Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Engineering

Date: Friday 25th January 2013
Time: 09:45 - 11:45

Please answer ONE Question from Section A and ONE Question from Section B

Use a SEPARATE answer book for each SECTION

This is an OPEN book examination

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

[PTO]
Section A

1. 
   a) In the complex information ecosystem of nowadays, how can we know who is changing which records? In your answer, discuss the causes of this emerging complexity, referring to the Data Lifecycle Management function of security in your answer, discuss the difficulties in guaranteeing Security in such an environment, and suggest any security mechanisms to overcome those difficulties, emphasising the functionality supported by each suggested mechanism.

   (10 marks)

   b) Consider an insurance company and its data sets, one of which is the company’s client forms, stored on-line. In compliance with government regulations, the company retains its client forms for at least two years.

   Suggest a Data Lifecycle Management solution that can help this insurance company keep high data accessibility and performance without increase in costs.

   (6 marks)

   c) Discuss how the evolution of information systems has changed the way in which business is carried out.

   (6 marks)

   d) In the complex information ecosystem of nowadays, how do we guarantee that data is accessible to our clients at all times? In your answer, refer to the Data Lifecycle Management function of recovery, including in your discussion aspects of volatility and speed of on and off-line storage, as well as data replication and its complications.

   (8 marks)
2.
   a) Web Services allow platform-independent communication between systems over a network. They have changed the face of business and commerce, by allowing communication between organisations and clients in an automatic and effective way, where one can provide services to another via a Web application. These services can include software and hardware services. When software services are provided, then a client can give access to his/her data for processing to the Web application.

   Discuss Web Services as a means for processing and exchanging data in modern information systems. In your discussion, mention at least two issues that would (i) prevent you from using Web services, and two issues that would (ii) drive you to use Web services in a business context. The issues should relate to basic data lifecycle management functions, such as Security, and should also relate to cost, responsibility and privacy.

   (8 marks)

   b) Assume that you are in charge of developing Data Lifecycle Management (DLM) procedures for Facebook Inc., the American multinational Internet Corporation which runs the social networking website Facebook. Give three examples of DLM procedures for Facebook, justifying your answer and stating any assumptions.

   (8 marks)

   c) Explain the reason why most data quality methodologies focus on structured and semi-structured data, illustrating your argument with an example.

   (4 marks)

   d) In the complex information ecosystem of nowadays, how do we guarantee that data is of high quality at all times? Refer to the Data Lifecycle Management function of Data Quality (DQ) management in your answer, including in your discussion aspects of the relevance of DQ, the difficulties of managing DQ in the information ecosystem of nowadays, emphasising the characteristics of the ecosystem, and the role of DQ methodologies in overcoming such difficulties.

   (10 marks)
Section B

3. a) Suppose a market shopping data warehouse consists of six dimensions: customer, time, product, from_location, to_location and shipper, and the following measures: quantity, weight, cost, and charge, where cost is the charge of shipping and charge is the shipping fee charged to the customer.

(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). How would a star schema differ? Briefly compare the advantages of each type of scheme. (4 marks)

(ii) Starting with the base cuboid, what specific OLAP operations (e.g., roll-up which dimension from which level to which level) should be performed in order to find the average monthly profit for shipping each brand of TV from Manchester to London since March 2011. (4 marks)

b) Consider the following data set for a binary class problem:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
<td>+</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>+</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>+</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>+</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>-</td>
</tr>
</tbody>
</table>

(i) Calculate information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose? (4 marks)

(ii) Compare and contrast information gain and gain ratio; in which circumstances would you use gain ratio? (2 marks)
c) The following contingency table shows a breakdown of transactions for tea and coffee drinkers:

<table>
<thead>
<tr>
<th></th>
<th>coffee</th>
<th>not coffee</th>
<th>sum(row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tea</td>
<td>150</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>not tea</td>
<td>650</td>
<td>150</td>
<td>800</td>
</tr>
<tr>
<td>sum(col)</td>
<td>800</td>
<td>200</td>
<td>1000</td>
</tr>
</tbody>
</table>

(i) Calculate support, confidence and lift for the association rule \{tea\} -> \{coffee\}.  
(3 marks)

(ii) Discuss the correlation between tea drinkers and coffee drinkers.  
(2 marks)

d) Explain how unsupervised clustering can be used to evaluate the likely success of a supervised learning model, and how supervised learning can be used to help evaluate the results of an unsupervised clustering.  
(3 marks)

e) 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.  
(8 marks)
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: **for** $i = 1$ to number of data objects **do**
2: Find the distances of the $i$th object to all other objects.
3: Sort these distances in decreasing order. (Keep track of which object is associated with each distance.)
4: **return** the objects associated with the first $K$ distances of the sorted list
5: **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.

(5 marks)

b) Using the Apriori algorithm, suppose that $L_3$ (the frequent 3-itemsets) is the list:

\{
\{a,b,c\}, \{a,b,d\}, \{a,c,d\}, \{b,c,d\}, \{b,c,w\}, \{b,c,x\}, \{p,q,r\}, \{p,q,s\}, \{p,q,t\},
\{p,r,s\}, \{q,r,s\}\}.

(i). At the join step of the algorithm, which itemsets are placed in $C_4$ (the candidate set)?

(5 marks)

(ii). Which itemsets are discarded by the prune step of the algorithm?

(4 marks)

Provide and explain your working.
c) 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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>allergic</td>
<td>normal</td>
</tr>
<tr>
<td>actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allergic</td>
<td>35</td>
<td>65</td>
</tr>
<tr>
<td>normal</td>
<td>20</td>
<td>500</td>
</tr>
<tr>
<td>total</td>
<td>55</td>
<td>565</td>
</tr>
</tbody>
</table>

Matrix B:

<table>
<thead>
<tr>
<th></th>
<th>predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>allergic</td>
<td>normal</td>
</tr>
<tr>
<td>actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allergic</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>normal</td>
<td>200</td>
<td>320</td>
</tr>
<tr>
<td>total</td>
<td>275</td>
<td>345</td>
</tr>
</tbody>
</table>

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

(5 marks)

d) Cross validation can be useful in both classification and clustering. What are the differences in each case?

(3 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?

(8 marks)