1070

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 / MAY - 2015

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. Define nominal, ordinal and ratio scaled variable.

2. Differentiate between the filter, wrapper and embedded approaches to feature selection.

3. Why is naïve Bayesian classification called “naïve”? 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?

4. What is antimonotone property in association analysis?

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. The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, ~hotdogs refers to the transactions that do not contain hot dogs hamburgers refers to the transactions containing hamburgers and ~hamburgers refers to the transactions that do not contain hamburgers.

	hotdogs	~ hotdogs	Σrow
hamburgers	2000	500	2500
~ hamburgers	1000	1500	2500
Σcol	3000	2000	5000

suppose that the association rule “hot dogs → hamburgers” is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong? Based on the given data is the purchase of hot dogs independent of the purchase of hamburgers? If not what kind of correlation relation ship exists between the two?

7. Define acore object and density reachability. Is density reachability a symmetric function?

8. Name one way that the clusters found by the agglomerative clustering algorithm differ to those found by the k-means clustering algorithm.

9. What is the difference between symmetric and asymmetric binary variables?

10. What is entropy and what is its significance in attribute selection? Why is gain ratio a better measure than information gain?

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

11. a) For the following group of data calculate

200,400,800,1000,2000

i. Mean and Variance

ii. Normalize the data using min max normalization

iii. In z score normalization what value should the first number 200 be transformed to?

b) The following list gives the scores of students in a midterm exam.

40, 36, 40, 40, 28, 36, 34, 34, 36, 34, 34, 38, 34, 24, 12

Draw the boxplot of the midterm scores. Write your observations.

12. What is the significance of feature selection in Data Mining? The sample dataset below contains the profile of 12 customers whose buy or no-buy responses to the new promotional email are listed below. Find the attribute which is the most dependent on the output class using chisquare method.

	Cust income	Cust uses high speed connection	Education level	Buy decision
1	Low	No	High School	No-buy
2	Low	Yes	High School	No-buy
3	Low	No	College	No-buy
4	Low	Yes	College	Buy
5	Medium	No	High School	No-buy
6	Medium	Yes	High School	No-buy
7	Medium	No	College	Buy
8	Medium	Yes	College	Buy
9	High	No	High School	No-buy
10	High	Yes	High School	Buy
11	High	No	College	Buy
12	High	Yes	College	Buy

13. a) Given the following data, discretize the age attribute into three groups: **0-25 YOUTH (Y)**; **26-40: MIDDLE (M)**; **41-99 OLD (O)**. Answer the following.

Age	Gender	Marital Status	Claims	Age
35	F	Married	LOW	M
20	F	Single	HIGH	Y
41	M	Married	LOW	O
22	M	Single	HIGH	Y
56	M	Married	LOW	O
27	F	Single	MEDIUM	M
38	M	Married	MEDIUM	M
43	M	Married	LOW	O

(i) Build the decision tree considering the Age Group, Gender, and Marital status as attributes, and Claims as the outcome.

(ii) Using the decision tree, determine the outcome for an unknown instance with: Age=35, Gender=M, and marital status = married.

14. 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?

b) Explain the methods for evaluating the accuracy of a classifier. Why is 10 fold cross validation the most preferred method for evaluation? Consider the confusion matrix of the classifier M

Data set 1		Predicted class	
		+	-
Actual	+	45	5
class	-	10	40

Calculate the various metrics for comparing classifiers based on the above data.

15. Give one advantage of hierarchical clustering over K-means clustering, and one advantage of K-means clustering over hierarchical clustering. Use single-link and complete-link agglomerative clustering to cluster the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Show the dendrograms.

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). Let the min support be 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/JU