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.

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 9 April 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 report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality.

Regression, part a: In this section, you are to solve a relevant regression problem for your data and statistically evaluate the result. We will begin by examining the most elementary model, namely linear regression.

1. Explain what variable is predicted based on which other variables and what you hope to accomplish by the regression. Mention your feature transformation choices such as one-of-K coding. Since we will use regularization momentarily, apply a feature transformation to your data matrix $\mathbf{X}$ such that each column has mean 0 and standard deviation 1[1].

2. Introduce a regularization parameter λ as discussed in Chapter 13 of the lecture notes, and estimate the generalization error for different values of λ. Specifically, choose a reasonable range of values of λ (ideally one where the generalization error first drop and then increases), and for each value use $K = 10$ fold cross-validation (Algorithm 5) to estimate the generalization error.

   Include a figure of the estimated generalization error as a function of λ in the report and briefly discuss the result.

3. Explain how a new data observation is predicted according to the linear model with the lowest generalization error as estimated in the previous question. I.e., what are the effects of the selected attributes in terms of determining the predicted class. Does the result make sense?

Regression, part b: In this section, we will compare three models: the regularized linear regression model from the previous section, an artificial neural network (ANN) and a baseline. We are interested in two questions: Is one model better than the other? Is either model better than a trivial baseline?. We will attempt to answer these questions with two-level cross-validation.

1. Implement two-level cross-validation (see Algorithm 6 of the lecture notes). We will use 2-level cross-validation to compare the models with $K_1 = K_2 = 10$ folds[2]. As a baseline model, we will apply a linear regression model with no features, i.e. it computes the mean of y on the training data, and use this value to predict y on the test data.

   Make sure you can fit an ANN model to the data. As complexity-controlling parameter for the ANN, we will use the number of hidden units[3] h. Based on

[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] If this is too time-consuming, use $K_1 = K_2 = 5$

[3] Note there are many things we could potentially tweak or select, such as regularization. If you wish to select another parameter to tweak feel free to do so.

Outer fold	ANN		Linear regression		baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	10.8	0.01	12.8	15.3
2	4	10.1	0.01	12.4	15.1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	3	10.9	0.05	12.1	15.9

Table 1: Two-level cross-validation table used to compare the three models

a few test-runs, select a reasonable range of values for h (which should include $h = 1$), and describe the range of values you will use for h and λ.

2. Produce a table akin to Table 1 using two-level cross-validation (Algorithm 6 in the lecture notes). The table shows, for each of the $K_1 = 10$ folds i, the optimal value of the number of hidden units and regularization strength (h_i^* and λ_i^* respectively) as found after each inner loop, as well as the estimated generalization errors E_i^{test} by evaluating on $\mathcal{D}_i^{\text{test}}$. It also includes the baseline test error, also evaluated on $\mathcal{D}_i^{\text{test}}$. Importantly, you must re-use the train/test splits $\mathcal{D}_i^{\text{par}}, \mathcal{D}_i^{\text{test}}$ for all three methods to allow statistical comparison (see next section).

   Note the error measure we use is the squared loss *per observation*, i.e. we divide by the number of observation in the test dataset:

$$E = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} (y_i - \hat{y}_i)^2$$

   Include a table similar to Table 1 in your report and briefly discuss what it tells you at a glance. Do you find the same value of λ^* as in the previous section?

3. Statistically evaluate if there is a significant performance difference between the fitted ANN, linear regression model and baseline using the credibility-interval method discussed in the lecture notes (Example 2 in section 10.4.3). These comparisons will be made pairwise (ANN vs. linear regression; ANN vs. baseline; linear regression vs. baseline). Since you evaluated on the same test/training splits, this can be done by using the numbers in the table you just computed.

   The output of the three comparisons are three intervals which reflect our estimate of the mean of the difference in generalization error. Include these intervals in your report and interpret what they say about your models. Is one model better than the other? Are the two models better than the baseline? Are some of the models identical? What recommendations would you make based on what you've learned?

Classification: In this part of the report you are to solve a relevant classification problem for your data and statistically evaluate your result. The tasks will closely mirror what you just did in the last section. The three methods we will compare is a baseline, logistic regression, and **one** of the other four methods from below (referred to as *method 2*).

Logistic regression for classification. Once more, we can use a regularization parameter $\lambda \geq 0$ to control complexity

ANN Artificial neural networks for classification. Same complexity-controlling parameter as in the previous exercise

CT Classification trees. Same complexity-controlling parameter as for regression trees

KNN k-nearest neighbor classification, complexity controlling parameter $k = 1, 2 \ldots$

NB Naïve Bayes. As complexity-controlling parameter, we suggest the term $b \geq 0$ from section 11.2.1 of the lecture notes to estimate[4] $p(x = 1) = \frac{n^+ + b}{n^+ + n^- + 2b}$

1. Explain which classification problem you have chosen to solve. Is it a multi-class or binary classification problem?

2. We will compare logistic regression[5], *method 2* and a baseline. For logistic regression, we will once more use λ as a complexity-controlling parameter, and for *method 2* a relevant complexity controlling parameter and range of values. We recommend this choice is made based on a trial run, which you do not need to report. Describe which parameter you have chosen and the possible values of the parameters you will examine.

   The baseline will be a model which compute the largest class on the training data, and predict everything in the test-data as belonging to that class (corresponding to the optimal prediction by a logistic regression model with a bias term and no features).

3. Again use two-level cross-validation to create a table similar to Table 1, but now comparing the logistic regression, *method 2*, and baseline. The table should once more include the selected parameters, and as an error measure we will use the error rate:

$$E = \frac{\{\text{Number of misclassified observations}\}}{N^{\text{test}}}$$

   Once more, make sure to re-use the outer validation splits to admit statistical evaluation. Briefly discuss the result.

[4]In Python, use the `alpha` parameter in `sklearn.naive_bayes` and in R, use the `laplacian` parameter to `naiveBayes`. We do *not* recommend NB for Matlab users, as the implementation is somewhat lacking.

[5]in case of a multi-class problem, substitute logistic regression for multinomial regression

4. Perform a statistical evaluation of your three models similar to the previous section. That is, compare the three models pairwise and include in your report the credibility intervals as estimated using the method in Example 2, section 10.4.3 of the lecture notes.

   Discuss your results. Based on the intervals, is one of the models performing better than the others? What recommendations would you make and what did you learn?

5. Train a logistic regression model using a suitable value of λ (see previous exercise). Explain how the logistic regression model make a prediction. Are the same features deemed relevant as for the regression part of the report?

Discussion:

1. Include a discussion of what you have learned in the regression and classification part of the report.

2. If your data has been analyzed previously (which will be the case in nearly all instances), find a study which uses it for classification, regression or both. Discuss how your results relate to those obtained in the study. If your dataset has not been published before, or the articles are irrelevant/unobtainable, this question may be omitted but make sure you justify this is the case.

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.

To have a report transferred, *do absolutely nothing*. Reports from previous semesters are automatically transferred. Therefore, please do not upload old reports to campusnet as this will lead to duplicate work. As a safeguard, we will contact all students who are missing reports shortly after the exam.

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.