Practical 3: Classification

Prof. M- Tahar Kechadi

School of Computer Science & Informatics
University College Dublin

The aim of this practical is to use RapidMiner to generate equations or classes from datasets using regression, classification algorithms, respectively. The datasets to be used can be found on Blackboard.

The deliverables for this practical are:

- A (results) document (e.g. a Word document) containing the equations, plots or classes (Text View, Plot View (Save Image...)) generated by RapidMiner, in a section per question. Any discussion should also be placed in this document.
- For each question, the RapidMiner process file should be exported and submitted.

All files generated should be placed in a zip file with the name <studentname> _<studentnumber>_ comp40370 practical3.zip, and submitted via blackboard.

Question 1 Prediction (I)

Using the MarkA.xls data set, generate a process that does the following:

1. Plot the data. Do Midterm Exam and Final Exam seem to have a linear relationship?

2. Add the W-SimpleLinearRegression operator with default parameters:
   a. Run the process to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

   b. Predict the final exam grade of a student who received an 86 on the midterm exam.

Question 2 Prediction (II)

Using the MarkB.xls data set, generate a process that does the following:

1. Add the Polynomial Regression operator with default parameters, run the process to find an equation for the prediction of a student's final exam grade based on the student's MCQ1 and MCQ2 grade in the course. Predict your final mark based on the first two MCQ marks.

2. Add the Polynomial Regression operator with default parameters, use local random seed with its default value, run the process to find an equation for the prediction of a student's final exam grade based on the student's MCQ1 and MCQ2 grade in the course. Compare this equation with the result of Question 2.1. Justify your answer.

Question 3 Classification with Decision Tree (I)

Using the borrower.xls data set, Defaulted Borrow is set as a label; generate a process that does the following:

1. Filter out the *TID* attribute, as its values are not useful for decision making.

2. Generate a decision tree with information gain (minimal size for split = 2, minimal leaf size = 2, minimal gain = 0.1, maximum depth = 20, select no pre/post pruning). Discuss the classification results.

3. Generate a decision tree with gain ratio (minimal size for split = 2, minimal leaf size = 2, minimal gain = 0.1, maximum depth = 20, select no pre/post pruning). Compare the classification results with the results of Question 3.1.

Question 4 Classification with Decision Tree (II)

Using the churn.xls data set, generate a process that does the following:

1. Filter out all attributes except CustServ Calls, Day Calls, Intl Calls and Churn?. Churn? is set as a label. Normalise the numerical data ([0..1]).

2. Generate a decision tree with gini_index and default parameters. Discuss the classification results.

3. Generate a decision tree with gini_index and default parameters, select no pruning. Discuss the classification results.

4. Generate a decision tree with information gain and default parameters. Compare the classification results with the results of Question 4.2. Select no pruning, discuss the classification results.