

THE AUSTRALIAN NATIONAL UNIVERSITY
First Semester Examination – June 2008

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 this examination 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. As a general guide, it is therefore 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 I 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
----	----	----	----	----	----	----	----	----	-------

Student Number:

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

(a) Data mining has many challenges. With one sentence each, describe four (4) of them.

[1 mark]

(b) (i) Draw a diagram that illustrates the data mining process. Your diagram should clearly show all the steps of the data mining process.

(i)

[1 mark]

(ii) Describe in which steps of the data mining process most of the time and efforts are often spent. Can you explain why this is so?

(ii)

[1 mark]

Question 1 (continued)

(c) Data mining is multi-disciplinary. List four (4) disciplines commonly used in data mining.

[1 mark]

(d) In one sentence each, list and describe four (4) different sources of data that are commonly used in data mining.

[1 mark]

(e) What is the main purpose of a data warehouse? Describe in one or two sentences.

[1 mark]

(f) Explain what the main difference between a *star* and a *snowflake* schema is.

[1 mark]

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

(a) List four (4) root conditions of data quality problems.

[1 mark]

(b) In one sentence each, describe the four (4) data quality measures *accuracy*, *completeness*, *consistency* and *timeliness*.

[1 mark]

(c) What are the four (4) main data pre-processing tasks?

[1 mark]

Question 2 (continued)

- (d) Explain why it is often not a good idea to replace missing values in a data set with a constant such as *'unknown'*, *'NA'*, *'n/a'* or *'missing'*.

[1 mark]

- (e) What is feature construction, and why can it be useful? Explain in one or two sentences, and give an example.

[1 mark]

- (f) What is data parsing and standardisation, and why is it important? Explain and give an example.

[1 mark]

- (g) Why do duplicate records appear in databases, and why can this become a problem? Explain in one or two sentences.

[1 mark]

Student Number:

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

(i) Following the *Apriori* algorithm, give all candidate and all frequent 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 of length three (3) from part (i) of this question (previous page), generate all rules with two items on the left-hand side and one item on the right-hand side (such as $\{a,b\} \rightarrow c$), and calculate their support and confidence (as ratios or percentage numbers).

(ii)

[1 mark]

(b) One of the main bottlenecks in the *Apriori* algorithm is that it requires many database scans. Describe two ways of how this can be improved.

[1 mark]

(c) Explain what a *redundant* association rule is, and how it can be generated. Also give an example.

[1 mark]

Question 4 [5 marks] Cluster analysis

- (a) There are many different clustering approaches. List four (4) of them, and describe with one sentence each the basic ideas of how they work.

[1 mark]

- (b) What are the main characteristics of a good clustering?

[1 mark]

- (c) What are the main advantages of k-means clustering compared to other clustering algorithms.

[1 mark]

Question 4 (continued)

- (d) Describe and illustrate (with a figure) three (3) different methods of how the distance between two clusters can be measured. What are these three methods called?

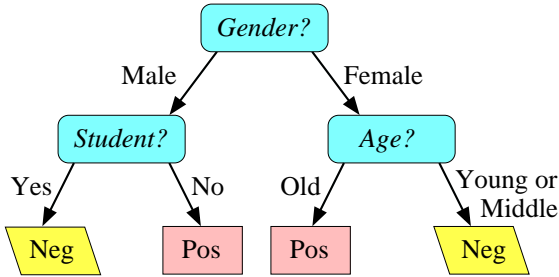
[1 mark]

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

[1 mark]

Question 5 [7 marks] Classification and prediction

(a) Below you can see a decision tree that has been constructed using a training data set (unknown to us). Also shown is a *test* data set with ten (10) test records. The attribute (variable) *Test* is the class (target) attribute with classes *Pos* and *Neg*.



Student	Age	Gender	Test
No	Young	Female	Pos
Yes	Young	Male	Neg
No	Old	Female	Pos
No	Young	Male	Neg
No	Middle	Female	Neg
No	Middle	Male	Pos
Yes	Old	Female	Pos
Yes	Old	Male	Pos
Yes	Young	Female	Neg
Yes	Middle	Female	Neg

(i) Write down the rules that can be generated from the above decision tree.

(i)

[1 mark]

(ii) For each of the ten (10) test records, determine if it is a *true positive*, a *true negative*, a *false positive* or a *false negative*. Then write down the resulting confusion matrix with the absolute counts (i.e. number of records) in each entry of the matrix.

(ii)

[1 mark]

(iii) What are the accuracy and the misclassification rate (both as percentage values or ratios) of the above decision tree on the above test data set?

(iii)

[1 mark]

Question 5 (continued)

(b) Explain why it is important to prune decision trees.

[1 mark]

(c) List three (3) advantages of support vector machines over artificial neural networks.

[1 mark]

(d) In his paper “*Comparing classifiers*”, Salzberg writes about the process of repeated *tuning*. Describe what this process is, and why it can be problematic.

[1 mark]

(e) Describe in one or two sentences how the prediction method of *linear regression* works.

[1 mark]

Student Number:

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

(a) List four (4) characteristics of data streams.

[1 mark]

(b) In one sentence each, list and describe four (4) major methods of data stream processing.

[1 mark]

(c) Describe how the *moving average* method works, as used for example for estimating the trend curve of a time series.

[1 mark]

(d) What is the process of de-seasonalising data? Explain and give an example where this is important.

[1 mark]

Question 7 [4 marks] Privacy-preserving data mining

- (a) Explain how re-identification is possible even when personal identifiers are removed from a data set.

[1 mark]

- (b) Explain what is meant by *horizontally* and *vertically* distributed data, and give one example each where such distributions of data can appear.

[1 mark]

- (c) What is *k*-anonymity? Describe in one sentence the basic principle behind it.

[1 mark]

- (d) Why is privacy-preserving data sharing important? Explain and describe an example application.

[1 mark]

Question 8 [8 marks] Web and text data mining

(a) What are the main differences between database systems and information retrieval systems?

[1 mark]

(b) Explain how the vector space model for document similarity calculation works.

[1 mark]

(c) Explain what word stemming is, and why it is important. Also give an example of word stemming.

[1 mark]

(d) Manual text categorisation is normally very labor-intensive and therefore does not scale up. Can you describe an alternative approach to manual text categorisation? How does it work?

[1 mark]

Question 8 (continued)

(e) (i) What is the aim of Web page layout structure mining?

(i)

[1 mark]

(ii) What are the challenges when mining the Web page layout structure?

(ii)

[1 mark]

(f) Describe in one or two sentences an example application of mining the Web's link structure.

[1 mark]

(g) According to a recent KDnuggets poll, Web usage mining is the most commonly used Web mining technique. Describe how this technique works, and give an example for what it can be used.

[1 mark]

Student Number:

Question 9 [3 marks] End-to-end data mining, social impacts and data mining trends

(a) Give two (2) examples of ensemble classification methods, and describe how they work.

[1 mark]

(b) Describe one example where data mining has changed most people's life.

[1 mark]

(c) Describe an area where data mining is likely to have an impact in the future.

[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