

Total No. of Questions :6]

SEAT No. :

P4933

[Total No. of Pages :4

BE/In Sem. - 79
B.E.(Computer Engineering)
DATA MINING TECHNIQUES & APPLICATIONS
(2012 Course) (410444D) (Elective - I) (Semester - I)

Time : 1Hour]

[Max. Marks : 30

Instructions to the candidates:

- 1) Answer Q1 or Q2, Q3 or Q4, Q5 or Q6.*
- 2) Neat diagrams must be drawn wherever necessary.*
- 3) Figures to the right side indicate full marks.*
- 4) Assume suitable data, if necessary.*

Q1) a) Compare OLTP and OLAP. [4]

b) Suppose that the data for analysis includes the attribute age. The values for attribute age for the data tuples are 4, 8, 15, 21, 24, 25, 28, 34. Using the following binning methods for data smoothing, show the resultant data. [6]

- i) Bin medians.
- ii) Bin boundaries.
- iii) Bin means.

Illustrate your steps.

OR

Q2) a) With a suitable diagram explain various steps of a Knowledge Discovery in databases (KDD) process and briefly explain each step. [4]

b) What is Predictive and descriptive data mining? [2]

c) Explain in brief any two predictive and two descriptive data mining tasks you know and provide one example application of each of these tasks.[4]

P.T.O.

Q3) a) W.r.t. Association Rule Mining Define **[4]**

i) Support

ii) Confidence

b) A database has five transactions. Let min-sup=60% and min-conf=80% **[6]**

TID	Items-bought
T100	{M,O,N,K,E,Y}
T200	{D,O,N,K,E,Y}
T300	{M,A,K,E}
T400	{M,U,C,K,Y}
T500	{C,O,O,K,I,E}

Find all frequent itemsets using Apriori algorithm and generated association rules along with their values for support and confidence.

OR

Q4) a) What do you meant by market basket analysis and how it can help a grocery shopper? **[2]**

b) Is the support and Confidence of an association rule $X \rightarrow Y$ the same as that for $Y \rightarrow X$? Why or why not?

If an itemset of 'n' items is frequent, are all subsets of this frequent itemset necessarily frequent as well? **[4]**

c) Explain the Apriori algorithm for generating Association Rules. What is it's time complexity? **[4]**

Q5) a) State Bayes Theorem. **[2]**

b) Consider following training data set

[6]

Age	Income	Student	Credit_rating	Buys_Computer
<=30	High	No	Fair	No
<=30	High	No	Excellent	No
31....40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31....40	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
<=30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<=30	Medium	Yes	Excellent	Yes
31....40	Medium	No	Excellent	Yes
31....40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

Consider Buys_Computer as a Class Attribute with values yes and no. What class label would be predicted by a Bayes classifier for a data sample $X = (\text{age} \leq 30, \text{Income} = \text{medium}, \text{Student} = \text{yes}, \text{Credit_rating} = \text{Fair})$ using Naïve Bayes Classifier. Write all necessary steps with computations.

c) Explain the term 10-fold cross validation. What is the significance of it?

[2]

OR

Q6) a) Several different classifiers such as Bayes, Decision Tree, KNN are available. State various performance metrics that are used to evaluate the classifiers. Compare the above three classifiers using the metrics you stated. [4]

- b) Consider the Sunburn dataset as given below. In this dataset, based on the attributes of a person namely the Hair Color, Height, Weight and Location, the person is classified as either as a case of Sunburn or not as case of Sunburn and it is recorded in the last column Class as Yes or No respectively. If one has to construct a decision tree, what splitting attribute will be used by such algorithm at the root? Justify your answer. [6]

Name	Hair	Height	Weight	Location	Class
Sunita	Blonde	Average	Light	No	Yes
Deepak	Blonde	Tall	Average	Yes	No
Arohi	Brown	Short	Average	Yes	No
Alka	Blonde	Short	Average	No	Yes
Farukh	Red	Average	Heavy	No	Yes
Pradnya	Brown	Tall	Heavy	No	No
Jigesh	Brown	Average	Heavy	No	No
Karthik	Blonde	Short	Light	Yes	No

EEE