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
(Repeat Paper)

Internal Examiner(s): Laura Keyes

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

Thursday 30th August 2007

1.00pm – 3.00pm

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 the steps involved in the Data Mining process when using the CRISP-DM methodology.

b) In what ways can the distribution of values in a variable be illustrated?

c) Why would a miner perform Exploratory Data Analysis (EDA) and data preparation on a dataset? Why not proceed directly into the modeling phase?

d) Describe two problems that arise when there are too many dimensions in a dataset.

e) Explain why you might normalize the values of a variable so that all values are in the range [0.0, 1.0].

f) Suppose the minimum and maximum values for a variable *age* are 15 and 70, respectively. We would like to map *age* to the range [0.0, 1.0]. Normalise a value of 35 for *age.*

g) When modeling data, explain why both a training set and test set are used.

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

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

j) Briefly describe the following:
- Clustering
- Classification.

Question 2:

a) 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.

(3 marks)

b) Explain what is meant by the term *outliers*. Illustrate your answer with an example.

(4 marks)

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

id	Age	Gender	Occupation	Income	Savings	Credit risk
001	47	C	Data Miner	40,500	20,500	Good
002	35	M	Software Engineer	30,000	14,000	Good
003		M	Marketing Consultant	35000000	15,500	Bad
004	37	M	Teacher	-30,000	8,500	Bad
005	35	F	Software Developer	32,450	7,000	Good

(i) For each column in the sample dataset above determine which variables are *categorical* and which are *continuous*. Why is it important to know the category of data for data mining?

(4 marks)

(ii) Discuss attribute by attribute any errors or anomalies in the above dataset.

(4 marks)

(iii) Explain how you would determine if a *sample* of data for the variable 'Income' was representative of the population for that variable.

(6 marks)

d) What is the primary objective of filling missing values in a dataset? Explain <u>two</u> techniques you could use to fill the missing values in a dataset. One technique should preserve the variables variability; the other technique should attempt to preserve the relationships between variables.

(9 marks)

Question 3:

a) Is data mining just a simple transformation of technology developed from databases, statistics, machine leaning and pattern recognition?

(3 marks)

b) 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)

c) What is Market Basket Analysis? What is the significance of support and confidence values with respect to market basket analysis?

(6 marks)

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

i. What a BPANN is *(3 marks)*

ii. The purpose of Input, Output and Hidden layers *(4 marks)*

iii. How the neural network trains itself *(4 marks)*

iii. What is happening in the hidden layer as the network trains itself *(5 marks)*

Question 4:

a) 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?

(9 marks)

b) Consider the following set of training examples:

Customer	Savings	Assets	Income (1000s)	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

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

(4 marks)

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

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

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

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

(9 marks)

c) Given the classification results in the following confusion matrix, complete the classification *accuracy*, *precision* and *recall* scores.

(8 marks)

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