INSTITUTE OF TECHNOLOGY

BLANCHARDSTOWN

Year	Year 3
Semester	Semester 2
Date of Examination	**Friday 29th August 2008**
Time of Examination	**1.00pm – 3.00pm**

Prog Code	BN302	**Prog Title**	Bachelor of Science in Computing in Information Technology	**Module Code**	Comp H3024

Module Title	Data Mining (Repeat)

Internal Examiner(s): *Mr. Markus Hofmann*
External Examiner(s): *Dr Richard Studdert, Mr John Dunnion*

Instructions to candidates:

1) **Question One Section A is COMPULSORY. Candidates should attempt Question One and ANY other <u>two</u> questions in Section B**
2) **This paper is worth 100 marks. Question One is worth 40 marks and all other questions are worth 30 marks each.**
3) **Show all your work**

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

SECTION A: COMPULSORY QUESTION

Question 1: This question is compulsory **(40 marks)**
Answer **ALL** eight parts.

a) Briefly define the term 'Data Mining' and give **two** applications.

(5 marks)

b) Explain the terms *binary*, *discrete* and *continuous* in terms of data types.

(5 marks)

c) Define *Noise* and *Outliers*. Distinguish between *Noise* and *Outliers*

(5 marks)

d) Missing values are a common occurrence in data sets. (1) Outline the reasons for *missing values*. (2) Explain how *missing values* can be handled?

(5 marks)

e) Visualisation is an important concept in data mining. Briefly explain why. Draw a box plot and explain its characteristics.

(5 marks)

f) List and briefly explain each of the processes in the *CRISP-DM* Methodology.

(5 marks)

g) Explain what is meant by *Overfitting* and *Underfitting*.

(5 marks)

h) Explain the concept of *Artificial Neural Networks*.

(5 marks)

SECTION B: Answer any TWO questions

Question 2: Data & Data Exploration **(30 marks)**

a) Discuss whether or not each of the following activities is a data mining task.

(i) Dividing the customers of a company according to their gender.

(ii) Dividing the customers of a company according to their profitability.

(iii) Computing the total sales of a company.

(iv) Predicting the outcomes of tossing a (fair) pair of dice.

(v) Predicting the future stock price of a company using historical records.

(vi) Monitoring the heart rate of a patient for abnormalities.

(6 marks)

b) Many sciences rely on observation instead of (or in addition to) designed experiments. Compare the data quality issues involved in observational science with those of experimental science and data mining.

(8 marks)

c) Explain the concept of *OLAP*. Describe the steps that are necessary to create a multidimensional array.

(8 marks)

d) Summary statistics are important in the data mining domain.

(i) Explain why.

(ii) Outline and briefly explain six *summary statistics*

(8 marks)

Question 3: Classification **(30 marks)**

a) Draw the full decision tree for the parity function of four Boolean attributes, A, B, C, and D. Is it possible to simplify the tree?

A	B	C	D	Class
T	T	T	T	T
T	T	T	F	F
T	T	F	T	F
T	T	F	F	T
T	F	T	T	F
T	F	T	F	T
T	F	F	T	T
T	F	F	F	F
F	T	T	T	F
F	T	T	F	T
F	T	F	T	T
F	T	F	F	F
F	F	T	T	T
F	F	T	F	F
F	F	F	T	F
F	F	F	F	T

(6 marks)

b) Explain the basic concept of Hunt's Algorithm

(6 marks)

c) When creating decision trees the splitting strategy is crucial for successful classification. Using examples, outline the concept of *multi-way* and *binary split* on *nominal attributes*, *ordinal attributes* and *continuous* attributes.

(6 marks)

d) Outline **four** advantages of decision trees.

(4 marks)

e) Performance evaluation of the predictive capability of a classification algorithm is a large aspect of data mining models. Describe in detail how performance can be evaluated using metrics.

(8 marks)

Question 4: Clustering & Evaluation **(30 marks)**

a) Differentiate between Partitional Clustering and Hierarchical Clustering.

(6 marks)

b) List and briefly explain five different types of clusters.

(5 marks)

c) Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is

(i) low for all clusters?
(ii) Low for just one cluster?
(iii) High for all clusters?
(iv) High for just one cluster?
(v) How could you use the per variable SSE information to improve your clustering?

(10 marks)

d) The following confusion matrices present the classification results that were produced using a decision tree and a neural network in the prediction of churn for an insurance company. Which data mining algorithm performs the best for this particular task? Discuss each evaluation measure you use to assess the performance of each model.

Decision Tree	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Neural Network	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

(9 marks)