UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2005

COMPUTING AND INFORMATION SYSTEMS

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

Duration: 2.15 hours

This paper contains five questions. ***You should answer 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.

Question 1

a. Define the terms "confidence" and "support" as they are used in association rules. **[2]**

b. Explain the main differences and similarities between association rules and production rules. **[3]**

c. The table below represents a snapshot of transactions in a grocery store. In order to find association rules with a minimum support of 20%, the Apriori algorithm may be used on this transactions set.

1. Trace the application of the algorithm. Show the candidate and frequent itemsets for each step when the database is scanned (Note that a candidate itemset is obtained through a join between itemsets obtained in the previous iteration). **[11]**
2. Indicate the association rules that will be generated with a minimum support of 20% and minimum confidence of 40%. **[9]**

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Question 2

a. Describe briefly the concept of feed-forward neural network (in no more than six statements). **[3]**

b. State the strengths and weaknesses of neural networks. **[7]**

c. The following subquestions are inter-related. You should answer them in the following order:

1. Draw a fully connected feed-forward neural network with three input nodes 1, 2, 3, one output node k and one hidden layer having two nodes i and j. **[3]**
2. Assume that the neural network from 1) has been trained to predict the future price of a favourite stock. The values of the input instances of the network correspond to three financial indicators whose values are in the range [100, 500]. In order to feed the network with one instance, convert the values of the three indicators 300, 180 and 500, to three input values between 0 and 1. **[4]**
3. Use the result of the conversion obtained at point 2) as input instance for the nodes 1, 2 and 3 respectively. Compute the output of the neural network using the sigmoid function (presented in the course). The weights of the connections are given in the table below, where $w_{x,y}$ denotes the weight of the connection between nodes x and y. **[6]**
4. Assuming that the range of the price of the favourite stock is [£20, £40], convert the output of the neural network obtained at 3) to a value that you can understand (that is, to a value between 20 and 40). **[2]**

Weight Values for the Neural Network

w_{1j}	w_{1i}	w_{2j}	w_{2i}	w_{3j}	w_{3i}	w_{jk}	w_{ik}
0.20	0.10	0.30	–0.10	–0.10	0.20	0.10	0.50

Question 3

a. Describe the major differences between OLTP databases and data warehouses. **[5]**

b. The data concerning credit card purchases are to be stored in a data warehouse. They contain details about the cardholders (name, gender and income range), the locations of the stores where credit cards are used (including addresses and region) and the category of the stores (supermarket, restaurant, etc).

1. Draw a star schema with no less than four dimensions, including also a time dimension. **[12]**
2. Sketch a three-dimensional OLAP cube from the above star schema. **[3]**
3. Express a concept hierarchy for the time dimension. **[1]**
4. Express in natural language on the cube from 2) four OLAP operations of the following types: slice, dice, roll-up and drill-down respectively. Note: you are required to state one operation per type. **[4]**

Question 4

a. Describe very briefly the six phases (up to three statements per phase) of the CRISP-DM process model. **[6]**

b. Describe the *Interpretation and evaluation* step of a KDD process model. **[5]**

c. Explain how supervised learning can be used to help evaluate the results of an unsupervised clustering. **[4]**

d. Perform two iterations of the K-Means algorithm in order to obtain two clusters for the input instances given in the table below. Assume that the first centres are the instances 1 and 3. Explain if more iterations are needed in order to get the final clusters. **[10]**

K-Means Input Values

Instance	X	Y
1	2.0	2.5
2	2.0	5.5
3	3.0	2.5
4	3.0	4.5
5	4.0	3.5
6	6.0	7.0

Question 5

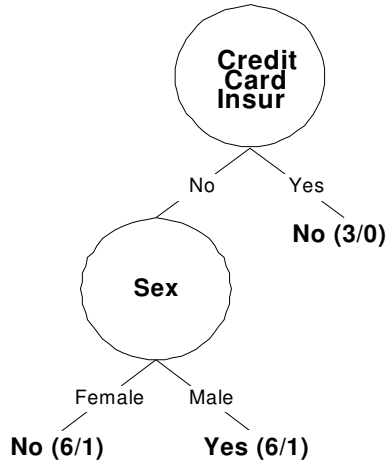
a. Describe briefly (in no more than four statements) the linear regression technique and its application to data mining. Illustrate the concept with an example. **[4]**

b. Apply the C4.5 algorithm to determine the top-level node of a decision tree to be built from the data in the table below, where Income Range, Credit Card Insurance and Sex are input attributes and Life Insurance Promotion is the output attribute. **[10]**

The Credit Card Promotion Database

Income Range	Life Insurance Promotion	Credit Card Insurance	Sex
40 – 50K	Yes	No	Male
30 – 40K	No	No	Female
40 – 50K	Yes	No	Male
30 – 40K	No	Yes	Male
50 – 60K	No	No	Female
20 – 30K	Yes	No	Female
30 – 40K	No	Yes	Male
20 – 30K	Yes	No	Male
30 – 40K	Yes	No	Male
30 – 40K	No	No	Female
40 – 50K	No	No	Female
20 – 30K	No	No	Male
50 – 60K	No	No	Female
40 – 50K	Yes	No	Male
20 – 30K	No	Yes	Female

c. Write the production rules for the decision tree below indicating the rule accuracy and coverage. **[7]**



d. In order to evaluate the quality of a supervised learning model, one builds the confusion matrix shown below.

1. Compute the number of instances correctly classified. **[1]**
2. Compute the accuracy and error levels of the model. **[1]**
3. Compute the lift for the model. **[2]**

Model	**Computed Accept**	**Computed Reject**
Accept	626	10
Reject	40	324