UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2006

COMPUTING AND INFORMATION SYSTEMS

IS53023A (CIS338)
Advanced Database Technologies – Data warehousing and Data mining

Duration: 2 hours 15 minutes

There are five questions in this paper. You should answer no more than THREE questions. Full marks will be awarded for complete answers to a total of THREE questions. Each question carries 25 marks. The marks for each part of a question are indicated at the end of the part in [.] brackets.

There are 75 marks available on this paper.

Electronic calculators must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Question 1

Suppose that a credit card company builds a dataset storing relevant information about its customers. This dataset is used in a particular application in which the problem to be solved, using a data mining strategy, is to characterize and distinguish the customers who have a credit card insurance from those who do not. Thus the dataset is divided in two classes each of which having 200 instances (the instances in one class contain information about individuals who currently have credit card insurance; the instances in the second class include information about individuals who have at least one credit card but without credit card insurance).

a) Which data mining strategy is most appropriate to be applied in order to form the two classes? Justify your answer. **[4]**

b) Propose a set of attributes for the dataset, giving examples of attributes among these that are useful for the mentioned problem. State if they are input or output attributes. **[4]**

c) Assume that the following rule is obtained in the data mining session differentiating the classes mentioned above:

IF Life Insurance=Yes AND Income > £30,000
THEN Credit Card Insurance=Yes
Accuracy = 80%
Coverage = 40%

i. Explain in one sentence what Accuracy = 80% means in this case. **[2]**

ii. Explain in one sentence what Coverage = 40% means in this case. **[2]**

iii. How many individuals represented by the instances in the class of credit card insurance holders have life insurance and make more than £30,000 per year? Justify your answer. **[3]**

iv. How many instances representing individuals who do not have credit card insurance have life insurance and make more than £30,000 per year? Justify your answer. **[6]**

v. If the condition ***Income > £30,000*** is replaced by ***Income >£35,000*** in the above rule, decide if the coverage of the new rule increases or decreases. In the same way, decide if the accuracy increases or decreases. Justify your answers. **[4]**

Question 2

a) Explain the difference between each of the following pairs of terms. In addition you may use examples to illustrate the differences.

i. Domain resemblance and class resemblance **[2]**
ii. Class predictiveness and class predictability **[3]**
iii. Domain predictability and class predictability **[3]**
iv. Typicality and class resemblance **[3]**
v. Within-class and between-class measure. **[3]**

b) Concept class C1 shows the following information for the categorical attribute *colour*. Use this information and the information in the table below, related to the class C1, to answer the following questions:

Name	**Value**	**Frequency**	**Predictability**	**Predictiveness**
colour	red	30		0.4
	green	20		1.0

i. What percent of the instances in class C1 have a value of *green* for the *colour* attribute? **[1]**
ii. Suppose that exactly one other concept class, C2, exists. In addition, assume all domain instances have a colour value of either *red* or *green*. Given the information in the table, determine the predictability score of *colour = red* for class C2. Justify your answer. **[2]**
iii. Using the same assumption as in point (ii) above, determine the predictiveness score of *colour = red* for class C2. Justify your answer. **[2]**
iv. Using the same assumption as in point (ii) above, determine the number of instances that reside in class C2. Justify your answer. **[3]**

c) Define the grade of cluster quality in the context of an unsupervised clustering session. What conditions the cluster quality grades should satisfy in order to the clustering to be acceptable? **[3]**

Question 3

a) Define the term *production rule*, taking into account the grades that accompany a production rule. Provide an example. **[3]**

b) Define the term *association rule*, taking into account the grades that accompany an association rule. Provide an example. **[3]**

c) Consider the set of five transactions displayed below:

Transaction 1: {bread, meat}
Transaction 2: {bread, beer}
Transaction 3: {beer, meat, potatoes}
Transaction 4: {beer, potatoes}
Transaction 5: {beer, eggs, potatoes}

i. Using a minimum support of 20%, apply the Apriori algorithm on the set of transactions above, by indicating the candidate and frequent itemsets in each phase. (*Note*: the candidate itemsets are obtained by joining the frequent itemsets from the previous phase of the algorithm.) **[11]**
ii. Generate all the association rules having a minimum confidence of 50% that can be built after the application of the Apriori algorithm. **[8]**

Question 4

a) Explain the difference between each of the following pairs of terms in the context of data warehouses:

i. Independent data mart and dependent data mart **[2]**
ii. Fact table and dimension table **[2]**
iii. Slice and dice **[2]**
iv. Drill-down and roll-up **[2]**
v. OLTP and OLAP. **[2]**

b) Consider the dataset given in the table below with the numeric attributes A and B, that has to be divided in two clusters through an unsupervised clustering session. In order to perform this, use the K-Means algorithm assuming that the first centres are the instances 1 and 4. Perform only two iterations, and then state if further iterations are needed in order to get the final clusters. **[10]**

INSTANCE NUMBER	A	B
1	3.0	3.0
2	3.0	6.0
3	4.0	5.0
4	4.0	3.0
5	5.0	4.0
6	7.0	7.5

c) Describe (in no more than ten sentences) the unsupervised clustering with self-organizing maps (known also as Kohonen networks). **[5]**

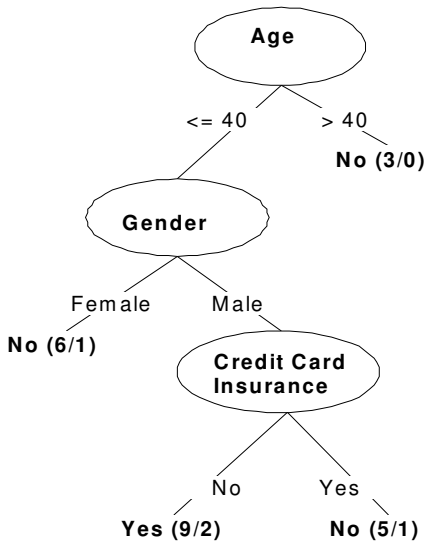
Question 5

a) Define very briefly (in no more than four sentences) the concepts of *star schema* and *snowflake schema* used in data warehouses. **[2]**

b) The decision tree below represents a generalization of data regarding customers that do or do not possess a margin account (this is a type of account through which the customer is allowed to borrow a limited amount of money for particular transactions). The attributes Age, Gender and Credit Card Insurance are input attributes while Margin Account is the output attribute. Answer the following:

i. Evaluate the error rate of the tree considering that the test instances are the same with the training instances (which were used to build the tree). **[1]**

ii. Extract all the production rules from the tree with the corresponding grades of accuracy and coverage. **[8]**



c) The small dataset given below contains a selection of data about customers that had applied for and obtained a loan. The company intends to develop a profile of the good credit risk customers, which will be used when processing new applications for a loan. A data mining strategy will be applied in order to build a reliable model.

i. Choose a data mining strategy that would be most appropriate for this application, justifying your choice. **[1]**

ii. In order to produce the model as a decision tree for this dataset, you are required to build the first level of the tree (that is, to find the root node), by applying a suitable algorithm. **[13]**

Income	Good credit risk	Home owner	Gender
40 – 50K	No	No	M
30 – 40K	Yes	No	F
40 – 50K	No	No	M
30 – 40K	Yes	Yes	M
50 – 60K	Yes	No	F
20 – 30K	No	No	F
30 – 40K	Yes	Yes	M
20 – 30K	No	No	M
30 – 40K	No	No	M
30 – 40K	Yes	No	F
40 – 50K	Yes	No	F
20 – 30K	Yes	No	M
50 – 60K	Yes	No	F
40 – 50K	No	No	M
20 – 30K	Yes	Yes	F

(End of examination)