Aston University
Birmingham B4 7ET
http://vle.aston.ac.uk/

Data Mining
Coursework Specification 2019-20

Felipe Campelo

Module DS40DM November 13, 2019

1 Introduction

The goal of this coursework is to give you experience with the full process of developing a data mining solution. You will be given a set of *training* data for model development, and *test* data to make predictions on.

Your goals are:

- to follow a sound data mining application development process;
- to develop models that will generalise well to new data;
- to write a clear report on your findings.

These goals relate directly to this module's learning outcomes, since to successfully complete the assigned tasks you will need to demonstrate your understanding of the complete data mining process and your ability to clearly articulate the pattern analysis principles underlying your proposed solution to the problem. The application of multiple data mining methods for data preprocessing, exploration and modelling of a realistic DM problem also relates directly to this module's learning outcomes.

This coursework must be completed **individually**.

2 Task Details

In this coursework you will explore a dataset related to customers default payments in Taiwan. The study that generated this dataset took payment data in the month of October 2005 from a large bank in Taiwan. From the total data collected approximately 22% are the cardholders with default payment. There is a single class attribute to be predicted (`DEFAULT`), and the following variables are available as explanatory attributes:

- `LIMIT_BAL`: Amount of the given credit (NT$).
- `SEX`: Male or female
- `EDUCATION`: Educational level (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- `MARRIAGE`: Marriage status
- `AGE`: Age in years
- `PAY_0`: Repayment status one month prior (-2 = no outstanding payment balance (same as 0), -1=pay duly (same as 0), 0=pay duly, 1=payment delay for one month, 2=payment delay for two months, ..., 9=payment delay for nine months and above)
- `PAY_2 ... PAY_6`: Repayment status 2,...,6 months prior (same scale as `PAY_0`)
- `BILL_AMT1 ... BILL_AMT6`: Amount of bill statement, 1,...,6 months prior (NT$):
- `PAY_AMT1 ... PAY_AMT6`: Amount of previous payment, 1,...,6 months prior (NT$)
- `DEFAULT`: Default payment

The data was randomly divided into two groups, one for model training and the other for validation. The dataset you should use for training is called `CWdata_train.arff`. The classification goal is to predict, as accurately as possible, whether or not a client will default in their payment.

Carry out the data mining process in a systematic way. Take good notes on what you have done, and save data and models regularly so that experiments can be repeated if necessary (you will need to make a note of the seed for the random number generators as well).

More information about the dataset is available in a few links:

- I-ChengYeh and Che-hui Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”, Expert Systems and Applications 36(2):2473-2480, 2009: `https://doi.org/10.1016/j.eswa.2007.12.020`
- UCI Machine Learning repository: `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`
- Kaggle (also contains interesting discussions and tips): `https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset`

Exploratory data analysis and data preprocessing are two aspects of your solution that are as important as the models themselves. Be sure to spend some time exploring the dataset with visualisation tools, and consider carefully how you will treat outliers, missing data, or any other specific aspect of the dataset. Document your analysis decisions, as these should be reported with your final analysis. You may need/wish to pre-process some attributes prior to modelling.

Notice that distinct models may require different preprocessing steps - take the time to explore your options (using the Explorer tab) before deciding which steps to take.

Feel free to experiment with the wide range of techniques that Weka provides. You may like to try out tools for attribute selection, model combination, or classifiers that we haven't covered in the module. All I ask is that you describe what you are doing clearly and that you evaluate the models consistently. **If you use preprocessing filters, then you should do so using a *Filtered Classifier***, which ensures that the same preprocessing is applied in a consistent way to all the datasets on which you run the model, including the prediction sets. This is important so that (i) your tests can be automated using Weka's Experimenter interface, and (ii) your analysis becomes more easily reproducible.

You should solve this classification problem in two distinct ways:

- one on the basis of *both* classes having equal misclassification costs.
- another for a cost matrix where the cost of misclassifying a client with `DEFAULT == 1` as a zero is **5 times** greater than that of misclassifying a client with `DEFAULT == 0` as a one.

You should explore some different models before you decide which one is the best. More specifically, **you are required to explore *at least* four different models for each classification task** (e.g., regular and cost-sensitive versions of kNN, J48, Naïve Bayes, Random Forests and Multilayer Perceptron - under a Filtered Classifier wrapper, if preprocessing is to be done prior to classification). Don't settle for the standard options of each model, but take the time to perform some parameter tuning for each candidate model.

After comparing the models, you must select the best one for the equal-costs classification and the best one for the cost-sensitive problem. These selected models must then be used to make predictions on the test set `CWdata_test`. The equal cost model will be evaluated on prediction **accuracy**; the unequal cost model will be evaluated on the total *misclassification cost* on the prediction set.

You can use whichever Weka interface you like - try to take the best out of each interface. For instance, the Explorer is great for the preliminary exploration of data preprocessing steps and modelling approaches, while the Experimenter interface will help you to automate some repetitive experiments ()freeing you to do other things while they run) as well as provide you with some statistical comparisons that may be useful for model selection.

On Blackboard you can find a 'model' solution to a loosely similar task: **do not follow this slavishly**, since the data and the general activities were different! It is only there to give you an idea of the sorts of thought processes I am looking for.

3 Assessment

The submission deadline for the assignment is **5:00pm on 13 December 2019.** Submission will be *online* through Blackboard. Late submissions will be treated under the standard rules for Computer Science: *the lateness penalty will be 10% of the available marks for each working day*, with an **absolute** deadline of one week (i.e. 5:00pm on 20 December 2019), after which submissions will not be marked.

Your report **must** be submitted as a PDF file.[1] The report should contain the following Sections:

Abstract A brief description of the key points in the report.

Introduction The background of the problem.

Data Exploration What you learned from your initial analysis of the data.

Data Preprocessing What data preprocessing steps (such as discretisation, standardisation, outlier removal, feature extraction etc.) were necessary. Please provide a brief justification for your preprocessing decisions (e.g., why did you choose to deal with outliers in a particular way)

Classification Models Which models you applied, their comparative performance, and a justification for your choice of the best model.

Conclusion What you have learned about the data (and the data mining process in general) from doing the coursework.

The following **five** files must be submitted (as standalone files - please **do not** submit a single .ZIP file):

- The Report as a PDF file - please clearly indicate your name on the report. Please also do not forget to include in your report the accuracy and cost measures of your final selected models on `CWdata_test.arff`.
- Two **prediction** files: one for the equal-cost model and one for the unequal-cost model. The predictions are generated by running the best model found for each classification task on the `CWdata_test.arff` dataset. Copy the relevant text from the “Classifier output” pane in Weka and past it into a **plain text** file which you can save.[2]
- **Two** model files: one for the equal-cost model and one for the unequal-cost model. Each model can be saved directly from the Weka Explorer interface, by right-clicking the trained model name in the ”Result List” pane in Weka and selecting ”Save model”.

Note that `CWdata\_test.arff` should *never* be used during the model training or parameter tuning process (otherwise your evaluation of the out-of sample prediction accuracy will be

[1] You can create a PDF file easily from any other format: For instance, if you’re using Microsoft Word just click *Save as* and select PDF as the file format.

[2] To store predictions from the Explorer, first go to the Classify tab. Select ‘*Supplied test set*’ and load the test set. Then select ‘*More options ...*’, click on the ‘*Output predictions*’ button and select ‘*PlainText*’. This will output a column of predictions in the output area (before the usual summary information). You must then copy the whole output area and paste the text into a log file.

biased). You should only use `CWdata\_test.arff` after your have chosen the best-performing models using the `CWdata\_train.arff` set.

The whole report should be no more than about 15 pages in length (or about 5000 words), including figures, tables and references. Submissions that exceed this threshold may be subject to a grade penalty (10% penalty if the submission exceeds 5500 words).

Submit your files to Blackboard (click the "Coursework" item in the left menu and then click the submission link under "Coursework Submission", you should see the page where you can upload your files).

Assessment criteria:

The breakdown of marks is: 50 for quality of the analysis process, 30 for presentation of results and discussion, and 15 for prediction accuracy.

- Excellent (71+): Data exploration finds all the key aspects of the data characteristics, well-described evaluation of data preprocessing steps in relation to the key aspects discovered, systematic development and evaluation of some sensibly chosen baseline models, professionally presented data mining process, some innovative ideas and/or original thoughts.
- Good (56-70): Data exploration finds some key aspects of the data characteristics, test of a few data preprocessing steps, development and evaluation of some baseline models, clearly presented data mining process.
- Pass threshold (50-55): Data exploration finds very few key aspects of the data characteristics, preprocessing steps tested without justification, development and evaluation of some baseline models, missing many details in the report.
- Fail: Incomplete data mining process, report is too brief to cover all the aspects of data mining.

Assessment feedback: Feedback for this coursework will be provided as follows:

- Individual feedback will be provided in the form of a marking summary report.
- Face-to-face (or Skype) formative feedback will also be available via on demand pre-bookable meetings.