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

Data Mining
Coursework Specification 2019

Module CS4850 February 8, 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 a model that will generalise well to new data;
- to write a clear report on your findings.

Please note that you should complete the laboratory materials up to Lab 5 before attempting the main part of the task, and the materials up to Lab 6 before attempting the unequal cost version of the task. The lab sessions after week 8 will be set aside to help you with this coursework by providing guidance and feedback on what you have done.

Also please note that this coursework is to be completed **individually**. While you are allowed (and indeed encouraged) to discuss your ideas and compare approaches with your peers, the deliverables (the report and all associated files, as detailed at the **Assessment** section below) should present only your individual work.

2 Task Details

Your task is to develop a model for the dataset which is related with direct marketing campaigns of a Portuguese banking institution available from the CS4850 course website on Blackboard:

http://vle.aston.ac.uk

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to confirm if the product (bank term deposit) would (or not) be subscribed to. The dataset you should use for training is called `cworkTrain.arff`. Each entry in the dataset corresponds to someone who has subscribed to a bank term deposit or not. The classification goal is to predict, as accurately as possible, whether or not a client will subscribe to a term deposit (outcome attribute "termDeposit").

- Carry out the data mining process in a systematic way. I suggest that you follow a standard process (for an example, see the KDD process illustrated in Unit 1). 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!).
- Use the information at the head of the training data file to get a better understanding of the data.
- 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 important features of the dataset. Document your analysis decisions, as these should be reported with your final analysis.
- You may wish to pre-process the attributes (e.g., discretising continuous variables or attribute extraction using PCA) in order to apply specific models.
- 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 do want to use pre-processing filters, then you should do so using a '`Filtered Classifier`'. This classifier type appears in the '`meta`' folder when a classifier is selected in either the Explorer or Experimenter interface. This ensures that the same pre-processing is applied in a consistent way to all the datasets on which you run the model, including the prediction set.
- You should develop *two* distinct models: one on the basis of *both* classes having equal misclassification costs, and another for a cost matrix where the cost of misclassifying a client who will subscribe to a term deposit is 10 times that of misclassifying a client who will not. You should select the best model under equal costs and the best model under unequal costs and use each of them to make predictions on the test set. The equal cost model will be evaluated on prediction set accuracy; the unequal cost model will be evaluated on the total cost on the prediction set.

  In the Explorer interface, you can find two *meta-learners* for cost-sensitive learning. The cost matrix can be supplied as a parameter or loaded from a file in the directory set by the `onDemandDirectory` property, named by the relation name with extension `cost`. `CostSensitiveClassifier` either reweights training instances according to the total cost, or predicts the class with the least expected misclassification cost. `MetaCost` generates a single cost-sensitive classifier from the base algorithm using bagging. You may prefer to start with the first of these meta-learners.

- You can use whichever Weka interface you like (or even several of them). The Experimenter interface will help you to automate some repetitive experiments, freeing you to do other things while they run.
- To store predictions from the Explorer, first go to the Classify tab. Select '`Supplied test set`' and load the prediction set `cworkPredict.arff`. Then select '`More options ...`' and click on the '`Output predictions`' button and select '`PlainText`'. This will generate an additional column of predictions in the output area which you should then copy and paste into into a log file. *You will need to submit an electronic version of this log file.*
- On the Blackboard site you can find a 'model' solution to a similar task: you should not follow this slavishly, since the dataset was different, but it gives an idea of the sorts of thought processes I am looking for. In particular, the pre-processing and analytical methods that you use will be different since the task is different.

3 Assessment

The submission deadline for the assignment is **5:00pm on Tuesday 30 April 2019.** Submission will be *on-line* through the Blackboard site; your submission should consist of a single .zip file. Late submissions will be treated under the standard rules for Computer Science, with an **absolute** deadline of one week (i.e. 5:00pm on 7 May 2019), after which submissions will not be marked. The lateness penalty will be 10% of the available marks for each working day.

Your report (which must be a PDF file[1]) 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 to create a dataset for subsequent analysis. 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.

[1]You can create a PDF file easily from any other format on the lab machines. Simply open the file in its normal programme (e.g. Microsoft Word) and select the 'Print' command; instead of sending it to your normal printer, select PDF-Creator (PDF-Exchange will probably also work) and it will be output as a PDF file. Alternatively, Word documents can be exported as PDF files.

The zip file must include (electronic) text files containing a list of predictions that each of your selected models makes on the dataset `cworkPredict.arff` (i.e. two files, one for equal cost and one for unequal cost). If you are not sure how to create these prediction files, be sure to ask me or the lab assistants in a lab session. In your report, please include the accuracy and cost measures of your final selected models on `cworkPredict.arff`. Note that `cworkPredict.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 biased). You should only use `cworkPredict.arff` after your have chosen the best-performing models through k-fold cross validation.

The whole report should be no more than 15–20 pages in length (or about 5000 words), including figures, tables and references. Submissions that exceed this threshold will be subject to a grade penalty (10% penalty if the submission exceeds 5500 words). In addition, you should also submit clearly identified Weka files for the two models you have selected so that I can replicate your analysis and check your results.

In summary, you need to submit five files:

- Report as a pdf file - please clearly indicate your name on the report.
- Two prediction files: one for the equal-cost model and one for the unequal-cost model. The predictions are generated on the `cworkPredict.arff` as described previously. Copy the relevant text from the "Classifier output" pane in Weka and past it into a plain text file which you can save.
- Two model files: one for the equal-cost model and one for the unequal-cost model. Each model can be saved directly from Weka by right clicking the model name in the "Result List" pane in Weka.

Zip all five files and submit it to the 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 zip file).

Assessment criteria:

The breakdown of marks is: 50 for quality of the analysis process, 30 for presentation of results and discussion, and 20 for prediction accuracy.

- Excellent (71+): Data exploration finds all the key aspects of the data characteristics, critical 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 with many innovative ideas and 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.