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

Course No. & Title: CS C415/CS F415, DATA MINING **Duration: 3 Hrs**
Weightage: 40% **Max Marks: 40** **Date: 26.05.14**

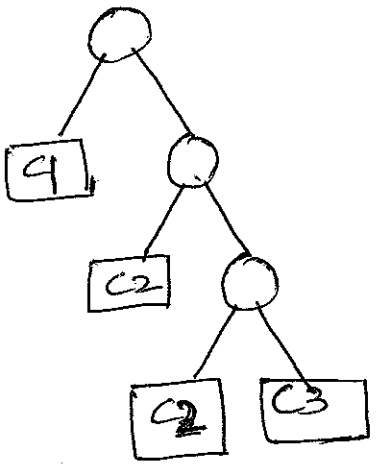
ANSWER ALL QUESTIONS SEQUENTIALLY

1. Explain in detail the following sampling techniques 5 M
   a) Simple Random Sampling b) Stratified Sampling and c) Adaptive Sampling

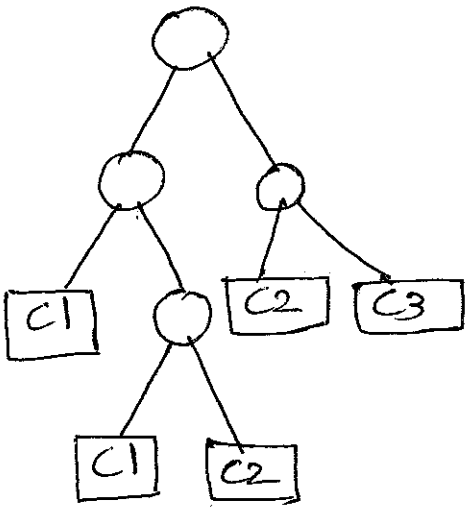
2. a) Define Z-score Normalization.
   b) Find the Z-score transformed values of 35, 36, 46, 68 and 70. 4 M

3. What is the total description length of a decision tree? For each of the decision trees A and B given below find their overall cost, assuming they are generated from a data set with 10 training instances, 16 binary attributes and C1, C2 ,C3 as class labels. Which tree is better as per the MDL principle? 5 M

Tree A with 7 errors



Tree B with 4 errors



4. Using the following training data build a decision tree classification model to classify bank loan applications into one of three risk classes. Use entropy for information gain measure 5 M

The Training Data Set

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

5. a) Define anti-monotone property of Apriori algorithm.

   b) Using min_sup as 30% and Apriori algorithm for candidate generation, find all frequent itemsets from the following transaction database. 5 M

Customer ID	Items Purchased
1	Milk, egg, bread, chips
2	Egg, popcorn, chips, beer
3	Egg, bread, chips
4	Milk, egg, bread, popcorn, chips, beer
5	Milk, bread, beer
6	Egg, bread, beer
7	Milk, bread, chips
8	Milk, egg, bread, butter, chips
9	Milk, egg, butter, chips

6. a) Using K-means clustering and Euclidean distance cluster the following 8 examples A1 to A8 into 3 clusters. Assume the initial seeds are A1, A4 and A7. Show the detailed working including the distance matrix, cluster centroids in each iteration and cluster membership at the end of each iteration.

A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8), A5 = (7,5), A6 = (6,4), A7=(1,2) and A8 = (4,9).

b) If Epsilon is 2 and minpoints is 2, write the epsilon neighborhood of each example mentioned above with respect to DBSCAN. Also write the clusters identified. 7 M

7. a) Define **Clustering Feature** with respect to BIRCH Scalable Clustering.

b) Write the algorithm to compute the SNN similarity for a database of n objects. Following table shows the two nearest neighbors of four points. Calculate the SNN similarity between each pair of points using this algorithm. 4 M

Point	First Nearest Neighbor	Second Nearest Neighbor
1	4	3
2	3	4
3	4	2
4	3	1

8. What are outliers? Write short notes on the various types of outliers? 3 M

9. Write short notes on the cross over and mutation genetic operators. 2 M

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

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

Course No. & Title: CS C415/CS F415, DATA MINING **Duration: 50 mins**
Weightage: 20% **Max Marks: 20** **Date: 23.04.14**

ANSWER ALL QUESTIONS SEQUENTIALLY

1. On a certain transaction database, the Apriori algorithm has identified the following F_3. What are the candidate 4-itemsets C_4?
   $F_3 = \{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{1,3,5\},\{2,3,4\}\}$ 2 M

2. Let $C_1, C_2, and\ C_3$ be the confidence values of the rules $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q,r\}\ and\ \{p,r\} \rightarrow \{q\}$ respectively. If it is assumed that $C_1, C_2, and\ C_3$ all have different values, what are the possible relationships that may exist among them? Which rule has the least confidence? 3 M

3. Following is a table of frequent itemsets and their respective supports in a transaction database. Complete the table by identifying each itemset as closed (yes or no), maximal(yes or no) and both closed and maximal (yes or no). 3 M

Itemset	Support	Closed?	Maximal?	Both?
{bread}	3			
{cheese}	3			
{juice}	4			
{milk}	3			
{egg}	3			
{bread, cheese}	2			
{bread, juice}	3			
{bread, milk}	2			
{cheese, juice}	3			
{juice, milk}	2			
{juice, egg}	2			
{milk, egg}	2			
{bread, cheese, juice}	2			
{bread, juice, milk}	2			

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

$\emptyset \rightarrow +$ (covers 100 positive and 100 negative examples) 2 M

5. Suggest any two methods of improving the performance of kNN classifiers in terms of the distance measure used to find kNN. 3 M

6. In the following table, original data represents points (x, y) where x is the attribute and y is the class label. Using Bagging with 1-nearest neighbor as the base classifier, classify each of the test data given below. 4 M

a) $x = 1.77$ b) $x = 1.69$

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

7. If support(A $\longrightarrow$B, C) is 75%, is the support(A $\longrightarrow$ C) greater, lesser or same? Give suitable explanation. 3 M

****

BITS PILANI, DUBAI CAMPUS
II Semester13-14
Test 1 – Closed Book

Course No. & Title: CS C415/CS F415, DATA MINING **Duration: 50 mins**
Weightage: 25% **Date: 05.03.14** **Max Marks: 25**

ANSWER ALL QUESTIONS SEQUENTIALLY

1. Draw the flow chart of the feature subset selection process. 2.5 M

2. Explain the techniques used to fill in missing values in a data set. 2.5 M

3. What are symmetric and asymmetric binary variables? Identify the following binary variables as symmetric or asymmetric. 2 + 2 = 4 M
   a) transaction type = {genuine, fraudulent}
   b) status of a message = {delivered, lost}
   c) occupancy status of an apartment = {free, occupied}
   d) gender = {male, female}

4. Given two binary vectors X = 0101010001 and Y = 0100011000, answer the following 2 M
   a) What is the *JC* between them?
   b) What is the *SMC* between them?
   c) Cosine measure is similar to which similarity measure you have studied. Justify.

5. Consider the following data set and answer the following

Instance No.	a_3	Class label
1	1.0	+
2	6.0	+
3	5.0	-
4	4.0	+
5	7.0	-
6	3.0	-
7	8.0	-
8	7.0	+
9	5.0	-

a) What are the candidate spilt points if entropy based discretization is used to discretize attribute a_3 ? 1 + 5 + 1 + 1 = 8 M
b) What is the entropy and information gain of split points 2 and 5.5?
c) Which is the best split 2 or 5.5?
d) What are the discretized features?

6. What is the new range of an attribute normalized using
   a) min-max 3 M
   b) Z-Score and
   c) Decimal Scaling

7. Differentiate the following
   a) Classification and Clustering 3 M
   b) Embedded and Filter approaches of feature selection
   c) Simple random sampling and stratified sampling

****

BITS PILANI, DUBAI CAMPUS
II Semester13-14
Quizt 1 – Closed Book

Course No. & Title: CS C415/CS F415, DATA MINING **Date: 26.03.14**
Duration: 20 mins **Weightage: 8%** **Max Marks: 8**
NAME: **ID No:**

ANSWER ALL QUESTIONS

1. Consider the following three class confusion matrix. It shows the classification results of a supervised model that uses previous voting records to determine the political party affiliation (Republican, Democrat or Independent) of Senate members. 1 Mark

	Computed Decision		
	Rep	Dem	Ind
Rep	42	2	1
Dem	5	40	3
Ind	0	3	4

a) What percent of the instances were correctly classified? **Ans**:
b) How many Democrats, Republicans and Independents are in the Senate individually?
Ans:

c) How many Republicans were classified as belonging to the Democrat party?
Ans:

d) How many Independents were classified as Republicans?
Ans:

2. If P and N are the number of positive instances and number of negative instances respectively of a training set, 2 Marks
a) Give the confusion matrix of a classifier which always predicts positive class.
Ans: what is its TPR and FPR ?

b) Give the confusion matrix of a classifier which always predicts negative class.
Ans: what is its TPR and FPR ?

3. Given (0.2, 0.7) as (fprate, tprate) of a classifier, what is the distance of this classifier from a perfect classifier? 1 Mark
Ans:

4. What is the Entropy and Gini Index measures when there are three classes with equal probabilities? 2 Marks
Ans:

5. What is the difference between cross validation and leave-one-out cross validation.
   Ans: 1 Mark

6. Write the formula to estimate the GainRatio of a split. 1 Mark
   Ans:

****