UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2012

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 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) Describe two advantages and two disadvantages of Neural Networks in the context of Data Mining. **[8]**

b)

i. Provide the definition of data warehouses and mention where they are used. **[3]**

ii. List the names of three OLAP operations. **[3]**

c)

i. Assume that $e_1, e_2, ..., e_{100}$ are the estimates of the prices of 100 houses that have been obtained by using a Data Mining algorithm, and that $p_1, p_2, ..., p_{100}$ are the prices with which these houses were actually sold, respectively. Using the above prices and estimates, you are required to choose and define only two of the following measures: the mean absolute error MAE; the mean squared error MSE; the root mean squared error RMSE. **[4]**

ii. Briefly explain the purpose of using these measures in Data Mining. **[1]**

d) Describe two assumptions that are made about data in the Naive Bayes algorithm. **[4]**

e) State whether or not the Naive Bayes algorithm can handle missing data in a dataset. Explain your answer. **[2]**

Question 2

a) Apply the k-Nearest Neighbour algorithm with k=3 on the dataset below, where the top ten instances form the training dataset, and the bottom two instances are **new** instances that need to be classified (that is, values for the output attribute need to be computed). The output attribute is *churn* and the first column called *No.* in the table below is an instance id whose only purpose is to identify instances and is not involved in computations. As a result of the application of the algorithm, for **each** of the two new instances you are required to:

i. Provide the distances computed by the algorithm. **[5]**

ii. Provide the ids of the 3 nearest neighbours. **[3]**

iii. Provide the decision. **[2]**

No.	region Nominal	marital Nominal	retire Nominal	gender Nominal	custcateg Nominal	**churn** Nominal
1	Zone 3	Unmarried	No	Female	Total service	No
2	Zone 2	Married	No	Male	E-service	Yes
3	Zone 3	Married	No	Female	E-service	Yes
4	Zone 2	Married	Yes	Female	Total service	No
5	Zone 1	Married	No	Male	Plus service	Yes
6	Zone 2	Married	No	Male	Basic service	No
7	Zone 1	Married	No	Male	Total service	No
8	Zone 1	Married	No	Female	E-service	Yes
9	Zone 2	Unmarried	No	Male	Plus service	No
10	Zone 3	Unmarried	No	Female	Basic service	No
11	Zone 2	Unmarried	No	Male	Basic service	
12	Zone 1	Married	Yes	Male	Total service	

b) A classification model is evaluated on a test dataset, leading to the confusion matrix provided below. You are required to: formally define the following measures by providing their calculation formulae, then provide their meaning in English, and finally calculate these measures using the concrete confusion matrix provided below. The measures are:

i. Accuracy **[3]**

ii. Sensitivity **[3]**

iii. Precision **[3]**

iv. Specificity **[3]**

v. Lift **[3]**

Confusion matrix:

	Computed Accept	**Computed Reject**
Accept	61	23
Reject	38	140

Question 3

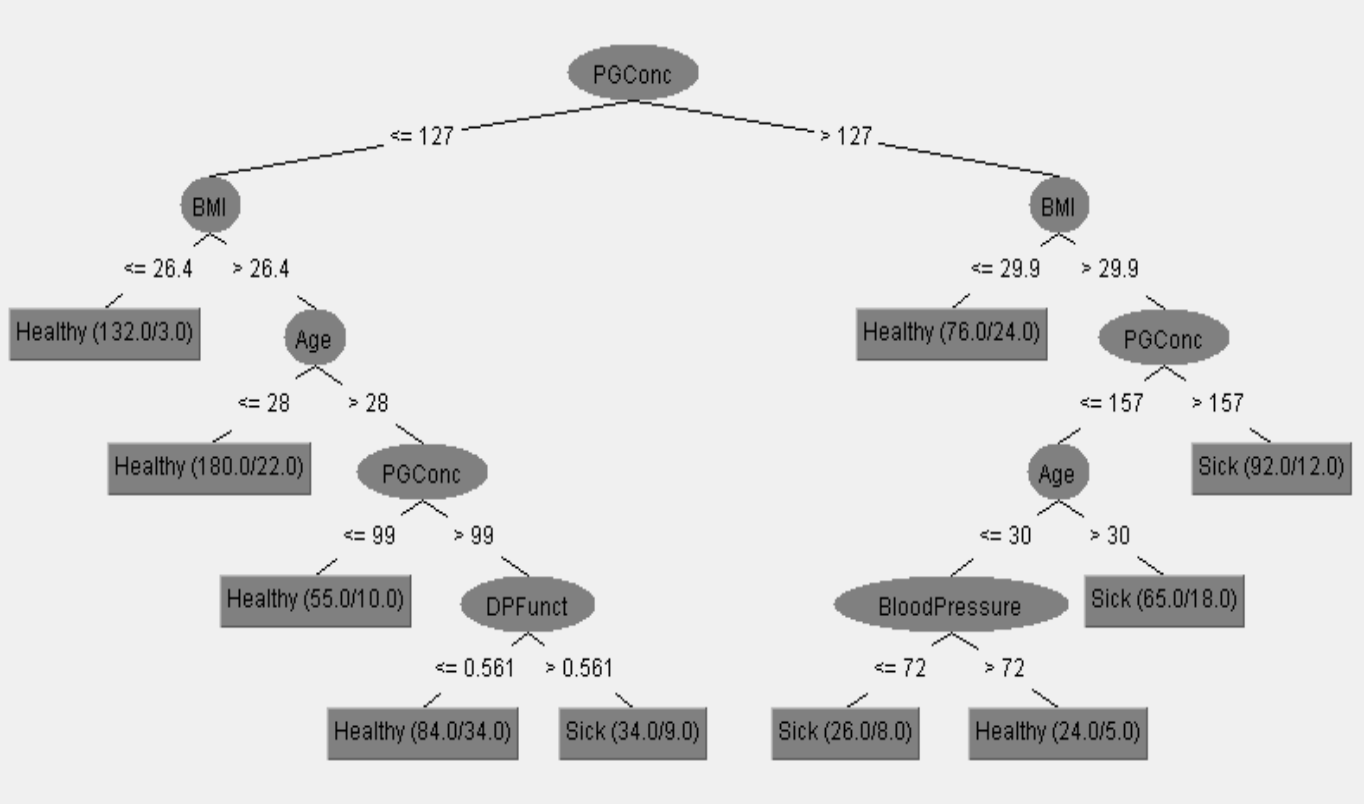
The decision tree provided below was built from a medical dataset with details about patients (one row per patient) in order to predict which future patients are likely to develop a specified illness. The output attribute, called *Outcome*, has the values *Sick* and *Healthy*. These values serve as decisions which are provided in the leaves of the decision tree. The figures/numbers in each decision leaf have the following meaning: the first figure represents the total number of patients that satisfy the conditions imposed on the input attributes on the path from the root to the leaf, and the second figure represents the number of those patients that are exceptions to the decision. For instance Healthy(132.0/3.0) in the leftmost decision leaf means that there are 132 patients satisfying the condition PGConc <=127 and the condition BMI<= 26.4, and all these patients are healthy except 3 of them (which are sick). You are required to:

a) Write down all the attributes. **[3]**

b) Calculate the number of patients which are sick and briefly justify your answer. **[4]**

c) Write all the production rules corresponding to the decision *Sick* that can be extracted from the decision tree (list the rules in the order they appear in the decision tree, from the leftmost to the rightmost). For each such rule compute the accuracy and coverage. **[12]**

d) Draw a new decision tree with 4 levels obtained by pruning the decision tree below (assume that level one is that of the root). In the new decision tree leaves, clearly indicate each decision and the two numbers (in brackets) associated to it. Briefly explain how you built the new decision tree. **[6]**



Question 4

a) In the context of the Agglomerative Clustering algorithm applied to a dataset with nominal values:

i. Define the similarity between two instances. **[2]**

ii. Define the similarity between two clusters. **[2]**

b) Illustrate the application of the Agglomerative Clustering algorithm on the dataset below (the first column is an id column which will not be used in the similarity computation). In particular show the computation process that is to start with five different clusters and is to terminate with one cluster. Then choose and write down the solution formed of two clusters as the result of the clustering. **[15]**

No.	Gender Nominal	Home_owner Nominal	Credit_card_type Nominal	Credit_card_insurance Nominal	**Life_insurance** Nominal
1	Female	No	Gold	Yes	No
2	Male	No	Standard	No	Yes
3	Female	Yes	Gold	Yes	No
4	Female	No	Standard	No	Yes
5	Male	No	Platinum	Yes	Yes

c) Name and provide the pseudo-code of a clustering algorithm that is based on cluster centre computations. **[6]**

Question 5

In the context of the Association Analysis / Market Basket Analysis you are required to:

a) Apply the Apriori algorithm on the dataset provided below knowing that the support threshold is 0.2. In particular you are required to provide the minimum count and the frequent itemsets with their counts, and to mention when and why the algorithm stops. **[18]**

b)

i. Briefly justify why higher values of the count of frequent itemsets correspond to higher values of the support of the association rules generated from those frequent itemsets. **[1]**

ii. Using the frequent itemsets found at point (a) above, write the association rules having the largest support. Calculate the confidence and the support for these rules precisely. **[6]**

Gender Nominal	CredCard Nominal	**LifeIns** Nominal
F	Basic	Yes
M	Platinum	No
M	Basic	Yes
F	Platinum	Yes
F	Gold	Yes
M	Gold	No
F	Basic	No
F	Diamond	Yes
M	Gold	Yes
F	Gold	Yes