BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
Second Semester 2011-12
Comprehensive Examination(Closed Book)

No. of Questions: 8

No. of Pages : 3

Course Number & Title : CS C415 – Data Mining **Weightage :40%**
Duration : 3Hrs. Date: 07.06.2012 Year : IV year **Marks : 40**

ANSWER ALL QUESTIONS

1. Write in detail the various methods of filling in missing values in a data set. 3 M
2. The glass data set in UCI repository has six classes. The confusion matrix of a classification model, with this data set is as below:

Actual class	Predicted Class					
	1	2	3	4	5	6
1	52	10	7	0	0	1
2	15	50	6	2	1	2
3	5	6	6	0	0	0
4	0	2	0	10	0	1
5	0	1	0	0	7	1
6	1	3	0	1	0	24

Using this write the confusion matrix for each of the following binary class problem
a) class 1 is considered as 'positive class' and all other classes as 'negative class'
b) class 2 is considered as 'positive class' and all other classes as 'negative class'
c) class 3 is considered as 'positive class' and all other classes as 'negative class'

4.5 M

3. Using the following as training data set, develop an ID3 model to classify birds with entropy and information gain measures.

Name	Eggs	Pouch	Flies	Feathers	Class
Cockatoo	Yes	No	Yes	Yes	Bird
Dugong	No	No	No	No	Mammal
Echidna	Yes	Yes	No	No	Marsupial
Emu	Yes	No	No	Yes	Bird
Kangaroo	No	Yes	No	No	Marsupial
Koala	No	Yes	No	No	Marsupial
Kookaburra	Yes	No	Yes	Yes	Bird
Owl	Yes	No	Yes	Yes	Bird
Penguin	Yes	No	No	Yes	Bird
Platypus	Yes	No	No	No	Mammal
Possum	No	Yes	No	No	Marsupial

Wombat	No	Yes	No	No	Marsupial

8 M

4. Using k-means algorithm, cluster the student objects shown in Table 3, each having four attributes namely, age, mark1, mark2 and mark3. The number of clusters are 3 and use the first three students as initial seeds. 7.5 M

Table 3. Objects for Clustering

Student	Age	Mark1	Mark2	Mark3
S1	18	73	75	57
S2	18	79	85	75
S3	23	70	70	52
S4	20	55	55	55
S5	22	85	86	87
S6	19	91	90	89
S7	20	70	65	60
S8	21	53	56	59
S9	19	82	82	60
S10	47	75	76	77

Show the detailed working after each iteration.

5. For the following transaction data set, using Apriori algorithm generate strong association rules with minsup = 30% and minconf = 80%. 7 M

The Transaction Data Set

Transaction – id	List of items
T1	Beef, Chicken, Milk
T2	Beef, Cheese
T3	Cheese, Boots
T4	Beef, Chicken, Cheese
T5	Beef, Chicken, Clothes, Cheese, Milk
T6	Chicken, Clothes, Milk
T7	Chicken, Milk, Clothes

6. Explain in detail the density based approach of outlier detection. 3 M

7. Given a document, drawn from a collection of 1000 documents, in which the four terms given in the table below occur, calculate the TFIDF weights for each one. 4 M

Term	**Frequency in current document**	**No. of documents containing the term**
Dog	2	800
Cat	10	700
Man	50	2
Woman	6	30

8. What is web mining? Write short notes on the various web mining tasks. 3 M

********** BEST OF LUCK **********

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
Second Semester 2011-12
Test – 2(Open Book)

No. of Questions: 4
No. of Pages : 1

Course Number & Title : CS C415 – Data Mining **Weightage :20%**
Duration : 50 minutes **Date:30.04.2012** **Year : IV year** **Marks : 20**

ANSWER ALL QUESTIONS

1. Table 1 below is a transaction data base with minimum support required to be 40%. Find all frequent item sets in it. Also fill up Table 2, by identifying each frequent item set as closed? or maximal? or both?

Table 1: Transaction Data Base

Transaction id	Items
100	Bread, Cheese, Juice
200	Bread, Cheese, Juice, Milk
300	Cheese, Juice, Egg
400	Bread, Juice, Milk, Egg
500	Milk, Egg

Table 2. Frequent Item sets

Frequent Item Set	Support	Closed? (yes / no)	Maximal? (yes / no)	Both? (yes / no)

(6 Marks)

2. What are the problems with the traditional association rule mining task which uses a single minimum support? (4 Marks)

3. Suggest a method for choosing the value of k, for the kNN classifier. Explain in detail. (6 Marks)

4. Consider a training set that has 100 positive examples and 400 negative examples. Find the accuracy, FOIL's information gain and the likelihood ratio statistic of the following rule

   $C \rightarrow +$ (covers 100 positive and 90 negative examples)

   Assume the initial rule

   $\phi \rightarrow +$ (covers 100 positive and 100 negative examples) (4 Marks)

********** BEST OF LUCK **********

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
Second Semester 2011-12

No. of Questions: 4
No. of Pages : 2

Test – 1(Closed Book)

Course Number & Title : CS C415 – Data Mining **Weightage :25%**
Duration : 50 minutes **Date:12.03.2012** **Year : IV year** **Marks : 25**

ANSWER ALL QUESTIONS

1. The values of **Age** attribute in a data base are 0, 4, 12, 16, 16, 18, 23, 26, 28. **Age** has to be discretized into 3 bins using a) equal width binning with bin width = 10 and b) Equal Frequency binning. Show the results in the following format 2.5 x 2 = 5 Marks

Bin #	Bin Elements	Bin Boundaries

2. The following table shows the diagnosis results of three patients for a certain disease.

Name	Fever	Cough	T1	T2	T3	T4
Han	Y	N	P	N	N	N
Kamber	Y	N	P	N	P	N
Vipin	Y	P	N	N	N	N

If Y and P are set to 1 and N to 0 (zero), Using Simple Matching Coefficient find the following

i) similarity(Han, Kamber)
ii) similarity(Han, Vipin)
iii) similarity(Vipin, Kamber) 3 x 1 = 3 Marks

3. Following table shows a list of attribute names, their type and their possible values respectively, in a certain data set. What are the possible ways of expressing test conditions on each one of the attribute during the decision tree induction process?
4 + 2 = 6 Marks

Attribute Name	Attribute Type	Possible values
Pay grade	Ordinal	A, B, C
Owns-Home	Nominal	Yes, No

4. Using ID3 algorithm with entropy and information gain, build a decision tree classification model to classify bank loan applications by assigning applications to one of three risk classes. Draw the final decision tree. 10 + 1 = 11 Marks

Owns Home?	Married	Gender	Employed	Credit Rating	Risk Class
Yes	Yes	Male	Yes	A	B
No	No	Female	Yes	A	A
Yes	Yes	Female	Yes	B	C
Yes	No	Male	No	B	B
No	Yes	Female	Yes	B	C
No	No	Female	Yes	B	A
No	No	Male	No	B	B
Yes	No	Female	Yes	A	A
No	Yes	Female	Yes	A	C
Yes	Yes	Female	Yes	A	C

********** BEST OF LUCK **********

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
Second Semester 2011-12
Quiz – 2(Closed Book)

No. of Questions: 3
No. of Pages : 3

Course Number & Title : CS C415 – Data Mining **Weightage :7%**
Duration :20 minutes **Date:14.05.2012** **Year : IV year** **Marks : 7**

NAME : **ID NO:**

ANSWER ALL QUESTIONS

1. What are the two most important disadvantages of the k-means clustering algorithm?

   Ans: 1 Mark

2. Find the WSS and BSS measures for each of the following clustering results.
   a) All 4 objects in one cluster i.e, K = 1 3.5 Marks

   Ans:

3. The following table shows the clustering results of labeled web pages as a confusion matrix. Find the entropy and purity of each cluster. Which is the best cluster?

Cluster	Science	Sports	Politics
1	250	20	10
2	20	180	80
3	30	100	210

Ans: 2.5 Marks

********** BEST OF LUCK **********

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
Second Semester 2011-12

No. of Questions: 8
No. of Pages : 3

Quiz – 1(Closed Book)

Course Number & Title : CS C415 – Data Mining **Weightage : 8%**
Duration : 20 minutes **Date:05.03.2012** **Year : IV year** **Marks : 8**

NAME: **ID No:**

---

1. Identify the type of the following attributes as one of **nominal/ordinal/ratio**

   a) Number printed on a sports person
   Ans:

   b) A set of countries
   Ans:

   c) The I, III and V persons in a race
   Ans:

   d) Pay bands in an organization denoted by A, B, C, D
   Ans:

   e) A person's weight
   Ans:

   f) The number of pizzas one can eat before fainting 0.25 x 6 = 1.5 M
   Ans:

2. Are the following data mining tasks supervised /unsupervised?

   a) Classification
      Ans:

   b) Clustering
      Ans:

   c) Z-score normalization
      Ans:

   d) Decimal Scaling
      Ans:

0.25 x 4 = 1 M

3. What is the difference between sequence data and sequential data? 1 M
   Ans:

4. What is the difference between time series data and spatial data? 1 M
   Ans:

5. What is the problem of "curse of dimensionality" for the mining algorithms? 1M
   Ans:

6. What are asymmetric attributes? For which data mining task, they are very important? 1 M

Ans:

7. What are the new ranges of the following normalization methods? 0.25 x 3 = 0.75 M

a) Min – max Ans:

b) Z-score Ans:

c) Decimal Scaling Ans:

8. What are the three ways of classifying feature selection algorithms? 0.75 M

Ans:

****