No of Pages : 4 Course Code : 09XT83

Roll No:
(To be filled in by the candidate)

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, MAY/JUNE 2011

PhD(PT) - MATHEMATICS

09XT83 DATA MINING

Time : 3 Hours **Maximum Marks : 100**

INSTRUCTIONS:

1. Answer **ALL** questions from PART - A and Answer any **4** questions from PART - B. Question under PART - C is Compulsory.

PART - A **Marks : 10 x 3 = 30**

1. Consider a naïve Bayes classifier with 3 boolean input variables X1, X2, X3 and one Boolean output, Y. How many parameters must be estimated to train such a naïve Bayes classifier?
2. What is the *apriori principle* and how is it applied to frequent itemset generation?
3. Why is Gain ratio said to be a better attribute selection measure than information gain?
4. What is the significance of feature selection in preprocessing stage of data mining. List the various techniques.
5. Consider the following training set in the 2-dimensional Euclidean space:

X	Y	class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

What is the prediction of the 3, 5, 7-nearest-neighbor classifier at the point (1,1)?

6. Differentiate between supervised and unsupervised learning with an example.
7. For the dataset shown below,

   i) Compute the support for itemsets {e}, {b,d} and {b,d,e} by treating each trasaction ID as a market basket.

No of Pages : 4 **Course Code : 09XT83**

ii) Compute the confidence for the association rules {b,d}→ {e} and {e} → {b,d}. Is confidence a symmetric measure?

Transaction ID	Items bought
0001	{a,d,e}
0024	{a,b,c,e}
0012	{a,b,d,e}
0031	{a,c,d,e}
0015	{b,c,e}
0022	{b,d,e}
0029	{c,d}
0040	{a,b,c}
0033	{a,d,e}
0038	{a,b,e}

8. Define a core object and density reachability. Is density reachability a symmetric function?
9. Given 5-dimensional numeric samples A= (1, 0, 2, 5, 3) and B= (2, 1, 0, 3, -1), find
   i.) The Euclidean distance between points
   ii.) The city block distance
   iii.) The Minkowski distance for p=3
10. What is Jaccard coefficient? What is its significance in clustering?

PART - B **Marks : 4 x 12.5 = 50**

11. a) Why is naïve Bayesian classification called "naïve"? Briefly outline the major ideas of naïve Bayesian classification. (4.5)

    b) Using the data given below perform classification using naïve Bayes classifier. Consider the following new instance to be classified (8)

    Magazine promotion =Yes
    Watch promotion = Yes
    Life insurance promotion = No
    Age = 45
    Credit card insurance = No
    Sex = ?

Magazine promotion	Watch promotion	Life insurance promotion	Age	Credit card insurance	Sex
Yes	No	No	45	No	Male
Yes	Yes	Yes	40	Yes	Female
No	No	No	42	No	Male
Yes	Yes	Yes	30	Yes	Male
Yes	No	Yes	38	No	Female
No	No	No	55	No	Female
Yes	Yes	Yes	35	Yes	Male
No	No	No	27	No	Male
Yes	No	No	43	No	Male
Yes	Yes	Yes	41	No	Female

12. Briefly explain the Apriori principle of frequent itemset mining. Consider the sample database shown below. Assume min support count =3.Show the step by step process for Apriori algorithm. A1 to A9 are the items purchased and 1 or 0 indicates whether the item is purchased(1) or not(0) for a given transaction.

	A1	A2	A3	A4	A5	A6	A7	A8	A9
1	1	0	0	0	1	1	0	1	0
2	0	1	0	1	0	0	0	1	0
3	0	0	0	1	1	0	1	0	0
4	0	0	1	0	0	0	0	0	0
5	0	0	0	0	1	1	1	0	0
6	0	1	1	1	0	0	0	0	0
7	0	1	0	0	0	1	1	0	1
8	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	1	0
10	0	0	1	1	1	0	1	0	0

13. a) Explain the fundamental difference between the Bagging and Boosting ensemble learning methods? How do these notions relate to the concept of generating a good ensemble? What are the advantages and disadvantages of each method? [8.5]

b) Explain any two methods for evaluating the accuracy of a classifier. [4]

14. The distance matrix shown below gives the distances in kilometers between some Italian cities. Use single link and complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendograms.

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

No of Pages : 4 Course Code : 09XT83

15. NASA wants to discriminate between Martians(M) and Humans(H) based on following characteristics: Green ? {N,Y}, Legs ? {2,3}, Height ? {S,T}, Smelly ? {N, Y) Our available training data is as follows:
Learn a decision tree using the ID3 algorithm and draw the tree.

	Species	Green	Legs	Height	Smelly
1)	M	N	3	S	Y
2)	M	Y	2	T	N
3)	M	Y	3	T	N
4)	M	N	2	S	Y
5)	M	Y	3	T	N
6)	H	N	2	T	Y
7)	H	N	2	S	N
8)	H	N	2	T	N
9)	H	Y	2	S	N
10)	H	N	2	T	Y

PART - C **Marks : 1 x 20 = 20**

16. What is Association rule mining? Explain using terms as support, confidence and frequent itemset. On what factors does FP Growth algorithm perform better than Apriori for Association rule mining? For the given transaction data set construct the FP tree and give a step by step process of generating the frequent item sets. (Show the conditional pattern base and conditional FP tree for every item). Min support =2.

Transaction Data Set

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

/END/

FD/RL