# **GOLDSMITHS, UNIVERSITY OF LONDON**

**Department of Computing**

# **B. Sc. Examination 2015-2016**

**IS53023B Data Mining**

**Duration: 2 hours 15 minutes**

**Date and time:**

This paper is in two parts: part A and part B.  You should answer ALL questions from part A and TWO questions from part B.  Part A carries 40 marks, and each question from part B carries 30 marks.  The marks for each part of a question are indicated at the end of the part in [.] brackets.

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**

**PART A**

**Question 1**

a) Mention four different solutions to handle missing data in a dataset. Do not use more than one statement for each solution, and provide your answer in an itemised list. [8]

b)

|  |  |  |
| --- | --- | --- |
|  | Actual  yes | Actual  no |
| Predicted yes | 300 | 200 |
| Predicted no | 500 | 1000 |

For a binary classification problem for which the confusion matrix is provided above, show how you calculate the following performance measures:  
  
accuracy, error rate, precision, sensitivity, specificity and lift.

In particular, for each such performance measure, provide the formula and calculate it using the confusion matrix above, and state in plain English what the performance measure means. Your answer should be provided with a separate entry for each of the performance measures mentioned above. [18]

c) Briefly describe the distinction between the terms of each of the following pairs. Do not use more than two sentences per pair of terms.

1. Data warehouse and operational database
2. Training data and test data
3. Input attribute and output attribute
4. Supervised learning and unsupervised clustering
5. Classification and regression
6. Supervised learning and classification
7. Numeric attribute and nominal attribute. [14]

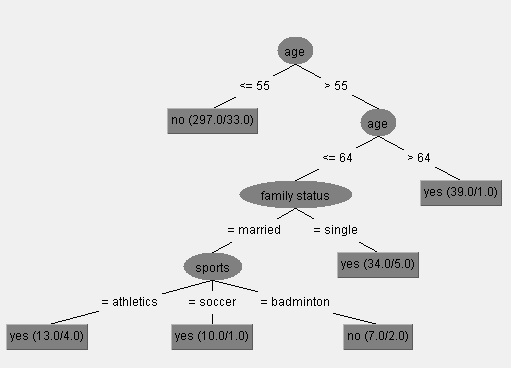
**PART B**

**Question 2**

The output attribute for the decision tree below is called LI, standing for life insurance, and its values can be “yes” or “no”.

Each terminal node (or leaf) in the decision tree indicates a class (“yes” or “no”) and a pair of figures which have the following meanings: the first figure shows how many instances satisfy the conditions appearing in the corresponding path from the root to the leaf; all these instances are in the class indicated in the leaf, except for a number of instances, which is indicated by the second figure. For instance, for the leaf of the leftmost path, the class is “no”, and there are 297 instances satisfying the condition age <=55. All these instances are in class “no” except for 33 of them, which are obviously in the other class (“yes”).

1. Using the given decision tree, you are required to compute the number of instances in the classes “yes” and “no” based on the dataset used to train the decision tree. [4]
2. Define the accuracy and the coverage of a production rule. Do not use more than one statement for each definition. [4]
3. Write all the production rules starting from the leftmost path in the decision tree, from the root to a leaf. For each rule provide its accuracy and coverage showing your work. [18]
4. Briefly explain how the splitting point for a numeric attribute such as “age” is calculated by the decision tree algorithm. Why different splitting points may appear for the same attribute in different nodes of the tree? Do not use more than two statements for your answer. [4]



**Question 3**

1. Explain the main steps of the DBSCAN algorithm for clustering. You may illustrate your explanation with an example. [10]
2. Which of the two algorithms DBSCAN and K-Means handle better outliers? Justify your answer. [4]
3. Illustrate the application of first iteration of the K-Means algorithm on the following dataset given by the points: A1=(20,100), A2=(20,50), A3=(80,40), A4=(50,80), A5=(70,50), A6=(60,40), A7=(10,20), A8=(40,90). The initial three cluster centres are A1, A4 and A7. After the first iteration is completed, stop the algorithm illustration and state if further iterations are needed and justify your answer. [16]

**Question 4**  
a) Explain the main steps of the Random Forest algorithm. Then give an example of a situation in which it is recommended to consider it in practice, and justify why. [12]

b) Provide two important differences between the techniques of Random Forest, and of Bagging with decision trees models. [4]

c) A feed-forward neural net with three layers has the following weights after training:  
  
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  
  
where 1,2, and 3 are the nodes in the input layer, 4 and 5 are nodes in the hidden layer, and 6 is the node in the output layer.

Using this neural net, you are required to predict the output attribute value based on the values 60, 36 and 100 of the three input attributes corresponding to the nodes 1, 2, and 3, respectively, and to show your work. It is known that all the three input attributes have the same range of values between 20 and 100, and that the output attribute has the range of values between 400 and 800.

*Note: you are to use the sigmoid function f(x)=1/(1+e-x) in the non-input nodes.* [14]