Project description for Continuing Education

Objective: The report will cover a selection of topics taught in the course and summarizes the report steps in the bottom of each exercise. The report will be divided into three parts, and for simplicity we recommend using a structure fairly closely matching the questions as outlined below.

Description

The report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality. Please match the format in terms of sections, etc. in your handin.

Part I

Understanding the data you are trying to model well is very important. You can apply very sophisticated machine learning methods but if you are not aware of potential issues with the data the further modeling will be difficult. Thus, the aim of this first part is to get a thorough understanding of your data and describe how you expect the data can be used for the later tasks.

This part will cover what you have learned in the section *"Data: Feature extraction, and visualization"* of the course. Your job is to make a useful description of the data set for your co-workers and make some basic plots.

1. **A description of your data set.**
   Explain
   - What the problem of interest is (i.e. what is your data about),
   - Where you obtained the data,
   - What has previously been done to the data. (i.e. if available go through some of the original source papers and read what they did to the data and summarize what were their results).
   - What the primary machine learning modeling aim is for the data, i.e. which attributes you feel are relevant when carrying out a classification, a regression, a clustering, an association mining, and an anomaly detection in the later reports and what you hope to accomplish using these techniques. For instance, which attribute do you wish to explain in the regression based on which other attributes? Which class label will you predict based on which other attributes in the classification task? If you need to transform the data to admit these tasks, explain roughly how you might do this (but don't transform the data now!).

2. **A detailed explanation of the attributes of the data.**

   - Describe if the attributes are discrete/continous, Nominal/Ordinal/Interval/Ratio,
   - give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so
   - describe the basic summary statistics of the attributes.

   If your data set contains many similar attributes, you may restrict yourself to describing a few representative features (apply common sense).

3. **Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).**
   Touch upon the following subjects, use visualizations when it appears sensible. *Keep in mind the ACCENT principles and Tufte's guidelines when you visualize the data.*

   - Are there issues with outliers in the data,
   - do the attributes appear to be normal distributed,
   - are variables correlated,
   - does the primary machine learning modeling aim appear to be feasible based on your visualizations.

   There are three aspects that needs to be described when you carry out the PCA analysis for the report:

   - The amount of variation explained as a function of the number of PCA components included,
   - the principal directions of the considered PCA components (either find a way to plot them or interpret them in terms of the features),
   - the data projected onto the considered principal components.

   If your attributes have very different scales it may be relevant to standardize the data prior to the PCA analysis.

4. **A discussion explaining what you have learned about the data.**
   Summarize here the most important things you have learned about the data and give also your thoughts on whether your primary modeling task(s) appears to be feasible based on your visualization.

Part II

This part should cover what you have learned in the lectures and exercises in section *"Supervised learning: Classification and regression"* of the course. You should include three sections: Two sections on regression and a section on classification.

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 14 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.

[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 $K_1 = K_2 = 5$

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
⋮	⋮	⋮	⋮	⋮	⋮
10	3	10.9	0.05	12.1	15.9

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

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 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 methods described in chapter 11. These comparisons will be made pairwise (ANN vs. linear regression; ANN vs. baseline; linear regression vs. baseline). We will allow some freedom in what test to choose. Therefore, choose either:

   setup I (section 11.3): Use the paired t-test described in Box 11.3.4)

   setup II (section 11.4): Use the method described in Box 11.4.1)

[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	Method 2		Logistic regression		baseline
i	x_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
⋮	⋮	⋮	⋮	⋮	⋮
10	3	10.9	0.05	12.1	15.9

Table 2: Two-level cross-validation table used to compare the three models in the classification problem.

Include p-values and confidence intervals for the three pairwise tests in your report and conclude on the results: 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,

[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

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 2, 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. Perform a statistical evaluation of your three models similar to the previous section. That is, compare the three models pairwise. We will once more allow some freedom in what test to choose. Therefore, choose either:

setup I (section 11.3): Use McNemera's test described in Box 11.3.2)

setup II (section 11.4: Use the method described in Box 11.4.1)

Include p-values and confidence intervals for the three pairwise tests in your report and conclude on the results: 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?

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.

Part III

The final part should include what you have learned in the third part of the course *"Unsupervised learning: Clustering and density estimation"*. In particular, you should perform clustering, outlier detection, and association mining on your data. We will therefore include three sections. A section on clustering, a section on outlier detection, and finally a section on association mining. **Clustering:** In this part of the report you should attempt to cluster your data and evaluate how well your clustering reflects the labeled information. If your data is a regression problem define two or more classes by dividing your output into intervals defining two or more classes as you did in report 2.

1. Perform a hierarchical clustering of your data using a suitable dissimilarity measure and linkage function. Try to interpret the results of the hierarchical clustering.

2. Cluster your data by the Gaussian Mixture Model (GMM) and use cross-validation to estimate the number of components in the GMM. Try to interpret the extracted cluster centers.

3. Evaluate the quality of the clustering in terms of your label information for the GMM as well as for the hierarchical clustering where the cut-off is set at the same number of clusters as estimated by the GMM.

Outlier detection/Anomaly detection: In this part of the exercise you should apply some of the scoring methods for detecting outliers you learned in Exercise 11. In particular, you should

1. Rank all the observations in terms of the Gaussian Kernel density (using leaveone-out), KNN density, KNN average relative density (ARD). (If the scale of each attribute in your data are very different it may turn useful to normalize the data prior to the analysis).

2. Discuss whether it seems there may be outliers in your data according to the three scoring methods.

Association mining: In this part of the report you are to investigate if there are associations among your attributes based on association mining. In order to do so you will need to make your data binary, see also exercise 12. (For categoric variables you can use the one-out-of-K coding format). You will need to save the binarized data into a text file that can be analyzed by the Apriori algorithm.

1. Run the Apriori algorithm on your data and find frequent itemsets as well as association rules with high confidence.

2. Try and interpret the association rules generated.