V/MCA/5.2/2020

Data Mining

2020

FM-70, Time-3 Hours

Answer any five

Numbers in the right hand side indicate marks

1. Explain the seven components of Data warehouse architecture with neat diagrams. 14

OR

Consider the following situation: the sales department of a super market chain wants to have a system to have strategic planning and promotion evaluations. For this, they need sales information from various stores of the super market chain. For computational analysis, they use average sales and total sales, for different product types (food, non-food) for different stores at different level: state and country, at different time periods: per year, month, quarter, semester and also by day of the week. Draw the star schema and identify the measures, dimensional attributes and hierarchies. 14

2. Suppose that the data for analysis include the attribute the frequency of stop words in documents. The values are given in increasing order:
   13,15,16,19,20,21,22,22,25,25,25,30,33,33,35,35,35,35,35,36,40,45,46,52, 70 3.5x4=14
   Apply the following methods:
   (i) Use smoothing by bins with a depth of 3
   (ii) Use min-max normalization to transform the value 35 into the range from 0.0 to 1.0
   (iii) Use normalization by decimal scaling to transform the value 35
   (iv) Use z-score normalization to transform the value 35 where the standard deviation of the above is 12.94

OR

Explain the various Data pre-processing methods with suitable examples. 14

3. A database has nine transactions with min-sup=30%, min-conf=60% 14

TID	list of items –IDs
1	a,b,e
2	b,d
3	b,c
4	a,b,d
5	a,c
6	b,c

7	a,c
8	a,b,c,e
9	a,b,c

Find all frequent itemsets using frequent itemset mining without candidate generation algorithm.

OR

Perform the FP-Tree algorithm on the above problem. 14

4. Explain how the Naïve Bayes classifier works? Find the class(X) for the following dataset by executing it in the given training set.
   X=(age≤30, income=medium, student=yes, credit rating=fair) 14
   Training set:

Age	income	student	credit rating	Class=buys-laptop
≤30	high	no	fair	no
≤30	high	no	excellent	no
31 to 40	High	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31.40	medium	no	excellent	yes
31.40	high	yes	fair	yes
>40	medium	no	excellent	no

OR

Perform Decision Tree algorithm on the above data using Information gain. 14

5. What is clustering? Briefly describe the partitioning and hierarchical methods of clustering. Group the following data points in which each point denotes the x and y coordinate of a location, into three clusters using K-means clustering. Use Euclidian distance measure. Use A1, B1, and C1 as the cluster center for each cluster. 14
   A1(2,10) A2(2,5) A3(8,4) B1(5,8) B2(7,5) b3(6,4) c1(1, 2) C2(4,9

OR

Discuss about the classification of major clustering methods in detail. 14