



Data Mining
COSC 2111/2110
Assignment 1 Practical Data Mining

	Assessment Type	This is an individual assignment, meaning that you must complete this assignment by yourself. Please submit your assignment online via “Canvas → Assignments → Assignment 1”. Clarifications/updates may be made via announcements/relevant discussion forums.
	Due Date	End of week 6, Friday 30 August 2024, 11:59pm
	Marks	40

1 Overview

This assignment is concerned with some kinds of tasks that occur in practical data mining situations. In this assignment you are asked to apply a number of algorithms to a number of data sets and write a report on your findings. Your assignment will be assessed on demonstrated understanding of concepts, algorithms, methodology, analysis of results and conclusions. Marks are awarded for meeting requirements as closely as possible (see rubrics). Please make sure your answers are labelled correctly with the corresponding part and sub-question numbers, to make it easier for the marker to follow. Please stick to the required page limits (penalty may apply).

2 Learning Outcomes

This assessment relates to the following learning outcomes of the course.

- CLO 1: Demonstrate advanced knowledge of data mining concepts and techniques.
- CLO 2: Apply the techniques of clustering, classification, association finding, feature selection and visualisation on real world data.
- CLO 4: Apply data mining software and toolkits in a range of applications.
- CLO 5: Set up a data mining process for an application, including data preparation, modelling and evaluation.
- CLO 6: Demonstrate knowledge of ethical considerations involved in data mining.

3 Assignment Details

3.1 Part 1: Classification (12 marks)

This part of the assignment is concerned with the file:

/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff/UCI/credit-g.arff.

The data was supplied by the Garavan Institute and J. Ross Quinlan, NSW, Australia. The main goal here is to achieve the highest classification accuracy with the lowest amount of overfitting.

1. 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. You can get the training error by selecting “Use training set” as the test option. What do you conclude from these results? Provide your explanation.

Run No	Classifier	Parameters	Training Error	Cross-validation Error	Overfitting
1	ZeroR	None	30.0%	30.0%	None
.

2. 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. What is your conclusion? Provide your explanation.
3. Reset J48 parameters to their default values. What is the effect of lowering the number of examples in the training set? Provide your explanation. Include your runs in your table of results.
4. Using the IBk classifier, can you find the value of k that minimizes the amount of overfitting? Provide your explanation. Include your runs in your table of results.
5. Try two other classifiers. Aside from ZeroR, which classifiers are best and worst in terms of predictive accuracy? Include 5 runs in your table of results. Provide your analysis on these results.
6. Compare the accuracy of ZeroR, OneR and J48. What do you conclude? Give your explanation on these results.
7. What “golden nuggets” did you find, if any?
8. [OPTIONAL for COSC2110] 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? Provide your explanation.

Report Length: Up to two pages, not including the table of runs.

3.2 Part 2: Numeric Prediction (8 marks)

This part of the assignment is concerned with the file:

`/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff/numeric/cholesterol.arff`.

The task is to predict the value of the “chol” attribute. The main goal is to achieve the lowest mean absolute error with the lowest amount of overfitting.

1. Run the following classifiers, 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? Give your explanation.
2. 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. Provide your explanation on these results.

- Investigate two 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? Provide your explanation.
- What golden nuggets did you find, if any?

Report Length Up to one page, not including the table of runs.

3.3 Part 3: Clustering (10 marks)

Clustering of the credit-g data of part 1. For this part use only the attributes `duration`, `age`, `credit_amount` and `job`. The aim is to determine the number of clusters in the data and assess whether any of the clusters are meaningful.

- Run the K -means 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? Provide your reasoning.
- Choose a value of K and run the algorithm with different seeds. What is the effect of changing the seed? Provide your explanation.
- Run the EM algorithm on this data with the default parameters and describe the output and your analysis.
- The EM algorithm can be quite sensitive to whether the data is normalized or not. Use the Weka normalize filter
(Preprocess --> Filter --> unsupervised --> normalize)
to normalize the numeric attributes. What difference does this make to the clustering runs? Provide your reasoning.
- The algorithm can be quite sensitive to the values of *minLogLikelihoodImprovementCV*, *minStdDev* and *minLogLikelihoodImprovementIterating*. Explore the effect of changing these values. What do you conclude?
- How many clusters do you think are in the data? Give a plain English language description of one of them.
- Compare the use of K -means and EM for these clustering tasks. Which do you think is best? Why?
- What golden nuggets did you find, if any?

Report Length Up to one page.

3.4 Part 4: Association Finding (5 marks)

Association finding in the files `groceries1.arff` and `groceries2.arff` in the folder:
/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff.

The main aim is to determine whether there are any significant associations in the data.

These files contain the same details of shopping transactions represented in two different ways. You can use a text viewer to look at the files.

- What is the difference in representations?

2. Load the file **groceries1.arff** into Weka and run the Apriori algorithm on this data. You might need to restrict the number of attributes and/or the number of examples.
 - What significant associations can you find? Provide your reasoning.
 - Explore different confidence values (note that confidence is the metric used here) and associated parameters. What do you find? Provide your explanation.
3. Load the file **groceries2.arff** into Weka and run the Apriori algorithm on this data. What do you find? Provide your explanation.
 - What significant associations can you find? Provide your reasoning.
 - Explore different confidence values (note that confidence is the metric used here) and associated parameters. What do you find? Provide your explanation.
4. Try the other associators. What are the differences in results that you find as opposed to those of Apriori?
5. What golden nuggets did you find, if any?
6. [OPTIONAL for COSC2110] Can you find any meaningful associations in the **credit-g** data?

Report Length Up to one page.

3.5 Part 5: Ethical issues in Data Mining (5 marks)

In this task, you will need to provide a recorded video presentation (3 to 5 minutes, with no more than 5 presentation slides). The topic can be an ethical or legal issue involving data mining. Your talk should address the following two criteria:

- Clearly articulate an ethical or legal issue in a real-world data mining application scenario;
- Make suggestions on how to handle the abovementioned issue.

You will need to record the presentation in MP4 format. Both the recorded video presentation and the presentation slides (PDF format) should be submitted via Canvas.

4 Submission Instructions

You need to submit the **following 3 files** via Canvas:

- one PDF file for the report covering Part 1 - Part 4.
- one MP4 video file in Part 5.
- one PDF file for your presentation slides in Part 5.

4.1 Late submission penalty

After the due date, you will have 5 business days to submit your assignment as a late submission. Late submissions will incur a penalty of 10% per day. After these five days, Canvas will be closed and you will lose ALL the assignment marks.

Assessment declaration:

When you submit work electronically, you agree to the assessment declaration - <https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/assessment-declaration>

5 Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to the following: <https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity>.

6 Marking guidelines

Factors contributing to the final mark will include the number of tasks attempted, the amount of exploration and demonstrated understanding of the algorithms, methodology, logical analysis, presentation of results and conclusions (see the marking rubrics on Canvas).