INSTITUTE OF TECHNOLOGY
BLANCHARDSTOWN

Institute of Technology
Blanchardstown
Institiúid Teicneolaíochta
Baile Bhlainséir

BACHELOR OF SCIENCE IN COMPUTING
(Information Technology)
BN302

Data Mining
COMP H3024

Semester 2

Internal Examiner(s): Laura Keyes

External Examiner(s): Mr. John Dunnion
Dr. Richard Studdert

Thursday 24th May 2007
12.30pm - 2.30pm

Instructions to candidates:

1) Question one is COMPULSORY. Candidates should attempt ALL parts of question one and any other two questions.
2) This paper is worth 100 marks. Question one is worth 40 marks. All other questions are worth 30 marks each.

DO NOT TURN OVER THIS PAGE UNTIL YOU ARE TOLD TO DO SO

Question 1: Compulsory
Answer all ten parts. Each part is worth 4 marks.

a) Outline and briefly describe a methodology for Data Mining with which you are familiar.

b) Measurements can be classified as: *Ratio scale; Nominal scale; Interval scale; Ordinal scale and Categorical scale*. Give examples for each of these five types of measurement, and order them in terms of information content. Why is it important to know the category of the data before mining?

c) *Bias* and *precision* are two categories of errors that may be present in a dataset. Explain what these errors are in the context of data mining.

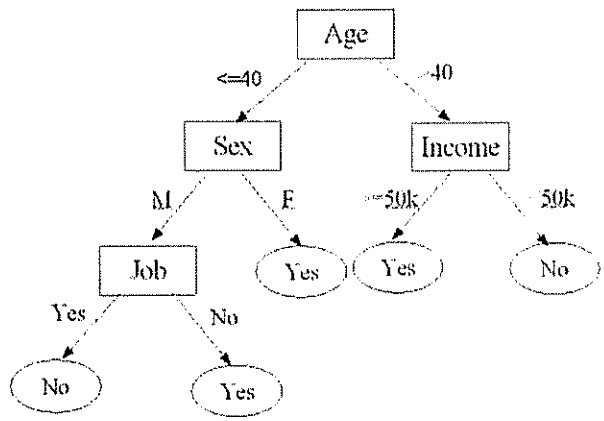
d) Give reasons why a data miner may need to reduce the dimensionality of a dataset.

e) Evaluate the following techniques for handling missing values in a dataset.
   i. Ignore the missing values
   ii. Omit columns containing missing values
   iii. Omit rows containing missing values
   iv. Use *Mean* or *Standard Deviation* functions to fill missing values

f) What is *noise*? Explain, with the aid of an example, one technique for reducing the impact of noise on a dataset.

g) Explain the role of a *dependent variable (*or *target variable)* when mining a data set.

h) Given the following decision tree, generate all classification rules from the tree. Note there are two classes, Yes and No.



i) What is the problem of *overfitting* in classification?

j) What is A*ssociation Analysis*? Include in your answer the significance of *support* and *confidence* values with respect to association analysis.

Question 2:

a) Discuss the need for the human direction of data mining. Describe the possible consequences of relying on completely automatic data analysis tools.

(4 marks)

b) In Data Understanding a miner may perform Exploratory Data Analysis (EDA). What is EDA? For each of the following descriptive methods explain how it may be applied in EDA and state whether it may be applied to categorical data, continuous data or both.

- Bar chart
- Histograms
- Cross-tabulation
- Scatter plots

(8 marks)

c) The table below shows an excerpt from a dataset.

id	Age	Gender	Marital Status	Occupation	Income	Savings	Credit risk
001	C	F	M	Data Miner	40,500	20,500	Good
002	35		W	Software Engineer	30,000	14,000	Good
003		M	S	Marketing Consultant	35000000	15,500	Bad
004	37	M	S	Teacher	-30,000	8,500	Bad
005	35	F	D	Software Developer	32,450	7,000	Good

(i) Discuss attribute by attribute any errors, anomalies or observations (if any) in the above dataset.

(4 marks)

(ii) What Data Preparation (data preprocessing) steps would you apply to this dataset before modeling (assume that data modeling will be carried out using a Neural Network). Outline the techniques you could use in each case.

(8 marks)

(iii) Using min-max normalization, normalise the data for the variable 'Savings' to values between 0 and 1. How might this technique for normalization be adapted to deal with out of range values?

(6 marks)

Question 3:

a) Is mining a sample of data as effective as mining the entire dataset. Discuss your answer.

(6 marks)

b) In data modeling, describe the difference between a training set and a test set.

(4 marks)

c) Outline the major steps of the C5.0 algorithm for Decision Tree classification. What concept does this algorithm use to decide which attribute to use to start splitting the data?

(8 marks)

d) Consider the following set of training examples:

Age	Income	Student	Credit rating	Buys Computer
<=30	High	No	Fair	No
<=30	High	No	excellent	No
31..40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	excellent	No
31..40	Low	Yes	excellent	Yes
<=30	Medium	No	Fair	No
<=30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<=30	Medium	Yes	excellent	Yes
31..40	Medium	No	excellent	Yes
31..40	High	Yes	Fair	No
>40	Medium	No	excellent	Yes

(i) What is the entropy of this collection of training examples with respect to the target variable classification?

(4 marks)

(ii) What is the information gain of *student* relative to these training examples?

$$E(x) = -\sum_{j} p_j \log_2(p_j) \qquad \text{eqn.1}$$

$$E_S(T) = \sum_{i=1}^{k} P_i E_S(T_i) \qquad \text{eqn.2}$$

$$I(S) = E(T) - E_S(T) \qquad \text{eqn.3}$$

(8 marks)

Question 4:

a) Explain the difference between supervised and unsupervised learning. Which data mining methods are associated with supervised learning? Which are associated with unsupervised learning?

(5 marks)

b) Explain with the aid of a diagram how a Back-Propagation Artificial Neural Network (BPANN) works. Your answer should refer to the following:

(i) What a BPANN is

(3 marks)

(ii) The characteristics of a ANN: *layered*, *feedforward*, *completely connected*

(3 marks)

(iii) How a neural network functions non-linearly

(2 marks)

(iv) How a BPANN trains itself

(4 marks)

(v) What happens in the hidden layer as the network trains itself

(4 marks)

c) The following confusion matrices present the classification results that were produced using a Decision Tree and Neural Network in the prediction of churn for a telecommunications company. Which data mining method performs the best for the particular task? Discuss each evaluation measure you use to assess the performance of each model.

(9 marks)

Table 1 Decision Tree classification

Classified As:		Correct
Positive	**Negative**	
40	5	**Positive**
15	205	**Negative**

Table 2 Neural Network classification

Classified As:		Correct
Positive	**Negative**	
50	10	**Positive**
5	200	**Negative**