BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2012-13
Comprehensive Examination(Closed Book)

No. of Questions: 12

No. of Pages : 4

Course Number & Title: CS C415 – Data Mining **Weightage: 40%**

Duration: 3 Hrs **Date: 06-01-2013** **Year: IV year/CS** **Marks: 40**

Answer All Questions

1. Explain the process of KDD with a neat diagram. 3 Marks

2. What are symmetric and asymmetric binary variables? Give two example for each. 3 Marks

3. The weights of 15 children are given below. Using equal interval binning, discretize them into 7 bins with suitable labels. 2.5 Marks

   26.2, 25.6, 25.1, 23.3, 23.7, 23.4, 29.7, 28.5, 25.2, 21.4, 28.3, 33.4, 27.8, 24.4, 25.9

4. The following data set gives the age of people (Age) and their corresponding class (Y). Using entropy based discretization, calculate the entropy at cut point 14 and 17. Which one would result in best split? 3 Marks

Age	4	28	0	16	24	26	12	18
Y	P	N	P	N	N	N	P	P

5. Using the weather data set in table1 as the training set, develop a Naïve Bayes Classification model to be used for the classification task, "To play or don't play golf". Using the model, predict the class of the test data set in Table 2. 6 Marks

Table 1: The training data set

Day	Outlook	Temperature	Humidity	Windy	Play
1	Overcast	83	78	False	Play
2	Overcast	64	65	True	Play
3	Overcast	81	75	True	Play
4	Overcast	72	90	False	Play

5	Rain	70	96	False	Play
6	Rain	68	80	False	Play
7	Rain	75	80	False	Play
8	Rain	65	70	True	Don't play
9	Rain	71	80	True	Don't play
10	Sunny	69	70	False	Play
11	Sunny	75	70	True	Play
12	Sunny	85	85	False	Don't play
13	Sunny	80	90	True	Don't play
14	Sunny	72	95	False	Don't play

Table 2: The test set

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
Sunny	66	90	True	???

6. Four classifiers are generated for the same training set, which has 100 instances. They have the following confusion matrices. 4 Marks

		Predicted Class	
		+	-
Actual Class	+	50	10
	-	10	30

		Predicted Class	
		+	-
Actual Class	+	55	5
	-	5	35

		Predicted Class	
		+	-
Actual Class	+	40	20
	-	1	39

		Predicted Class	
		+	-
Actual Class	+	60	0
	-	20	20

It is required to identify the best classifier of the four. Explain in detail the ROC graph method used for comparing the classifiers. Also identify the best classifier using an ROC plot.

7. What is the significance of ensemble classifiers? Differentiate bagging and boosting in tabular form. When does the AdaBoost classifier fail? 3 Marks

8. Define anti monotone property? What is it also called as? A transaction data base is given in table 3. Using min_sup as 30% and Apriori algorithm identify the frequent itemsets. 4 Marks

Table 3: Transaction data base

Customer	**Items**
C1	Milk, egg, bread, chip
C2	Egg, popcorn, chip, beer
C3	Egg, bread, chip
C4	Milk, egg, bread, popcorn, chip, beer
C5	Milk, bread, beer
C6	Egg, bread, beer
C7	Milk, bread, chip

C8	Milk, egg, bread, butter, chip
C9	Milk, egg, butter, chip

9. Explain how association rule mining can be applied to categorical data? 1.5 Marks

10. The mutual distance of four objects are as given below. Using agglomerative clustering with a) complete linkage and b) average linkage as inter cluster distances, cluster the objects. Show the detailed working and the dendrogram in each case. What are the advantages of a dendrogram? 6 Marks

Objects	O1	O2	O3	O4
O1	0			
O2	1	0		
O3	11	2	0	
O4	5	3	4	0

11. What are the types of outliers? Explain the clustering based outlier detection method in detail. 3 Marks

12. What are tf, idf and tf-idf in text mining? What is their significance? 1 Mark

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

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2012-13
Test – 2(Open Book)

No. of Questions:	5
No. of Pages	: 2

Course Number & Title: CS C415 – Data Mining **Weightage: 20%**
Duration: 50 minutes **Date: 27-11-2012** **Year: IV year/CS** **Marks: 20**
ONLY PRESCRIBED TEXT BOOK AND HANDWRITTEN CLASS NOTES ARE ALLOWED

ANSWER ALL QUESTIONS SEQUENTIALLY

1. Assuming minsup count to be 2, for the transaction data base given below find
   a) All infrequent itemsets b) maximal frequent itemsets c) closed frequent itemsets and d) non-closed frequent itemsets. 5.5 Marks

TID	Items
1	s1,s3
2	s2
3	s4
4	s2,s3,s4
5	s2,s3
6	s2, s3
7	s1, s2, s3, s4
8	s1,s3
9	s1,s2,s3
10	s1, s2, s3

2. If the confidence of the rule $S \rightarrow T$ is equal to the confidence of $T \rightarrow S$, what is the support of S and support of T? Justify your answer. 3 Marks

3. Prove that support is always ≤ confidence. 4 Marks

4. Consider a multi-class problem, where each class is encoded using a 6-bit code word as in the following table. If a certain test data is classified as **1 1 0 0 1 1** by all binary classifiers, using error-correcting output coding approach to multi-class problem, what is the actual class of the test data? 3 Marks

Class	Code word
0	0 0 0 1 0 0
1	1 0 0 0 0 0
2	0 1 1 0 1 0
3	0 0 0 0 1 0
4	1 1 0 0 0 0
5	1 1 0 0 1 0

5. Following Table 1 has a set of original data of the form (X, Y), where X is the data and Y is the class label. The table also has three bootstrap samples to be used in three bagging rounds. Using the ensemble method of bagging, and 3-nearest neighbor as base classifier, find the actual class of all the test data in Table 2. 4.5 Marks

Table 1: Original Data and Bootstrap Samples

Original Data	Bootstrap Sample 1	Bootstrap Sample 2	Bootstrap Sample 3
(1.76,1)	1.76	1.76	1.73
(1.84,1)	1.76	2.01	1.69
(1.69,0)	2.01	1.76	1.73
(1.82, 1)	1.82	1.76	1.73
(2.01, 1)	2.01	1.69	1.82
(1.73,0)	1.76	1.82	1.69

Table 2: Test Data Set

Test Data	Actual Class
1.77	
1.69	

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

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2012-13
Test – 1(Closed Book)

No. of Questions:	7
No. of Pages	: 2

Course Number & Title: CS C415 – Data Mining **Weightage: 25%**
Duration: 50 minutes **Date: 09-10-2012** **Year: IV year/CS** **Marks: 25**

Answer All Questions

1. Describe any three measures that can be used as the splitting criteria at a tree node.
3 Marks
2. If there are ***m*** classes, what is the maximum possible value of the Gini index? When does it attain its maximum? 2 Marks

3. It is needed to design a decision tree model, using the **ID3** algorithm (with **entropy as impurity measure**) for the following weather data set. What is i) the entropy of the set and 2) what is the root node of the tree? 3) Draw the partial-tree with the root node.
2 + 6 + 2 = 10 Marks

[Show the detailed working]

Day	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

4. Describe the method to identify attributes that contribute too little to classification in a decision tree? (Note: without using the measures of node impurity and gain)

2 Marks

5. What is the difference between k-fold cross validation and leave-one-out cross validation?

2 Marks

6. Construct the confusion matrix for a binary classification model, where the target values are either democrat or republican. The model correctly predicted a democrat 81 times and incorrectly predicted a democrat 2 times. It correctly predicted a republican 46 times and incorrectly predicted a republican 6 times. 2 Marks

7. A data set has **P** positive examples and **N** negative examples. Draw the confusion matrix of a perfect classifier for this data set. What are recall, FP rate, precision, F1 score and accuracy of this classifier? 4 Marks

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

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2012-13

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

Quiz – 2(Closed Book)

Course Number & Title : CS C415 – Data Mining **Weightage : 7%**
Duration : 20 minutes **Date:18-12-2012** **Year : IV year/CS** **Marks :7**
NAME : **ID NO :**

Answer All Questions

1. What are the two most important disadvantages of the k-means clustering algorithm?
   Ans: 0.5 M

2. What is the centroid of the following points with 6 attributes 0.5 M

8.0	7.2	0.3	23.1	11.1	-6.1
2.0	-3.4	0.8	24.2	18.3	-5.2
-3.5	8.1	0.9	20.6	10.2	-7.3
-6.0	6.7	0.5	12.5	9.2	-8.4

Ans:

3. The sequence of merges of a hierarchical clustering process is shown as a _____. 0.25 M

4. Define a core point in a density based clustering. 0.5 M
   Ans:

5. When two points P and Q are density connected in a density based clustering?
   Ans: 0.5 M

6. What is the cluster compactness and isolation measure for the following data set with 4 points in each of the following cases: 2 M

   a) K = 1 cluster with all five points:
      Ans:

   b) K = 2 clusters, with cluster 1 = {1, 2} and cluster 2 = {4,5}
      Ans:

7. Single-link clustering algorithm can be done in _____ time and k-means can be done in _____ time. So, the _____ clustering algorithm is slower than the _____ clustering algorithm. 1 M

8. The clustering results of an animal data set are as below. 1.75 M

Cluster	Dogs	Cats	Cows
1	250	20	10
2	20	180	80
3	30	100	210

a) What is the entropy and purity of cluster 1?

Ans:

b) What is the entropy and purity of cluster 2?
Ans:

c) What is the entropy and purity of cluster 3?
Ans:

d) Which one of the three is a highly pure cluster?
Ans:

****

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2012-13

No. of Questions: 7
No. of Pages : 3

Quiz – 1(Closed Book)

Course Number & Title : CS C415 – Data Mining **Weightage : 8%**
Duration : 20 minutes **Date:25-9-2012** **Year : IV year/CS** **Marks : 8**
NAME : **ID NO :**

Answer All Questions

1. Assume you have a heterogeneous group of people to each of whom the following questions are asked. Write down the type of variable you will use and possible values as answers to each of the following: 0.5 x 7 = 3.5 Marks

a) What is your current religious affiliation?
Ans:

b) What is your family's income?
Ans:

c) What is the severity of your fever?
Ans:

d) Did the symptom completely vanish with the medication?
Ans:

e) Which languages can you read?
Ans:

f) What is the power of your eye glasses?
Ans:

g) Which T.V channels do you watch every week?
Ans:

2. What is the least informative of the scales? 0.5 M
Ans:

3. In a study to attract uninsured persons to a new insurance scheme, how will you code the insured variable? 0.5 M

Ans:

4. Are the following nominal variables? Write in one line why or why not? 0.5 x 2 = 1 M

a) Geographic position = {longitude, latitude, altitude}
Ans:

b) Tea type = {light, medium, strong}
Ans:

5. If the equal frequency binning and equal interval binning results in nearly identical partitions, what can you infer from the data? 0.5 M
Ans:

6. List any three data mining algorithms that use the concept of distances. 0.5 M
Ans:

7. The body mass index (BMI) of 15 patients are as follows:

26.2, 25.6, 25.1, 23.3, 23.7, 23.4, 29.7, 28.5, 25.2, 21.4, 28.3, 33.4, 27.8, 24.4, 25.9.

Discretize the data using equal interval binning into **3 bins**, with the labels U = Underweight, N = Normal and O = Overweight.

a) What is the bin width ? 0.5 x 3 = 1.5 M
Ans:

b) What is the cut-off of each bin?
Ans:

c) What are the transformed values?
Ans:

****