RMIT University
School of Computer Science and Information Technology
COSC2110/COSC2111 Data Mining
Assignment 1

This assignment counts for 25% of the total marks in this course.

Due Date 9:00am Monday 14 April 2014

PART 1: CLASSIFICATION

20 marks

1. Classification of
   `/public/courses/DataMining/data/arff/UCI/hypothyroid.arff`.
2. Run the following classifiers, with the default parameters, on this data: ZeroR, OneR, J48, IBK and construct a table of the training and cross-validation errors. What do you conclude from these results?

Run No	Classifier	Parameters Parameters	Training Error	Cross-valid Error	Over- Fitting
1	ZeroR	None	61.5%	61.5%	None
.	.	.	.	.	

3. Using the J48 classifier, can you find a combination of the C and M parameter values that minimizes the amount of overfitting? Include the results of your best five runs, including the parameter values, in your table of results.
4. Reset J48 parameters to their default values. What is the effect of lowering the number of examples in the training set? Include your runs in your table of results.
5. Using the IBk classifier, can you find the value of k that minimizes the amount of overfitting? Include your runs in your table of results.
6. Try a number of other classifiers. Which one gives the best performance in terms of predictive accuracy and overfitting? Include your best five runs in your table of results.
7. Compare the accuracy of ZeroR, OneR and J48. What do you conclude?
8. What golden nuggets did you find, if any?
9. [OPTIONAL] Use an attribute selection algorithm to get a reduced attribute set. How does the accuracy on the reduced set compare with the accuracy on the full set.

   Submit: Up to two pages that describes what you did for each of the above questions and your results and conclusions.

PART B: NUMERIC PREDICTION 10 marks

1. Numeric Prediction of `/public/courses/DataMining/data/arff/numeric/breastTumor.arff`
2. Run the following classifers, with default parameters, on this data: ZeroR, MP5, IBk and construct a table of the training and cross-validation errors. You may want to turn on "Output Predictions" to get a better sense of the magnitude of the error on each example. What do you conclude from these results?
3. Explore different parameter settings for M5P and IBk. Which values give the best performance in terms of predictive accuracy and overfitting. Include the results of the best five runs in your table of results.
4. Investigate three other classifiers for numeric prediction and their associated parameters. Include your best five runs in your table of results. Which classifier gives the best performance in terms of predictive accuracy and overfitting?

   Submit: Up to one page that describes what you did for each of the above questions and your results and conclusions.

PART 3: CLUSTERING 10 marks

1. Clustering of `yallara:/public/courses/DataMining/data/arff/student-data.arff`

   This file contains some data for some hypothetical students studying at RMIT.
2. Run the Kmeans clustering algorithm on this data for the following values of K: 1,2,3,4,5,10,20. Analyse the resulting clusters. What do you conclude? How many clusters do you think there are in the data?
3. Choose a value of K and run the algorithm with different seeds. What is the effect of changing the seed?
4. Run the EM algorithm on this data with the default parameters and analyse the output. Give an English language description of the clusters.
5. Run the algorithm with different seeds. What is the effect of changing the seed?
6. Explore the effect of changing the standard deviation parameter. Carry out runs with values from 100 to E-10. What do you conclude?
7. Compare the use Kmeans and EM for clustering tasks. Which do you think is best? Why?
8. What golden nuggets did you find, if any?

   Submit: Up to one page that describes what you did for each of the above questions and your results and conclusions.

PART 4: ASSOCIATION FINDING 10 marks

1. The files `supermarket1.arff` and `supermarket2.arff` in the folder `/public/courses/DataMining/data/arff` contain the same details of shopping transactions represented in two different ways. You can use a text viewer to look at the files.
2. What is the difference in representations?
3. Load the file `supermarket1.arff` into weka and run the Apriori algorithm on this data. You will need to restrict the number of attributes and/or the number of examples. What significant associations can you find?
4. Explore different possibilities of the metric type and associated parameters. What do you find?
5. Load the file `supermarket2.arff` into weka and run the Apriori algorithm on this data. What do you find?
6. Explore different possibilities of the metric type and associated parameters. What do you find?
7. Try the other associators. What are the differences to Apriori?
8. What golden nuggets did you find, if any?

Submit: Up to one page that describes what you did for each of the above questions and your results and conclusions.

Submission instructions: Submit through Blackboard assessment tasks.