UNIVERSITY OF LONDON
GOLDSMITHS COLLEGE

Department of Computing

B. Sc. Examination 2014

IS53023B
Data Mining

Duration: 2 hours 15 minutes

Date and time:

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

a. Do supervised learning and unsupervised clustering require:

i. input attributes?
ii. output attributes?

Justify your answer. [5]

b. The dataset in the table below is used to build a Naive Bayes Classifier model. The output attribute is Car and all the other attributes are input attributes. You are required to illustrate how the algorithm works in order to classify the following instance defined by MagazinePromotion=Yes, WatchPromotion=Yes, LifeInsurancePromotion=No, and CreditCardInsurance=No. [20]

MagazinePromotion	WatchPromotion	LifeInsurancePromotion	CreditCardInsurance	Car
Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	No
No	No	No	No	Yes
Yes	Yes	Yes	Yes	Yes
Yes	No	Yes	No	No
No	No	No	No	No
Yes	Yes	Yes	Yes	Yes
No	No	No	No	Yes
Yes	No	No	No	Yes
Yes	Yes	Yes	No	No

Question 2

a. Explain why one can obtain more than one result (or clustering) with the k-Means algorithm applied on a dataset to cluster. Then define the criterion used to choose the best result/clustering, and briefly explain it. [6]

b. Apply the k-Means algorithm on the dataset below for k=2. You are required to start with the instances 1 and 3 as the centres of the two clusters, and to perform two iterations only. Mention whether or not you obtained the final clusters after the two iterations, and justify your answer. [19]

Instance	Att1	Att2
1	20	25
2	20	55
3	30	25
4	30	45
5	40	35
6	60	70

Question 3

a. What is the purpose of the sensitivity analysis for neural networks for supervised learning? Provide the steps of this method. [8]

b. A fully connected feed-forward neural network has three layers: the input layer has the nodes 1, 2, 3, the hidden layer has the nodes 4 and 5, and the output layer has the node 6. The neural network is used to predict the price of a share based on the input values 600, 360 and 1000 of the three input attributes corresponding to the nodes 1, 2, and 3, respectively. It is known that all the three input attributes have the same range of values, namely the interval between 200 and 1000, and that the output attribute has the range of values the interval between 40 and 80. Moreover, the weights w(i,j) of the connections between the nodes i and j, obtained from the training of the neural network are:

w(1,4)=0.2, w(1,5)=0.4, w(2,4)=-0.2, w(2,5)=0.6, w(3,4)=0.4, w(3,5)=-0.2, w(4,6)=1, and w(5,6)=0.2

You are required to draw the neural network above, and to calculate the predicted price of the share (that is, the value of the output attribute) based on the values provided above for the three input attributes. Illustrate in detail all the steps of your computation. *Note: as presented in the course, you are required to normalise the input values to the interval [0,1], and to use the sigmoid function $f(x)=1/(1+e^{-x})$ in the non-input nodes.* [17]

Question 4

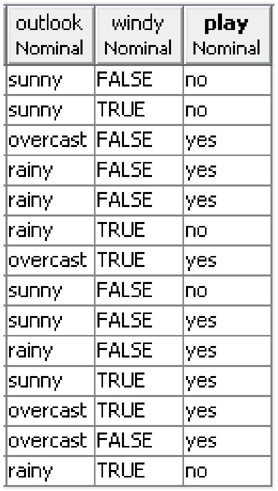
a. Name three OLAP operations and illustrate each of them with a concrete example. Use one statement per example. [6]

b.

i. Give one example of concept hierarchy for a time dimension in a data warehouse. [1]

ii. Explain what granularity in the context of the example from i. above. [2]

c. Apply the Apriori algorithm on the dataset below with the minimum support of 0.35 and the minimum confidence of 0.75. In particular display all the frequent itemsets, and all the candidates. Display all the association rules you get and decide which ones are strong rules. Finally you are required to compute also the support for the strong rules. [16]



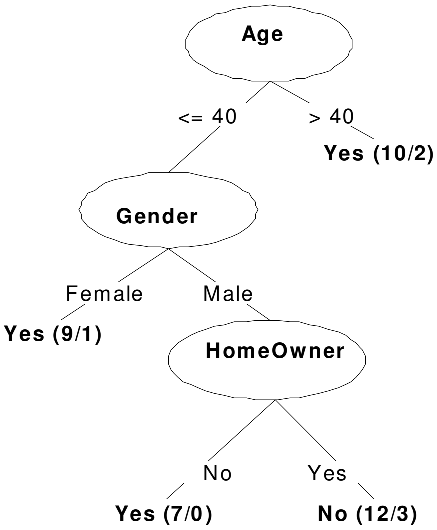
outlook Nominal	windy Nominal	**play** Nominal
sunny	FALSE	no
sunny	TRUE	no
overcast	FALSE	yes
rainy	FALSE	yes
rainy	FALSE	yes
rainy	TRUE	no
overcast	TRUE	yes
sunny	FALSE	no
sunny	FALSE	yes
rainy	FALSE	yes
sunny	TRUE	yes
overcast	TRUE	yes
overcast	FALSE	yes
rainy	TRUE	no

Question 5

a. Briefly explain three advantages of decision trees (use an itemised list to provide your answer). [6]

b. The decision tree below, in which the output attribute is called LifeInsurancePromotion and takes the values Yes or No, has been trained/built on a dataset D with customer details. You are required to:

i. Briefly explain what the notation Yes(10/2) from the rightmost terminal node in the decision tree represents. [1]
ii. Compute how many instances the dataset D contains. How many instances are in class Yes and how many are in class No? [3]
iii. Compute the error rate of this decision tree if it were tested on the same dataset D that was used to build it. [3]
iv. Extract all the production rules from the decision tree and compute their accuracy and coverage. [12]


Age
<= 40
> 40
Yes (10/2)
Gender
Female
Male
Yes (9/1)
HomeOwner
No
Yes
Yes (7/0)
No (12/3)