Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Engineering

Date: Friday 28th January 2011
Time: 09:45 - 11:45

Please answer ONE Question from Section A (worth 30 marks)
and ONE Question from Section B (worth 30 marks)

For full marks your answers should be concise as well as accurate.
Marks will be awarded for reasoning and method as well as being correct.

This is an OPEN book examination

The use of electronic calculators is permitted provided they are
not programmable and do not store text.
1. A1

   a) An organisation’s data lifecycle management plan defines a management agreement on how the organisation creates, organises and resources a strategy for managing its data lifecycle. Describe two of the main issues involved in the development of such a plan. (5 marks)

   b) What are the two main reasons behind the phenomenon of data retention in today’s business world? Explain. (7 marks)

   c) The interactive ease of networks, including the Internet, means that once access to specific data is acquired, the data can be transmitted to countless destinations around the globe in a matter of seconds.

      Based on this scenario, provide an example (fictitious or not) of a case involving the un-consented dissemination of personal and private data in a business environment, emphasising any legal ethical or privacy issues that may relate to the case. (8 marks)

   d) Provide an example and discuss the main advantages and the main disadvantages of the use of the Web as a means for disseminating data. (10 marks)
2. A2

a) Describe three dimensions of data quality and provide a data example for each dimension. (6 marks)

b) Describe the differences between data-driven and process-driven data quality improvement strategies and provide an example of a technique for each. (6 marks)

c) Data lifecycle management solutions enable IT managers to assemble the appropriate combination of storage devices, media types, and network infrastructure to create a proper balance of performance, data accessibility, retrieval and reliability based on the relative value of data at different phases of its lifecycle. Provide an example of a company and data lifecycle management solution that can help a company keep high data accessibility and retrieval performance for its clients. (8 marks)

d) Discuss the issue of data security in today's complex information ecosystem. Provide examples to support your views. (10 marks)
Section B

1. B1

a) Data Warehousing.

A data warehouse for a University consists of the following dimensions, each with its hierarchy as shown:

- student < major < university < all
- course < theme < PG-course < FacultyCourse < all
  where one of the PG-course attributes is "School", and its values include "Computer Science"
- semester < year < all
- instructor < all

and measures count and grade.

i) draw either a snowflake or a star schema to represent the data model for this warehouse, and justify your design. (2 marks)

ii) calculate the total number of cuboids in this data cube (including the base and apex cuboids) (3 marks)

iii) describe the following query in terms of primitive OLAP operations (drill-down, roll-up, slice/dice, pivot):
  "list the average grade of each CS major across all PG-course for Computer Science, for this semester" (3 marks)

b) Two classifiers designed to predict patients’ susceptibility to allergy are being designed and tested, independently of one another. Each of the classifiers predict that a patient is either positive (allergic) or negative (normal) based on a combination of observable factors. The tests result in the following two confusion matrices, one for each classifier:

Matrix A:

<table>
<thead>
<tr>
<th>actual</th>
<th>allergic</th>
<th>normal</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>allergic</td>
<td>30</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td>normal</td>
<td>20</td>
<td>500</td>
<td>520</td>
</tr>
<tr>
<td>total</td>
<td>50</td>
<td>570</td>
<td>620</td>
</tr>
</tbody>
</table>

Matrix B:

<table>
<thead>
<tr>
<th>actual</th>
<th>allergic</th>
<th>normal</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>allergic</td>
<td>70</td>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td>normal</td>
<td>200</td>
<td>320</td>
<td>520</td>
</tr>
<tr>
<td>total</td>
<td>270</td>
<td>350</td>
<td>620</td>
</tr>
</tbody>
</table>

i) Calculate the precision, recall, and F-measure for each of the classifiers based on these tables. (4 marks)

ii) Based on their values, can you recommend one classifier over the other, considering the type of application they are used for? Justify your answer (2 marks)
c) A new classifier that can associate probabilities to class assignments is being tested, and it has produced the following list of class assignments:

<table>
<thead>
<tr>
<th>instance</th>
<th>class probability distribution</th>
<th>predicted</th>
<th>error?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1 0.55 0.35</td>
<td>c2</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>0.2 0.7 0.1</td>
<td>c2</td>
<td>N</td>
</tr>
</tbody>
</table>

Quadratic and information loss functions are two ways to assess the prediction accuracy of this classifier.

i) Calculate both of them for this list (the list is partial but, for the sake of the exercise, we may assume it is complete). (3 marks)

ii) Briefly discuss why you would use one of them rather than the other to report on the classifier accuracy. (3 marks)

d) Suppose you are in the market to purchase a data mining system. Regarding the coupling of a data mining system and/or a warehouse system:

i) What are the differences between no coupling, loose coupling, semitight, and tight coupling? (3 marks)

ii) What is the difference between row scalability and column scalability, and why is column scalability difficult to achieve? (3 marks)

iii) Which features, amongst those listed above, would you look for when selecting a data mining system? Discuss specifically desirable data visualization capabilities. (4 marks)

Justify all of your answers.
2. B2

a) Data Warehousing.

Suppose a warehouse model consists of three dimensions:
- time, with hierarchy: day < month < year,
- doctor, and
- patient

and two measures: count (how many times a patient has been seen by a doctor), and charge (i.e., the fee charged by a doctor for a visit).

Consider the query $Q: \text{sum}(\text{charge}) \text{ where month} = \text{May group by (doctor, patient)}$.

i) List all the cuboids in this data cube \hspace{1cm} (2 marks)

ii) Describe the query $Q$ in terms of primitive OLAP operations (drill-down, roll-up, slice/dice, pivot) \hspace{1cm} (3 marks)

iii) Suppose you can afford to materialise only one of the following cuboids from (A):
- (year, doctor)
- (doctor, patient)
- (month = June, patient)

Which one would you choose? Justify your answer. (3 marks)

b) Suppose a shop sells a number of products and each transaction record consists of a set of such products. The goal is to discover strong rules that may help the shop managers to better place their products on the shelves. The following contingency shows a breakdown of transactions for two products, A and B.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>not B</th>
<th>sum(row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7000</td>
<td>500</td>
<td>7,500</td>
</tr>
<tr>
<td>not A</td>
<td>500</td>
<td>2000</td>
<td>2,500</td>
</tr>
<tr>
<td>sum(col)</td>
<td>7500</td>
<td>2500</td>
<td>10000</td>
</tr>
</tbody>
</table>

i) Use the table to determine whether attributes A and B are negatively correlated, positively correlated, or independent. (4 marks)

ii) If you were the shop manager, what would you conclude regarding the joint sales or products A and B? (2 marks)
c) Given the following dataset:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>x</td>
<td>c1</td>
</tr>
<tr>
<td>b</td>
<td>x</td>
<td>c1</td>
</tr>
<tr>
<td>a</td>
<td>x</td>
<td>c2</td>
</tr>
<tr>
<td>a</td>
<td>y</td>
<td>c2</td>
</tr>
<tr>
<td>b</td>
<td>y</td>
<td>c2</td>
</tr>
</tbody>
</table>

use the *information gain* metric to determine which of the two attributes is the better one to split on when constructing a decision tree. (6 marks)

d) Intrusion detection is an area of active research in data mining, with important applications for computer security. Association rules, classification, and other approaches have been used to detect intrusions in computer systems.

i) Briefly describe the problem, trying to cast it in terms of pattern detection. (2 marks)

ii) Are techniques such as frequent pattern mining likely to be effective to address the problem? Please provide an example to support your answer. (2 marks)

iii) Can you think of specific characteristics of the data to be analysed that make it particularly challenging? Explain your answer. (2 marks)