Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Engineering

Date: Tuesday 17th January 2012
Time: 09:45 - 11:45

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

Use a SEPARATE answer book for each SECTION

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.
Section A

1.  
   a) Explain the evolution of information systems into today’s complex information “ecosystems” and its consequences.  
      (5 marks)

   b) Discuss the reasons behind the phenomenon of data retention, its disadvantages, and how data retention can be overcome.  
      (10 marks)

   c) Discuss Web Services as a means for processing data. In your discussion, mention at least two issues that would:
      i. Prevent you from using Web Services
      ii. Drive you to use Web Services in a business context.  
      (11 marks)

   d) Define the term ‘data quality methodology’ and describe the three common phases of a data quality methodology.  
      (4 marks)
2. 

a) Explain the following: “Nowadays, Business deals with ‘data’, not with ‘documents’.”  

(6 marks)

b) Give an example of a company, describing its business. Provide a data lifecycle management solution that can help the company keep high data accessibility and performance without increase in costs.  

(8 marks)

c) Enumerate and explain three advantages and three disadvantages of the Web-DBMS integration approach, providing examples.  

(11 marks)

d) Explain the reason why most data quality methodologies focus on structured and semi-structured data.  

(5 marks)
Section B

3.

a) Suppose a market shopping data warehouse consists of four dimensions: customer, date, product, and store, and two measures: count, and avg sales, where avg sales stores the real sales in pounds at the lowest level but the corresponding average sales at other levels.

(i) Draw a snowflake schema diagram (you do not have to mark every possible level, but make clear your implicit assumptions on the levels of a dimension).

(3 marks)

(ii) Starting with the base cuboid [customer, date, product, store], what specific OLAP operations (e.g., roll-up student to department (level)) should be performed in order to list the average sales of each cosmetic product since January 2005?

(4 marks)

(iii) If each dimension has 5 levels (excluding all), such as store-city-state-region-country, how many cuboids does this cube contain (including base and apex cuboids)?

(2 marks)

b) (i) 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></th>
<th>predicted</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>allergic</td>
<td>normal</td>
<td>total</td>
</tr>
<tr>
<td>actual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allergic</td>
<td>20</td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>normal</td>
<td>30</td>
<td>490</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></th>
<th>predicted</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>allergic</td>
<td>normal</td>
<td>total</td>
</tr>
<tr>
<td>actual</td>
<td>allergic</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>normal</td>
<td>190</td>
<td>330</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>270</td>
<td>350</td>
</tr>
</tbody>
</table>

Calculate the accuracy, recall, and F-measure for each of the classifiers based on these tables. Based on their values, can you recommend one classifier over the other, considering the type of application they are used for? Justify your answer.

(ii) A new classifier that can associate probabilities to class assignments is being tested, and it has produced the following table 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></td>
<td>c1</td>
<td>c2</td>
<td>c3</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.55</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.7</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Quadratic and information loss functions are two ways to assess the prediction accuracy of this classifier. Calculate both of them for this table (the table is partial but, for the sake of the exercise, we may assume it is complete), and briefly discuss why you would use one of them rather than the other to report on the classifier accuracy.

(c) Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining and anomaly detection can be applied. Give examples to support your answer.
4.  

a) Consider the problem of finding the \( K \) nearest neighbours of a data object. A programmer designs the following algorithm for this task.

1: \textbf{for} \( i = 1 \) to \textit{number of data objects} \textbf{do}  
2: \hspace{1em} Find the distances of the \textit{i}th object to all other objects.  
3: \hspace{1em} Sort these distances in decreasing order.  
   \hspace{1em} (Keep track of which object is associated with each distance.)  
4: \hspace{1em} \textbf{return} the objects associated with the first \( K \) distances of the sorted list  
5: \textbf{end for}  

Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same. How would you address this problem? Make clear any assumptions you make.  

(4 marks)

b) 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>6500</td>
<td>1000</td>
<td>7500</td>
</tr>
<tr>
<td>not A</td>
<td>1000</td>
<td>1500</td>
<td>2500</td>
</tr>
<tr>
<td>sum(col)</td>
<td>7500</td>
<td>2500</td>
<td>10000</td>
</tr>
</tbody>
</table>

Use the table and the \textit{lift} metric to determine whether attributes A and B are negatively correlated, positively correlated, or independent. If you were the shop manager, what would you conclude regarding the joint sales or products A and B?  

(5 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. Show your working. (6 marks)

d) Both decision-tree induction and associative classification may generate rules for classification. What are their major differences? Why is it that in many cases an associative induction may lead to better accuracy in prediction? (5 marks)

e) Distinguish between noise and outliers. In your answer address the following questions (justify your answers):
- Is noise ever interesting or desirable?
- Can noise objects be outliers?
- Are noise objects always outliers?
- Are outliers always noise objects?
- Can noise make a typical value into an unusual one, or vice versa? (10 marks)