1198

No of Pages : 3 Course Code : 09XT83

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

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, APRIL - 2013

MSc – THEORETICAL COMPUTER SCIENCE Semester : 8

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. What are the different kinds of data used in data mining?
2. Differentiate the following terms:

   a. supervised learning

   b. unsupervised learning

   c. semi supervised learning
3. What is normalization? What is its role in classification and clustering? Give any two normalization techniques?
4. Given a dataset containing 500 positive and 500 negative instances, and classification algorithm *A*, what are the different ways to evaluate the performance of *A*?
5. What is meant by over-fitting? How does decision tree solve over-fitting?
6. How is data compression done using wavelets?
7. What is the basic property behind the Apriori algorithm? Give few techniques to improve the efficiency of Apriori algorithm?
8. Comment on 'k-nearest neighbor classifiers are lazy learners'
9. Use the method of least squares to find an equation for the prediction of students final exam grades (Y) based on midterm grade (X) in the course and predict the final exam grade of a student who received an 86 on the midterm exam.

X	72	50	81	74	94	86	59	83	65	33	88	81
Y	84	63	77	78	90	75	49	79	77	52	74	90

10. What is the need of CLARA and CLARANS clustering algorithms?

PART - B Marks : 4 x 12.5 = 50

11. a. Explain in detail about the different dissimilarity measures for different types of attributes? (7.5)

b. Find the best pair of pen-pals from the following data with the assumption that data is symmetric and asymmetric: (5)

Name	Trait1	Trait2	Trait3	Trait4
X	0	1	1	0
Y	1	1	1	1
Z	1	0	0	1

12. Discuss in detail about the following:

a. Principal Component Analysis (4)

b. Entropy and information gain for feature selection (4.5)

c. OLAP Vs OLTP (4)

13. Explain the FP tree construction algorithm for finding the frequent itemsets? Trace out your algorithm with the following set of transactions.

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

14. a) Explain in detail about box-plot and binning methods for noisy data. (7.5)

b) Draw a box-plot for the following 10 fold cross validation experimental results of different algorithms. Find any outliers which denote the result is due to the bias. (5)

Experiments Algorithms	A1	A2	A3
E1	98	67	98
E2	97	54	95
E3	98	32	91
E4	96	89	92
E5	95	76	87
E6	98	75	88
E7	97	77	85
E8	99	78	82
E9	92	79	65
E10	91	80	90

15. a) Discuss the Density Based Clustering algorithm? Cluster the following datapoints using your algorithm. Assume epsilon = 2 minpoints = 2

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

b) Write a short note on web mining and it applications. (4.5)

PART – C **Marks : 1 x 20 = 20**

16. What is the basic assumption behind the Naïve Bayes classifier? Construct a mathematical model using Naïve Bayes classifier for the following weather data. Predict the label for the following test data.

    Rainy, 71, 91, true

Outlook	Temperature	Humidity	Windy	Play?
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes

/END/

FD/RL