INSTITUTE OF TECHNOLOGY

BLANCHARDSTOWN

Year	Year 3
Semester	Semester 2
Date of Examination	**Friday 23rd May 2008**
Time of Examination	**12.30pm – 2.30pm**

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

Module Title	Data Mining

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

nstructions 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) Data Mining is an important concept in many industry sectors. Give **two** examples of applications and briefly outline the term *Data Mining.*

(5 marks)

b) Explain the terms *Nominal* and *Ordinal* as well as *Interval* and *Ratio* in terms of data types.

(5 marks)

c) Briefly explain the terms *Precision*, *Accuracy* and *Bias.*

(5 marks)

d) What is the *Curse of Dimensionality*? Explain how this can be addressed.

(5 marks)

e) Explain the characteristics of *Contour Plots* and *Scatter Plots.*

(5 marks)

f) Briefly explain the importance of the *CRISP-DM* methodology.

(5 marks)

g) List *four* advantages of *Rule Based Classifiers.*

(5 marks)

h) Explain the concept of *Association Rules.* Describe the *Two Step Approach.*

(5 marks)

SECTION B: Answer any TWO questions

Question 2: **(30 marks)**

a) Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

(i) Time in terms of AM or PM.

(ii) Brightness as measured by a light meter.

(iii) Brightness as measured by people's judgments.

(iv) Angles as measured in degrees between 0° and 360°.

(v) Bronze, Silver, and Gold medals as awarded at the Olympics.

(vi) Height above sea level.

(vii) Number of patients in a hospital.

(7 marks)

b) Define the terms *Noise* and *Outliers*. Distinguish between *Noise* and *Outliers*. Be sure to consider the following questions.

(i) Is noise ever interesting or desirable? Are Outliers ever interesting or desirable?

(ii) Can noise objects be outliers?

(iii) Are noise objects always outliers?

(iv) Are outliers always noise objects?

(v) Can noise make a typical value into an unusual one, or vice versa?

(7 marks)

c) Calculate the Euclidian Distance of the points shown in the table below. What needs to be done if the scales of attributes differ?

p1	1	4
p2	2	3
p3	3	3
p4	7	1

Note: $dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$

(8 marks)

d) List and explain **two** statistics that describe Measures of Location and **two** statistics that describe Measures of Spread.

(8 marks)

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

a) Describe the concept of a Rule Based Classifier. Further outline the applications of such classification algorithms.

(6 marks)

b) Consider a binary classification problem with the following set of attributes and attribute values:

- Air Conditioner = {Working, Broken}
- Engine = {Good, Bad}
- Mileage = {High, Medium, Low}
- Rust = {Yes, No}

Suppose a rule-based classifier produces the following rule set:

Mileage = High → Value = Low
Mileage = Low → Value = High
Air Conditioner = Working, Engine = Good → Value = High
Air Conditioner = Working, Engine = Bad → Value = Low
Air Conditioner = Broken → Value = Low

Justify all answers:

(i) Are the rules mutually exclusive?

(ii) Is the rule set exhaustive?

(iii) Is ordering needed for this set of rules?

(iv) Do you need a default class for the rule set?

(8 marks)

c) Outline **five** advantages of Rule Based Classifiers.

(5 marks)

d) Describe Nearest Neighbour Classification:

(i) Basic definition **(*2 marks*)**

(ii) Which three things are required **(*6 marks*)**

(iii) How can the value of k be chosen? **(*3 marks*)**

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

a) Define Clustering and list **three** applications.

(5 marks)

b) Explain the K-means clustering algorithm. Use pseudo code to define the algorithm.

(5 marks)

c) Describe a method that can be used to evaluate k-means clusters.

(5 marks)

d) Consider the training examples shown below for a binary classification problem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

(i) Compute the Gini index for the overall collection of training examples.

(ii) Compute the Gini index for the Customer ID attribute.

(iii) Compute the Gini index for the Gender attribute.

(iv) Which attribute is better Car Type (0.1625) or Shirt Size (Gini = 0.4914)? Justify your answer.

(v) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

Note: $\boxed{GINI(t) = 1 - \sum_j [p(j \mid t)]^2}$

(15 marks)