 vs. model 2, model 1 vs. baseline, model 2 vs. baseline) using the techniques in Chapter 9 (see Example 1). As for regression, remember to re-use the test splits. Include a small discussion of your results.

5. Include a small section that discuss if your data has previously been analyzed by regression or classification in the literature. If so, report what methods have been used 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.

Transferring/reusing reports from previous semesters

If you are retaking the course, you are allowed to reuse your previous report. You can either have the report transferred in it's entirety, or re-work sections of the report and have it evaluated anew.

If you wish to have your report transferred without changes, please do not upload it to campusnet (as this will lead to unnecessary work), but rather contact the teacher shortly after the written exam and make sure to include your study number and which semester the report was originally handed in and evaluated.

If you wish to redo parts of a report you have already handed in as part of a group in a previous semester, then to avoid any issues about plagiarism please keep attribution to the original group members for those sections you choose not to redo.