INSTITUTE OF TECHNOLOGY

BLANCHARDSTOWN

Year	Year 3
Semester	Semester 2
Date of Examination	Thursday 21st May 2009
Time of Examination	12.30pm - 2.30pm

Prog Code	BN302	**Prog Title**	Bachelor of Science in Computing in Information Technology	**Module Code**	Comp H3024
Prog Code	BN013	**Prog Title**	Bachelor of Science in Computing in Information Technology	**Module Code**	Comp H3024
Prog Code	BN104	**Prog Title**	Bachelor of Science (Honours) in Computing	**Module Code**	Comp H3024
Prog Code	BN997	**Prog Title**	Erasmus foreign students	**Module Code**	Comp H3024

Module Title	Data Mining

Internal Examiner(s): *Ms. Geraldine Gray*
Ms. Laura Keyes
Dr. 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) **Define** the term Data Mining. Provide **two** examples of data mining applications.

(5 marks)

b) What are **box plots**? Draw a box plot and explain all of its features.

(5 marks)

c) **Differentiate** between Discrete, and Continuous data. Provide **two** examples for each category and **outline** the main characteristics.

(5 marks)

d) **Explain** how the **classification** process takes place in a data mining project?

(5 marks)

e) As part of rule based reasoning the terms **rule coverage** and **accuracy** are used. Explain **each** of the two terms and provide an example.

(5 marks)

f) *The **CRISP-DM*** methodology is the de facto standard for data mining projects. Outline **each** of the processes and give at least one practical example for **each** process.

(5 marks)

g) Outline the difference between **partitional** and **hierarchical clustering.** Give **one** example each.

(5 marks)

h) **k** Nearest Neighbour is a classification technique. **Outline** the reasoning and pitfalls when choosing the correct value for **k**.

(5 marks)

SECTION B: Answer any TWO questions

Question 2: Data preparation and Exploration (30 marks)

ExampleSet (3333 examples, 1 special attribute, 20 regular attributes)

Type	Name	Value Type	Statistics	Range	Unknown
label	Churn?	nominal	mode = False. (2850)	False. (2850), True. (48	0
regular	State	polynominal	mode = WV (106)	KS (70), OH (78), NJ (68	0
regular	Account Length	integer	avg = 101.060 +/- 39	[1.000 ; 243.000]	3
regular	Area Code	integer	avg = 437.182 +/- 42	[408.000 ; 510.000]	0
regular	Phone	nominal	mode = 382-4657 (1)	382-4657 (1), 371-71	1
regular	Int'l Plan	binominal	mode = no (3010)	no (3010), yes (323)	0
regular	VMail Plan	binominal	mode = no (2411)	yes (922), no (2411)	0
regular	VMail Message	integer	avg = 8.099 +/- 13.6	[0.000 ; 51.000]	0
regular	Day Mins	real	avg = 179.749 +/- 54	[0.000 ; 350.800]	1
regular	Day Calls	integer	avg = 100.436 +/- 20	[0.000 ; 165.000]	0

Figure 1. Dataset representing customer phone call patterns

a) Given the meta data displayed above, recommend THREE pre-processing techniques that would be appropriate to use on this dataset. In your answer, explain both why that technique would be advisable, and what would be the effect of applying the technique.

12 marks

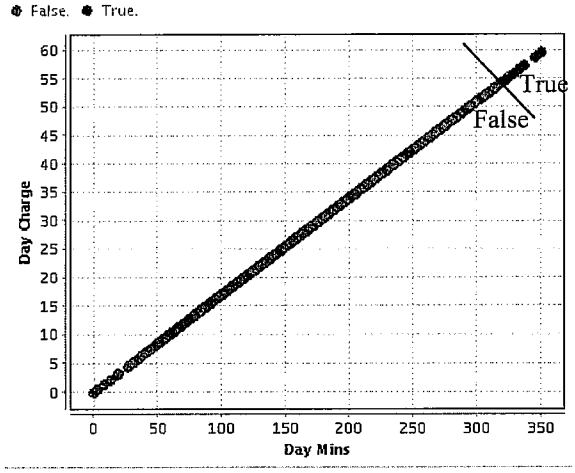
b) Explain each of the data types listed in the meta data above. **5 marks**

c) Briefly explain the role of the data exploration phase of a data mining process

3 marks

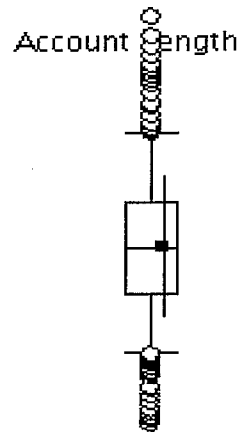
d) Interpret the following three plots generated from the dataset above.

(i) Scatter plot of **day minutes** versus **day charge**, overlaid with the binary class label, **churn** (all true values of the class label are on the top right corner).

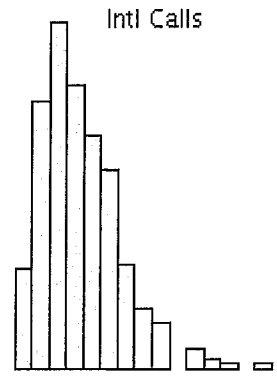


[continued on the next page]

(ii) Box plot on **account length**



(iii) Histogram of **International calls made**.



10 marks

Question 3. Classification (30 marks)

accuracy: 61.20% +/- 6.14% (mikro: 61.20%)

	true one	true two	true three	true four	class precision
pred. one	156	23	19	51	62.65%
pred. two	32	55	0	0	63.22%
pred. three	20	0	24	0	54.55%
pred. four	48	1	0	71	59.17%
class recall	60.94%	69.62%	55.81%	58.20%	

a) Interpret the confusion matrix given above. The class label has four possible values, namely 'one', 'two', 'three' and 'four'. In your answer explain the individual cell entries, the class **precision** entries and the class **recall** entries. Also explain how these figures could be used to estimate the **cost of the c**lassifier.

8 marks

b)

(i) Explain how an **impurity measure** can be used to decide on spilt points in a decision tree. **4 marks**

(ii) Using the data given below, calculate an impurity measure for **student** and **gender**. **8 marks**

(iii) Based on your results in part (ii) above, advise which attribute should be used as the next split point for the data given.

2 marks

Student	Gender	Label
yes	M	high
yes	F	low
yes	M	low
no	M	high
no	F	high
no	F	high
yes	M	low

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

c) Explain what is meant by model **over-fitting**. Your answer should discuss causes of model over fitting and how the problem can be addressed when building a **decision tree**.

8 marks

Question 4. Clustering **(30 marks)**

a) Explain why the notion of a **cluster** can be ambiguous.

5 marks

b) Calculate the Euclidean distance between the following three data points:

Note:

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

	Attribute 1	Attribute 2	Attribute 4
Point a	10	2	100
Point b	15	3	70
Point c	12	7	90

12 marks

c) Explain the benefit of **normalising** attributes before calculating distances. Make reference to the data in part (b) above to illustrate you answer.

4 marks

d) Explain in detail how the **k-means clustering** algorithm works.

9 marks