UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination May 2010

COMPUTING AND INFORMATION SYSTEMS

IS53023B (CIS338B)
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 may be used but 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:

a) Apply the *Apriori* algorithm on the dataset below, using for *minimum support* the value of 0.25. **[14]**
b) Using the *frequent itemsets* outputted by the algorithm from (a), generate the strong (or interesting) association rules with a minimum confidence of 0.7. Compute the support of each generated strong association rule. **[11]**

humidity Nominal	windy Nominal	**play** Nominal
high	FALSE	no
high	TRUE	no
high	FALSE	yes
high	FALSE	yes
normal	FALSE	yes
normal	TRUE	no
normal	TRUE	yes
high	FALSE	no
normal	FALSE	yes
normal	FALSE	yes
normal	TRUE	yes
high	TRUE	yes
normal	FALSE	yes
high	TRUE	no

Question 2:

a) Formally define the notion of association rule and its quality measures of confidence and support. Then provide an example of an association rule with confidence and support, and explain its meaning. **[6]**
b) Formally define the notion of production rule and its quality measures of accuracy and coverage. Then provide an example of a production rule with accuracy and coverage, and explain its meaning. **[6]**
c) What is the main difference between association rules and production rules? **[2]**
d) Mention the data mining strategies in which association rules and production rules appear, and name two data mining techniques/algorithms that generate association rules and production rules, respectively. **[4]**
e) Assume that, after you have applied a data mining technique to profile the customers of a telecom company that have the tendency to churn (i.e. to leave the service), you get the production rule ***IF last_month_call_minutes < 100 and service=basic THEN churn=yes***, with the accuracy=0.6 and the coverage 0.3. If last month the company had 10,000 customers of which 500 churned (that is, for these customers, the value of attribute churn is yes; let us call them churners), compute the following:
    i. The number of customers that churned last month, and had a basic service and called less than 100 minutes. **[2]**
    ii. The number of all customers (churners and non churners) last month, that had a basic service and called less than 100 minutes. **[2]**
    iii. The number of customers that did not churn last month, and had a basic service and called less than 100 minutes. **[1]**
    iv. The accuracy and coverage of the rule ***IF last_month_call_minutes < 100 and service=basic THEN churn=no*** . **[2]**

Question 3:

a) State in pseudocode the *K-Means* algorithm and explain it. **[8]**
b) Consider the dataset formed of the following instances:

X	**Y**
4	20
4	10
16	8
10	16
14	10
12	8
2	4
8	18

The K-Means algorithm is to be used to cluster the above dataset in three clusters. The algorithm starts with the first centres given by the instances (4,20), (10,16) and (2,4), respectively. You are required to compute the first clusters and their new centres. Are the computed clusters the final ones? Explain your answer. **[17]**

Question 4:

a) Briefly define the following concepts in the context of data warehouses (do not use more than two statements per concept):

i. Granularity of a dimension **[2]**
ii. Star schema **[2]**
iii. Snowflake schema **[2]**
iv. Constellation schema **[2]**
v. Independent data mart **[2]**

b) Suppose that a data warehouse consists of the four dimensions *date, spectator, location*, and *game*, and the two measures, *count* and *charge* (to be stored in the fact table), where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. You are required to:

i. Provide a concept hierarchy for each of the four dimensions. **[4]**
ii. Choose an appropriate set of attributes for the four dimensions and draw a star schema diagram for the data warehouse. **[7]**
iii. Provide two examples of slice and dice operations, respectively, on the data warehouse. **[4]**

Question 5:

a) State in pseudocode the *Agglomerative Clustering* algorithm and explain it. **[8]**
b) Apply the Agglomerative Clustering algorithm on the dataset below, formed of five instances (denote them by I1 to I5 from top to bottom, for easy identification). Finally choose as clustering result the solution formed of two clusters. **[17]**

Income Range	Magazine Promotion	Watch Promotion	Life Insurance Promotion	Sex
40–50K	Yes	No	No	Male
25–35K	Yes	Yes	Yes	Female
40–50K	No	No	No	Male
25–35K	Yes	Yes	Yes	Male
50–60K	Yes	No	Yes	Female