Aston University
Birmingham B4 7ET

Data Mining
Coursework Specification

Ian T. Nabney

1 Introduction

The goal of this coursework is to give you experience of the whole lifecycle of developing a data mining application. You will be given a set of *labelled* data for model development, and *unlabelled* 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.

2 Task Details

Your task is to develop a model for the credit assessment dataset.

The dataset you should use for training is called `cwork06trainv2.arff`. Each entry in the dataset corresponds to someone who is a good or a bad credit risk, and your model should predict the outcome as accurately as possible.

- Carry out the data mining in a systematic way. I suggest that you follow a standard process (for example, see the CRISP-DM standard shown in lecture 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!).

- You should develop models on the basis of *both* equal costs and a cost matrix where the cost of misclassifying a bad credit risk is 5 times that of misclassifying a good credit risk. 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 test set accuracy; the unequal cost model will be evaluated on total cost.

  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 use the first of these meta-learners to start with.

- 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 treat outliers and missing data.

- You may wish to pre-process the attributes (for example discretising continuous variables) 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.

- You may use whichever Weka interface you like (or even several of them). To make predictions visible in the Explorer, first go to the Classify tab. Select 'Supplied test set' and load the prediction set. Then select 'More options ...' and click on the 'Output predictions' tick box. This will generate an additional column of predictions in the output area which you can then copy and paste into your report (or save in a log file).

3 Assessment

The submission date for the assignment is 3:30pm on Wednesday 29th March. Late submissions will be treated under the standard rules for Computer Science, with an **absolute** deadline of one week, after which submissions will not be marked. The lateness penalty will be 10% of the available marks for each working day.

Your report (which may be typed or handwritten, provided it is legible) should contain the following sections:

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

Introduction The background of the problem.

Data Preparation What data manipulation was necessary to create a dataset for analysis.

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

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 machine learning from doing the coursework.

Appendix A list of predictions that your selected model makes on the unlabelled dataset `cwork06predict.arff`. Note that the class variable has been set to unknown (the `?` character) so the error results are not meaningful.

The whole report (not including the appendix) should be no more than 10 pages in length (or about 5000 words), including figures and tables. In addition, you should also submit a disc or CD/DVD *firmly* attached to your report that contains Weka files for the two models you have selected, clearly identified.

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