

THE AUSTRALIAN NATIONAL UNIVERSITY
First Semester Examination – June 2007

COMP8400
Algorithms and Techniques for Data Mining

Study Period: 15 minutes

Time Allowed: 3 hours

*Permitted Materials: One A4 page with notes on both sides.
NO calculator permitted.*

Questions are NOT equally weighted.

The questions are followed by labelled, framed blank panels into which your answers are to be written. Additional answer panels are provided (at the end of the paper) should you wish to use more space for an answer than is provided in the associated labelled panels. If you use an additional panel, be sure to indicate clearly the question and part to which it is linked.

The marking scheme will put a high value on clarity so, as a general guide, it is better to give fewer answers in a clear manner than to outline a greater number in a sketchy, half-answered fashion.

Please write clearly – if we cannot read your writing you might lose marks!

Name (family name first):

Student Number:

The following are for use by the examiners.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Total
----	----	----	----	----	----	----	----	----	-------

Question 1 [7 marks] The data mining process, and data issues in data mining

(a) (i) Illustrate the data mining process.

(i)



[1 mark]

(ii) Provide percentage estimates for time and effort spent in each of the steps in the data mining process.

(ii)



[1 mark]

(b) List five (5) major data mining challenges.



[1 mark]

Question 1 (continued)

(c) Define data mining in your own words in one or two sentences.

[1 mark]

(d) (i) Describe in one or two sentences what a data warehouse is.

(i)

[1 mark]

(ii) Explain how a data warehouse is different from a relational database, and give one example application each where you would use one or the other.

(ii)

[1 mark]

(iii) Explain how a data warehouse can be useful for data mining.

(iii)

[1 mark]

Question 2 [7 marks] Data pre-processing, and data integration and linkage

- (a) What is meant by *real world data is dirty*? Explain in one sentence and give an example of why this is an important issue.

[1 mark]

- (b) In one sentence each, list and describe four (4) different data quality measures.

[1 mark]

- (c) In one sentence each, list and describe four (4) different methods how missing values can be handled.

[1 mark]

Question 2 (continued)

(d) (i) In one or two sentences, explain what dimensionality reduction is and what it is used for.

(i)

[1 mark]

(ii) In one sentence each, describe two approaches to dimensionality reduction.

(ii)

[1 mark]

(e) (i) Why are data integration and data linkage becoming increasingly important?

(i)

[1 mark]

(ii) In one sentence each, describe the two main data linkage techniques.

(ii)

[1 mark]

Question 3 [5 marks] Mining frequent patterns and associations

- (a) Assume you have the following small data set with 7 transactions containing items sets made of items **a** to **e**.

TID	Items
1	c, d, e
2	a, c, e
3	b, d
4	a, c, e
5	a, b, d, e
6	b, d, e
7	a, c, e

- (i) Following the Apriori algorithm, give all frequent and all candidate item sets of lengths 1, 2 and 3 with a minimum support of 2 transactions.

(i)

[2 marks]

Question 3 (continued)

(ii) For all frequent item sets containing three (3) items from part (i) of this question, generate all possible rules, and give their support and confidence (as ratios or percentage numbers).

(ii)

[1 mark]

(iii) What is the main bottleneck of the Apriori algorithm? Describe in one or two sentences.

(iii)

[1 mark]

(b) What is multi-dimensional association mining? Describe in one or two sentences and give an example.

[1 mark]

Question 4 [5 marks] Cluster analysis

- (a) A good clustering will produce clusters with two (2) characteristics. Describe these characteristics in one sentence each.

[1 mark]

- (b) List three (3) drawbacks of the k-means clustering algorithm.

[1 mark]

- (c) List and describe five (5) different clustering requirements in data mining.

[1 mark]

Question 4 (continued)

(d) What is a dendrogram? Explain in one or two sentences, and describe how it is used in clustering.

[1 mark]

(e) What is constraint based clustering? Explain in one or two sentences, and give an example application where constraint based clustering can be useful.

[1 mark]

Question 5 [7 marks] Classification and prediction

- (a) The following table contains a small training data set with 10 records, two attributes with two values each (**0** and **1**), and the class label attribute (with two classes **Y** and **N**).

RecID	Attr1	Attr2	Class
1	0	0	Y
2	1	0	N
3	0	1	Y
4	0	0	Y
5	1	1	N
6	1	1	N
7	1	0	Y
8	0	1	Y
9	0	0	Y
10	0	1	Y

- (i) Build a decision tree based on the training data given in the table. At each step, use the attribute which results in the purest splitting of the data (i.e. results in sub sets of the data with all – or most – of the records being in one class). Show your workings.

(i)

[2 marks]

- (ii) What is the accuracy (as percentage) of your decision tree classifier from part (i) of this question on the given training data?

(ii)

[1 mark]

Question 5 (continued)

(iii) What is overfitting, and how can it be prevented for decision trees?

(iii)

[1 mark]

(b) In one sentence each, describe one advantage and one disadvantage of the naïve Bayesian classifier.

[1 mark]

(c) In “*Classifier technology and the illusion of progress*”, David Hand discusses several important issues that need to be considered when dealing with classifiers. Summarise two (2) of these issues.

[1 mark]

(d) How is prediction different from classification? Explain in one or two sentences.

[1 mark]

Question 6 [4 marks] Mining data streams and time series

(a) (i) Compared to static databases, what are the main limiting factors when mining data streams?

(i)

[1 mark]

(ii) List four (4) application areas where data stream mining techniques can be used.

(ii)

[1 mark]

(b) (i) What is a main characteristic of time series data?

(i)

[1 mark]

(ii) List the four (4) main categories of time-series movements.

(ii)

[1 mark]

Question 7 [4 marks] Privacy-preserving data mining

(a) What is the goal of privacy-preserving data mining?

[1 mark]

(b) Explain in one or two sentences what is meant by re-identification.

[1 mark]

(c) What is the aim of secure multi-party computation? Explain and give a simple example.

[1 mark]

(d) Give two (2) approaches to data perturbation and obfuscation.

[1 mark]

Question 8 [7 marks] Web and text data mining

- (a) A stop word list is commonly used in information retrieval and text data mining. Explain in one or two sentences what a stop word list is, and give an example.

[1 mark]

- (b) Explain what TF-IDF stands for, what it is, and what it is used for.

[1 mark]

- (c) Explain in one or two sentences how keyword based association analysis on a set of text documents works.

[1 mark]

- (d) Give two (2) applications where text classification can be being applied.

[1 mark]

Question 8 (continued)

(e) Describe a main challenge when using names for text data mining.

[1 mark]

(f) (i) In one sentence each, list and describe four (4) main challenges that the Web poses for data mining.

(i)

[1 mark]

(ii) In one sentence each, list and describe four (4) different types of Web data mining.

(ii)

[1 mark]

Question 9 [4 marks] End-to-end data mining, social impacts and data mining trends

(a) What are ensemble methods in supervised classification? Explain in one or two sentences.

[1 mark]

(b) Describe in one or two sentences how random forest works.

[1 mark]

(c) Describe one area where data mining has made a social impact.

[1 mark]

(d) List four (4) trends in data mining.

[1 mark]

Student Number:

Continuation of answer to Question Part

Continuation of answer to Question Part

Continuation of answer to Question Part

Student Number:

Continuation of answer to Question Part

Continuation of answer to Question Part

Continuation of answer to Question Part