UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2010/11

COMPUTING AND INFORMATION SYSTEMS

IS53023B (CIS338B)
Data Mining

Duration: 2 hour 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 in order to generate all the sets of frequent itemsets L1, L2, L3 and so on, knowing that the support threshold is 0.25. Clearly mention when the algorithm stops, and briefly explain why it stops. **[13]**

b) Using the result of the Apriori algorithm from point (a), build all the association rules that can be obtained from L2 only, and compute their confidence and support. **[12]**

SalaryRange Nominal	CreditCardIns Nominal	**Gender** Nominal
40-50K	No	Female
30-40K	No	Male
40-50K	No	Female
30-40K	Yes	Female
50-60K	No	Male
20-30K	No	Male
30-40K	Yes	Female
20-30K	No	Female
30-40K	No	Female
30-40K	No	Male
40-50K	No	Male
20-30K	No	Female
40-50K	No	Female
20-30K	Yes	Male

Question 2

a)

i. State Bayes' theorem that is used in naive Bayesian classifiers, and briefly explain each of the terms involved in the theorem. **[5]**

ii. Name one application in which the naive Bayesian classifiers are largely used. **[1]**

b) Provide an itemised list of up to three advantages of machine learning techniques over statistical techniques when used in data mining. **[3]**

c) Apply the decision tree algorithm based on goodness scores on the dataset below, in order to only build the root of a decision tree, knowing that the last column (called churn) in the dataset below is the output attribute. **[16]**

region Nominal	marital Nominal	retire Nominal	custcateg Nominal	**churn** Nominal
Zone 3	Married	No	E-service	Yes
Zone 2	Unmarried	No	Basic service	Yes
Zone 3	Married	No	Plus service	No
Zone 2	Unmarried	No	Basic service	Yes
Zone 3	Unmarried	No	Total service	No
Zone 2	Married	No	Plus service	No
Zone 2	Married	No	Total service	No
Zone 1	Married	No	Total service	Yes
Zone 3	Married	No	Total service	Yes
Zone 1	Unmarried	Yes	E-service	No
Zone 1	Married	No	E-service	No
Zone 3	Married	No	E-service	No
Zone 1	Unmarried	No	Total service	No
Zone 3	Unmarried	No	Basic service	No
Zone 2	Married	No	E-service	Yes
Zone 2	Married	No	Basic service	Yes
Zone 1	Unmarried	No	Basic service	No
Zone 1	Married	No	Plus service	Yes
Zone 1	Unmarried	No	Plus service	No
Zone 1	Married	No	Plus service	Yes

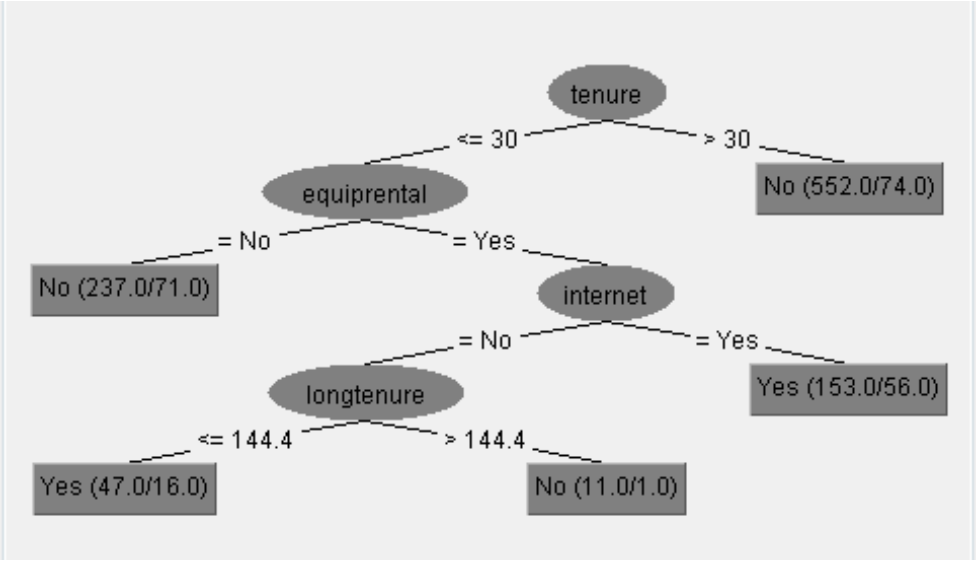
Question 3

a) Briefly state what a production rule is, and define the rule accuracy and coverage. **[3]**

b) The decision tree below is a model used to predict the customers that are likely to churn (assume that the dataset from which the tree was built contains an attribute called "churn" whose values are *Yes* and *No,* corresponding to the fact that a customer churned or did not churn in the past, respectively).

i. Write all the attributes in this model (one per line), and for each of them specify whether it is an input or an output attribute, and whether its type is numeric or nominal. **[5]**

ii. Write all the production rules from the decision tree. For each rule compute the accuracy and coverage. Note that the figures in each decision leaf have the following meaning: the first figure represents the total number of instances that satisfy the conditions on the path from the root to the leaf, and the second figure (if it exists) represents the number of those instances that are exceptions to the decision (if these exceptions exist). **[17]**



Question 4

In the context of Genetic Learning for training a Neural Network you are required to answer the following questions:

a)

- **i.** Define the mean absolute error and mention its purpose here. **[3]**
- **ii.** Define the squared root mean error and mention its purpose here. **[3]**
- **iii.** Define the fitness score of a population element and mention its purpose here. **[2]**
- **iv.** Explain the Genetic Learning algorithm for training a Neural Network. **[7]**

b) The following four weight elements belong to a population used to train a Neural Network with Genetic Learning:

element E1:	0.22	-0.26	0.29	0.26	-0.25	0.27
element E2:	0.14	0.19	0.17	0.14	-0.11	0.16
element E3:	0.45	0.46	0.42	0.43	0.49	-0.47
element E4:	0.81	0.89	-0.85	0.83	0.82	-0.86

- **i.** Generate two new elements only via the Mutation operation, indicating the initial elements from which they are obtained, and briefly stating how they are used to generate the new elements. **[4]**
- **ii.** Generate two new elements only via the Crossover operation, indicating the initial elements from which they are obtained, and briefly stating how they are used to generate the new elements. **[4]**
- **iii.** How many new elements can be obtained from the initial elements E1, E2, E3, E4 by applying one Mutation operation only on each initial element, and one Crossover operation only on each pair of initial elements? Briefly justify your answer. **[2]**

Question 5

a) You are required to cluster the dataset provided in the table below by using the Agglomerative Clustering algorithm (note that the first column of the table indicates the instance number and is not to be involved in the clustering). Before the algorithm application, clearly define the similarity between two instances, and between two clusters. At the end of the algorithm application, choose as final clustering the solution which is formed of three clusters, and clearly indicate it. **[21]**

No.	gender Nominal	income Nominal	internet Nominal	equipment Nominal	**ebill** Nominal
1	M	Medium	Yes	No	No
2	F	Low	Yes	Yes	Yes
3	M	Medium	No	No	No
4	M	Low	Yes	Yes	Yes
5	F	High	Yes	No	Yes

b) Briefly define the term "outlier" in terms of clustering. **[2]**

c) Mention a practical data mining application of the unsupervised clustering strategy, and briefly describe it in one sentence. **[2]**