BITS PILANI, DUBAI CAMPUS
I Semester13-14
Comprehensive Examination - Closed Book

Course No. & Title: CS C415, DATA MINING | Date: 29.12.13 | Duration : 3 Hrs
Weightage: 40% | Max Marks: 40

ANSWER ALL QUESTIONS SEQUENTIALLY

1. With a neat illustration explain the process of KDD. 2 M

2. With a neat flow chart explain the feature subset selection process. 2 M

3. Online time in seconds spent by six e-commerce customers to a floral site for a firm order are [60, 90, 118, 150, 165, 170]. Transform the data into the [0, 1] range using Min-Max normalization. 2 M

4. Given X = [32, 80, 56, 75, 69, 26, 44, 50]. Normalize vector X into vector Y where each $y_i = \frac{(\bar{x} - x)}{s}$ where $\bar{x}$ is the mean of x and s is its standard deviation. 2 M

5. Calculate the information gain using entropy when splitting the following data set using A and B. Which attribute will the decision tree induction algorithm choose? 3 M

A	B	Class
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	_
F	F	_
F	F	_
T	T	_
T	F	_

6. Define the 3 measures of node impurity used in decision tree induction algorithms. 1.5 M

7. In a two-class data set, P is the number of positive examples and N is the number of

negative examples. In each of the following cases, draw the confusion matrix and also find the TP rate, FP rate, Precision, F1 Score and Accuracy.

A) Worst possible classifier.

B) A classifier that always predicts the positive class. 5 M

8. Following Table 1 is a credit card promotion database. Using Naïve Bayes classifier classify the credit card customers given in Table 2 as male/female. 8 M

Table 1: Credit Card Promotion Data Base

Magazine Promotion?	Watch Promotion?	Life Insurance Promotion?	Credit Card Insurance?	Sex?
Yes	No	No	No	Male
Yes	Yes	Yes	Yes	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	Yes	Yes	Yes	Male
No	No	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female

Table 2: Test Instances

Magazine Promotion?	Watch Promotion?	Life Insurance Promotion?	Credit Card Insurance?	Sex?
Yes	Yes	No	No	?
Yes	Missing/Unknown	No	No	?

9. Prove that support is always ⍰ confidence. 2.5 M

10. Suppose that $L_3 = \left\{ \begin{matrix} \{a,b,c\}, \{a,b,d\}, \{a,c,d\}, \{b,c,d\} \{b,c,w\}, \\ \{b,c,x\} \{p,q,r\}, \{p,q,s\}, \{p,q,t\}, \{p,r,s\}, \{q,r,s\} \end{matrix} \right\}$. Answer the following.

A) Fill up the following table

Itemset after the join step in C_4	**Subsets all in L_3 ? If No, which subsets are not in L_3 ?**

B) Which remain after the pruning step? 3 M

11. Define the entropy and purity measures of a cluster and a clustering. 2 M

12. Following is the distance matrix of 6 objects. Use agglomerative single linkage clustering to group them. Show the detailed working and the dendrogram. 5 M

	A	**B**	**C**	**D**	**E**	**F**
A	0					
B	12	0				
C	6	19	0			
D	3	8	12	0		
E	25	14	5	11	0	
F	4	15	18	9	7	0

13. Write short notes on web mining and its associated tasks. 2 M

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

BITS PILANI, DUBAI CAMPUS
I Semester13-14
Test 2 – Open Book

Course No. & Title: CS C415, DATA MINING Date: 10.11.13 Duration: 50 mins
Weightage : 20% Max Marks: 20

---

1. A set of original data of the form (X, Y) where X is the data and Y is the class label is given in Table 1. The bootstrap samples for the bagging rounds are also given. Using kNN as the base classifier with k = 3, find the actual class of the test data set in Table 2. 2 Marks

Table 1: Original Data and Bootstrap Samples

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

Table 2: Test Data Set

Test Data	Actual Class
1.77	
1.69	

2. Give two suggestions with detailed explanation in each one of the following cases with respect to kNN classifiers
   a) Method to choose the value of k.
   b) Method to improve the Euclidean distance measure. 3 Marks

3. Given a data set with 10,000 text messages, the task is to predict if a text message has positive or negative opinion associated (sentiment analysis task). The data set has 250 positive messages and 9750 negative messages. A classification model for this task correctly predicts 9700 negative messages and 100 positive messages.
   a) What is its confusion matrix?
   b) What is its sensitivity and specificity in percentage? 3 Marks

4. Following is the confusion matrix of a classifier that classifies training documents into one of three categories.

Actual Class	Predicted Class		
	Sports	Music	Science
Sports	5	3	0
Music	2	3	1
Science	0	2	11

a) What is the confusion matrix for the Sports category?
b) What is the confusion matrix for the Music category?
c) What is the confusion matrix for the Science category? 5 Marks

5. Given 'fprate' and 'tprate' of a classifier
   a) What is its Euclidean distance from a perfect classifier?
   b) What is the smallest possible value of this Euclidean distance and what is 'fprate' and 'tprate' in this case? What does this signify?
   c) What is the largest possible value of this Euclidean distance and what is 'fprate' and 'tprate' in this case? What does this signify?
   d) What is the disadvantage of using Euclidean distance in this context to assess the classifier's performance? 4 Marks

6. What is model overfitting? With respect to decision tree algorithms, suggest any three criteria that can be applied to a node to determine whether or not pre-pruning should take place. 3 Marks

****

BITS PILANI, DUBAI CAMPUS

I Semester13-14

Test 1 – Closed Book

Course No. & Title: CS C415, DATA MINING Date: 13.10.13 Duration: 50 mins

Weightage : 25% Max Marks: 25

---

1. Complete the following table of similarity and dissimilarity for simple attributes . **5 Marks**

Attribute Type	Dissimilarity	Similarity

2. Given a similarity measure with values in the interval [0,1], describe a method to transform this similarity value into a dissimilarity value in the interval [0,∞]. **3 Marks**

3. Given two objects x = {1, 1, 1, 1} and y = {0, 1, 0, 0}
   a) find the Jaccard coefficient and Jaccard distance between them.
   b) Find the simple matching coefficient and the simple matching distance. **5 Marks**

4. Construct the decision tree classification model using the ID3 algorithm, for the following weather data set. **12 Marks**

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

****

Name :

ID. No :

BITS PILANI, DUBAI CAMPUS

I Semester13-14

Quiz 1 – Closed Book

Course No. & Title: CS C415, DATA MINING | Date: 01.10.13 | Duration: 20 mins

Weightage : 8% | Max Marks: 8

---

1. Identify the type of variable in each of the following case as nominal/binary/continuous 1 M
   a) Year = {regular, leap} Ans:
   b) Entertainment = {TV, music, radio, movies} Ans:
   c) Blood-pressure = {normal, abnormal} Ans:
   d) Travel time to college (measured in decimal places) Ans:

2. A disadvantage of equal interval binning is that, it is sensitive to data _____ on both sides. 0.5 M

3. An education psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each. 1.5 M
   a) How would you convert this data into a form suitable for association analysis.
      Ans:

   b) What type of attributes would you have?
      Ans:

   c) How many attributes would you have?
      Ans:

4. Differentiate traditional and non-traditional data. Give an example. 1 M
Ans:

5. Following is a mobile user's profile data set. Using min-max Normalization, transform the attributes in the given data set into the scale [0.0, 1.0] 1.5 M

User. Id.	Call duration (mins)	SMS (total)	Data Counter (MB)
1	25000	24	4
2	40000	27	5
3	55000	32	7
4	27000	25	5
5	53000	30	5

Ans:

6. Differentiate the embedded and filter approaches for feature selection. Give 1M an example.
Ans:

7. Distinguish between the local and global estimation of missing values in a data set. 0.5 M
   Ans:

8. When the 'replace by most frequent' strategy cannot be used to fill in the missing value of a categorical attribute. 1 M
   Ans:

****