UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2007

COMPUTING AND INFORMATION SYSTEMS

IS53023A (CIS338)
Advanced Database Technologies – Data warehousing and 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. Write and explain the pseudo-code for the Nearest Neighbour algorithm and mention what data mining strategy it is part of. **[4]**

b. Explain in no more than three statements the similarity and the difference between the K-Nearest Neighbour algorithm and the Nearest Neighbour algorithm. **[3]**

c. Consider the following application of the Nearest Neighbour algorithm used as a classifier for diagnosing patients based on their symptoms. In order to evaluate the resulting data model, the algorithm is used with the training dataset from Table 1 below and the test dataset from Table 2 below. Note that the attribute PatientID is to be ignored in computing classes as it is not relevant to this application. Note also that SoreThroat, Fever, SwollenGlands, Congestion, Headache are input attributes and Diagnosis is the output attribute. You are required to

i. List the classes. **[1]**

ii. Generate the computed classes for the instances from the test dataset using the algorithm. **[10]**

iii. Build the confusion matrix corresponding to the test dataset. **[2]**

iv. Evaluate the accuracy and error rates of the classifier. **[2]**

d. Mention three techniques for supervised learning. **[3]**

Table 1: The training dataset

PatientID Numeric	SoreThroat Nominal	Fever Nominal	SwollenGlands Nominal	Congestion Nominal	Headache Nominal	**Diagnosis** Nominal
1.0	No	No	No	No	No	Healthy
2.0	Yes	Yes	Yes	Yes	Yes	Strepthroat
3.0	No	No	No	Yes	Yes	Allergy
4.0	Yes	Yes	No	Yes	No	Cold
5.0	Yes	No	Yes	No	No	Strepthroat
6.0	No	Yes	No	Yes	No	Cold
7.0	No	No	No	Yes	No	Allergy
8.0	No	No	Yes	No	No	Strepthroat
9.0	Yes	No	No	Yes	Yes	Allergy
10.0	No	Yes	No	Yes	Yes	Cold
11.0	Yes	Yes	No	Yes	Yes	Cold

Table 2: The test dataset

PatientID Numeric	SoreThroat Nominal	Fever Nominal	SwollenGlands Nominal	Congestion Nominal	Headache Nominal	**Diagnosis** Nominal
21.0	No	No	Yes	Yes	Yes	Allergy
22.0	No	No	No	No	Yes	Healthy
23.0	Yes	Yes	No	No	Yes	Strepthroat
24.0	No	Yes	No	Yes	No	Cold
25.0	No	Yes	No	No	Yes	Allergy

Question 2

a. Describe the concept of Self-Organising Maps, mentioning what data mining strategy they can be applied to. Provide an example to illustrate how they work. Do not use more than ten statements in your answer. **[6]**

b. Provide the two formulas for computing the Lift, using conditional probabilities and frequencies. Illustrate how the Lift can be computed considering an example of data model testing. Do not use more than five statements in your answer. **[4]**

c. Tackle the following points in the order shown below:

i. Provide an example of a feed-forward neural network with three layers: the input layer should have three nodes noted 1,2 and 3, the hidden layer should have two nodes noted 4 and 5 and the output layer should have one node noted 6. In addition the neural network should have the maximum number of connections. **[3]**

ii. Assume that the neural network from i) has been trained to estimate the value of a depended variable (or output attribute). The values of the input instances of the network correspond to three independent variables (or input attributes) whose values are in the range [50, 250] (the range is the same for all three independent variables). The network has as input the instance 150, 90 and 250. Normalize these values and compute the actual values that feed the network. The corresponding normalised values will be used as an input for the nodes 1, 2, 3, respectively. You are required to compute the output of the neural network from point i) using the sigmoid function for the nodes evaluations. The weights of the connections are as follows, where $w_{x,y}$ denotes the weight of the connection between nodes x and y: $w_{1,4}$=0.1; $w_{1,5}$=0.2; $w_{2,4}$=-0.1; $w_{2,5}$=0.3; $w_{3,4}$=0.2; $w_{3,5}$= -0.1; $w_{4,6}$=0.5; $w_{5,6}$=0.1. **[10]**

iii. Convert the output of the neural network obtained at ii) to a value in the range [10,20], which is the range of the dependent variable. **[2]**

Question 3

a. Describe the distinction between the terms of each of the following pairs. Do not use more than three sentences per pair of terms.

i.	Data warehouse and operational database	**[2]**
ii.	Training data and test data	**[2]**
iii.	Input attribute and output attribute	**[2]**
iv.	Shallow knowledge and hidden knowledge	**[2]**
v.	Exemplar view and probabilistic view	**[2]**
vi.	Probabilistic view and classical view	**[2]**
vii.	Supervised learning and unsupervised clustering	**[2]**
viii.	Prediction and estimation	**[2]**
ix.	Supervised learning and classification	**[2]**

b. Define the concept of confusion matrix and then explain how it is employed for assessing the quality of a data model built with a supervised learning algorithm. Illustrate your explanation with an example. **[7]**

Question 4

a. Provide the C4.5 algorithm used for building decision trees. Define the concept of information gain of an attribute and describe and explain in detail the criterion of attribute selection based on this concept. **[10]**

b. Describe and explain the conceptual clustering technique, providing the standard conceptual clustering algorithm. **[6]**

c. Define the concepts of association rule, confidence, support and strong association rule in the context of the market basket analysis (providing formulas for computing the confidence and support). Then provide and explain the Apriori algorithm. **[9]**

Question 5

a. State Bayes' theorem and briefly explain the meaning associated to the probability and conditional probability terms used in the theorem, in the context of a classification problem. **[5]**

b. Show how missing data are handled by a Bayes classifier (using no more than two statements). **[2]**

c. Illustrate how numeric data is handled by a Bayes classifier, providing an example that should consider a dataset containing an output attribute and two input attributes, of which one is nominal and the other is numeric having a particular probabilistic distribution. **[6]**

d. Use the dataset contained in the table below in order to:

i. Determine the counts and probabilities necessary to build a Bayes classifier for the output attribute *Life Insurance Promotion*. **[6]**

ii. Determine the value of *Life Insurance Promotion* for the following instance

Magazine Promotion = Yes
Watch Promotion = Yes
Credit Card Insurance = No
Sex = Female
Life Insurance Promotion = ? **[3]**

iii. Repeat the evaluation from point ii) of d, but assume that the gender of the customer is unknown (that is, assume the same values for the attributes excepting attribute Sex). **[3]**

Dataset for Question 5.d

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Yes	No	No	No	Male
Yes	Yes	Yes	Yes	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	Yes	Yes	Yes	Male
No	No	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female