Practical III: Classification

Prof. M-Tahar Kechadi

School of Computer Science
University College Dublin

The aim of this practical is to use RapidMiner to generate equations or classes from data sets, using regression and classification algorithms. The data sets to be used can be found on Blackboard at the same location.

All the generated files should be placed in a zipfile with the name **<student name>_<student number>_comp40370_practical03.zip**, and submitted via Blackboard.

Question1 Prediction (1)

Using the MarkA.xls data set, generate a process which 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 students final exam grade based on the students midterm grade in the course.
   b) Predict the final exam grade of a student who received an 86 on the midterm exam.

   Note:

   Before setting the `W-SimpleLinearRegression` operator, you will need to set the Set Role operator first, with the following parameters (`AttributeName:FinalExam, TargetRole = Prediction, SetAdditionalRole: AttributeName = FinalExam, TargetRole = Label`).

Question2 Prediction (2)

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 students final exam grade based on the students 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 predic-

tion of a students final exam grade based on the students MCQ1 and MCQ2 grade in the course. Compare this equation with the result of Question 2.1. Justify your answer.

Question3 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.

Question4 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.

   Note:

   - You will need to use the `Set role` operator right after your data and before using any prediction or classification operators.
   - Take screenshots of all your results and include them in your document that you submit.