# **GOLDSMITHS, UNIVERSITY OF LONDON**

**Department of Computing**

# **B. Sc. Examination 2016-2017**

**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.

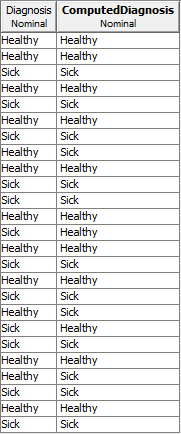
Electronic calculators must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

**THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM**

**PART A**

**Question 1**

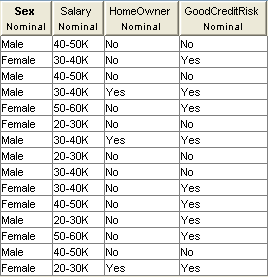
1. Define and explain the accuracy, error rate, precision, sensitivity, specificity, and the lift in a 2-class classification problem. In particular, for each performance measure, provide the formula, and a brief interpretation of that performance measure in plain English. **[12]**
2. The picture below shows only a part of a test dataset, namely the values of the output attribute Diagnosis (1st column) and the predicted values for Diagnosis (2nd column) which were made by using a data mining algorithm. Each row corresponds to a diagnosed patient. You are required to:
3. Build the confusion matrix, and calculate the accuracy, error rate, precision, sensitivity, specificity, and lift with respect to the class *Sick*. **[10]**
4. What is the most important performance measure for this problem from those calculated in (i)? Explain your choice. **[4]**
5. Mention 7 data mining algorithms that are appropriate to be applied in the classification problem mentioned above to predict diagnoses. Mention 3 algorithms that are not appropriate to use in this problem to predict diagnoses. Briefly justify your choice. **[14]**



**PART B**

**Question 2**

a) Apply the decision tree algorithm based on goodness score attribute selection criterion studied in class, in order to find out the root node of the decision tree model. The output attribute is GoodCreditRisk, and all the other attributes are input attributes. **[23]**



1. Explain the concept of overfitting. In addition, explain pruning, prepruning and postpruning as methods for decision tree overfitting prevention. Give an example of prepruning and an example of postpruning. **[7]**

**Question 3**

1. Name two algorithms for training feed-forward neural networks, and briefly explain them in no more than 4 statements per algorithm. **[8]**
2. Draw a fully connected feed-forward neural network with an input layer of three nodes denoted 1,2 and 3, a hidden layer of two nodes denoted 4 and 5, and an output layer of one node noted 6. **[4]**
3. A feed-forward neural network as that of (b) has been trained. The values of the input instances of the neural network correspond to three input attributes each of which having values in the range [100, 500]. You are required to:
   1. Normalize the instance 300, 180 and 500 by applying the min-max normalization, and compute the actual values that feed the network. The corresponding normalised values will be used as an input for the nodes 1, 2, and 3, respectively, in the next point (ii). **[6]**
   2. Compute the output of the neural network when using the input from (i). Employ the sigmoid function f(x)=1/(1+exp(-x)). The weights of the connections are as follows, where wx,y denotes the weight of the connection between nodes x and y: w1,4=0.1; w1,5=0.2; w2,4=-0.1; w2,5=0.3; w3,4=0.2; w3,5= -0.1; w4,6=0.5; w5,6=0.1. **[9]**
   3. Convert the output of the neural network obtained at (ii) to the range of the dependent variable [1000, 2000]. **[3]**

**Question 4**

a) Compare and contrast the association rules with the production rules. **[6]**

b) Apply the Apriori algorithm on the dataset below with the minimum support of 0.37. In particular, you are required to generate all the frequent itemsets, and all the strong association rules formed of two items, knowing that the minimum confidence threshold is 0.72. Compute the support for these strong rules. **[24]**

