INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN

Academic term	2013-14
Year of study	Year 3
Semester	SEMESTER ONE – REPEAT PAPER
Date of examination	
Time of examination	

Programme code	Programme title	Module code
BN302	Bachelor of Science in Computing in Information Technology	COMP H3027
BN013	Bachelor of Science in Computing in Information Technology	COMP H3027
BN104	Bachelor of Science (Honours) in Computing	COMP H3027

Module title	**Repeat Paper - Data Mining**

Internal Examiner(s)	**Geraldine Gray**
External Examiner(s)	Dr. Tom Lunney Mr. Michael Barrett

Instructions to candidates:

1.	**To ensure that you take the correct examination, please check that the module and programme which you are following is listed in the table above.**
2.	**Questions 1 in Section A is COMPULSORY. Candidates should attempt Questions 1, and any two of three questions in Section B.**
3.	**There are 100 marks on the papers. Question 1 is worth 40 marks. All other questions are worth 30 marks each.**
4.	**Show all your work**

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

SECTION A

Question 1: (Compulsory)

a) Give a brief explanation of four of the six phases of the CRISP-DM methodology.

(4 marks)

b) Explain the term **outlier** in the context of classification. How could you identify an outlier value?

(4 marks)

c) Discuss the importance of having a test dataset when evaluating the performance of a classification algorithm.

(4 marks)

d) Explain the term 'confusion matrix'. Illustrate your answer with an example.

(4 marks)

e) Why is ***k***-Nearest Neighbour described as a lazy classifier?

(4 marks)

f) Compare **Decision Trees** and **Support Vector Machines** in terms of the type of attributes and class label each can model, and the scalability of each.

(4 marks)

g) What does **imputing** a missing value mean? Advise when this would be an appropriate choice for handle missing values.

(4 marks)

h) Explain how **progressive sample** identifies an appropriate sample size for a dataset.

(4 marks)

i) Explain the difference between a **noise point** and a **border point** in the context of a DBScan clustering algorithm.

(4 marks)

j) Explain how to use a decision tree to evaluate the performance of a clustering algorithm.

(4 marks)

Total: 40 marks

SECTION B

Question 2:

The table below shows the meta data for a dataset of titanic passengers, and whether or not they survived. The dataset has 8 attributes, and 891 rows:

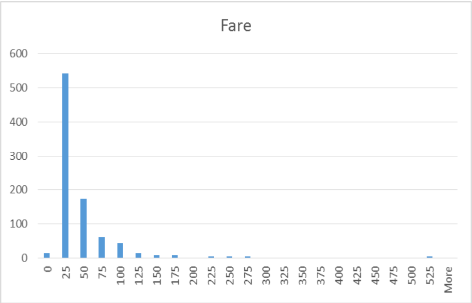
Role	Name	Data Type	Statistic	Range	Missing values
Label	Survived	Binominal	Mode=No (549)	No (549),Yes(342)	0
Regular	PClass	Integer	2.3±0.8	[1,3]	0
Regular	Name	Polynominal	Mode = Harris (1)		0
Regular	Gender	Binominal	Mode=male(577)	Male(577), Female(314)	0
Regular	Age	Real	29.7±14.5	[0,80]	177
Regular	Ticket	Polynominal	Mode=1601(4)		0
Regular	Fare	Real	32.2 ± 49.6	[0,512]	0
Regular	Cabin	Polynominal	Mode=G6(4)		687
Regular	Embarked	Polynominal	Mode=SouthHampton(644)	SouthHampton(644), Queenstown(77), Cherbourg(168)	2

a) Three of the attributes in the table above have missing values. Explain how you would handle the missing values in each case. Justify the choices you make. **(8 marks)**

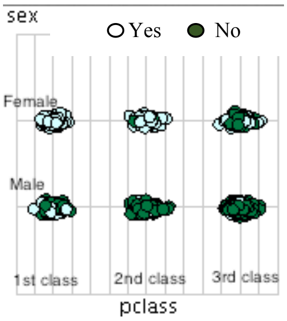
b) Discuss each of the data types in the table above with reference to how useful they are to a classification algorithm. Are there any attributes you would remove from the dataset at this point, based on the meta data? **(12 marks)**

c) Interpret each of the plots below. The first histogram is for **fare.** The scatter plot is **sex** by **passenger class (pclass)**, colour coded by the class label, **survived**, **yes** or **no**. The final histogram is for **age**.

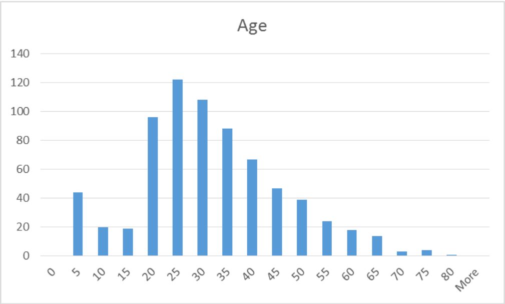
(**10 marks)**



(i)



(ii)



(iii)

Total: 30 marks

Question 3:

The dataset below is based on an online auction site such as e-Bay. The two attributes represent the duration of the auction, and whether or not the starting price was high. The class label, **good price**, determines if the item was sold for a good (high) price or not.

Start Price	**Length of Auction**	**Good price**	Note: Entropy(2,1)=0.92 Entropy(2,3)=0.97
High	Short	yes	
High	Long	yes	
Low	Long	no	
Low	Short	no	
High	Short	no	

a) Explain how an impurity measure such as **entropy** can be used to decide which attribute to select for each node on a decision tree. Use the data given above to illustrate your answer by calculating the entropy for **Start Price** and **Length of Auction**. Based on your calculations, which attribute should be at the root of the tree?

(14 marks)

b) Explain what is meant by **pre-pruning** a decision tree. If mining a dataset that is known to be noisy, would you recommend generating a full decision tree or a pruned decision tree? Explain your answer.

(6 marks)

c) Interpret the following confusion matrix from a decision tree, trained on 50 rows of the online auction dataset:

	Predicted Yes	**Predicted No**
Actual Yes	10	20
Actual No	0	20

a. What is the overall **accuracy** of the classifier? (**2 marks)**

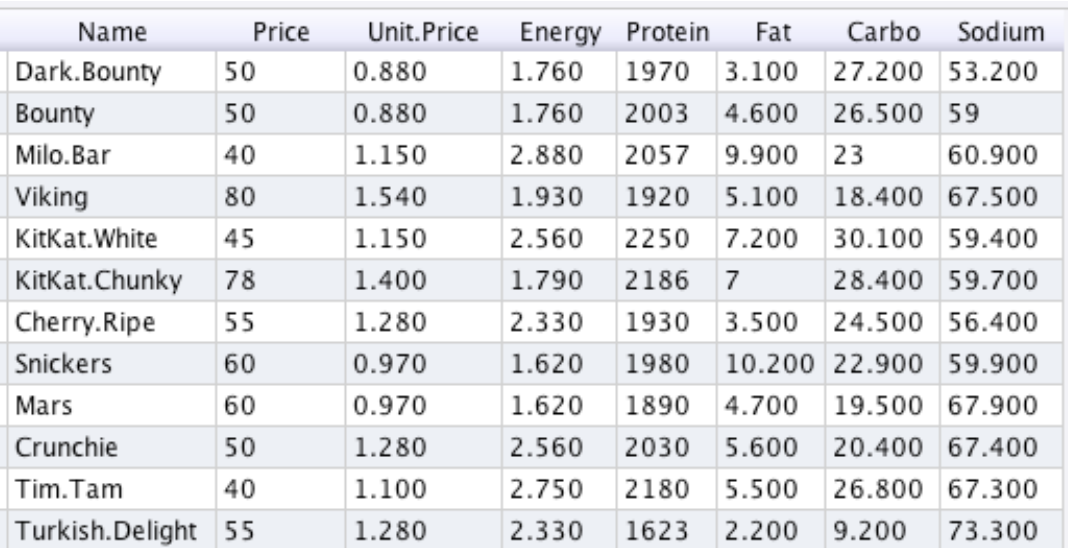
b. Calculate the **precision** for each class. Which class has the best precision? (**4 marks)**

c. Calculate the **recall** for each class. Which class has the best recall? (**4 marks)**

Total:30 marks

Question 4:

Below is a screen shot showing 12 rows of data from a dataset recording characteristics of chocolate bars. There are 8 attributes and no class label. Answer the questions below based on this dataset.



Name	Price	Unit.Price	Energy	Protein	Fat	Carbo	Sodium
Dark.Bounty	50	0.880	1.760	1970	3.100	27.200	53.200
Bounty	50	0.880	1.760	2003	4.600	26.500	59
Milo.Bar	40	1.150	2.880	2057	9.900	23	60.900
Viking	80	1.540	1.930	1920	5.100	18.400	67.500
KitKat.White	45	1.150	2.560	2250	7.200	30.100	59.400
KitKat.Chunky	78	1.400	1.790	2186	7	28.400	59.700
Cherry.Ripe	55	1.280	2.330	1930	3.500	24.500	56.400
Snickers	60	0.970	1.620	1980	10.200	22.900	59.900
Mars	60	0.970	1.620	1890	4.700	19.500	67.900
Crunchie	50	1.280	2.560	2030	5.600	20.400	67.400
Tim.Tam	40	1.100	2.750	2180	5.500	26.800	67.300
Turkish.Delight	55	1.280	2.330	1623	2.200	9.200	73.300

a) What is **unsupervised lea**rning, and why is it appropriate for this dataset?

5 marks

b) Explain in detail one preprocessing technique that should be applied to the dataset above prior to using a clustering algorithm. Justify your choice.

6 marks

c) Calculate the **Manhattan** distance between the first two rows in the chocolate dataset above. How would you include categorical attributes in a distance calculation?

7 marks

d) Recommend one algorithm you could use to identify groups of chocolate bars that have similar characteristics. Explain in detail how your chosen algorithm identifies clusters in the dataset.

12 marks

Total: 30 marks