Qn. Set Code-1

Semester: 5th
Programme: B.Tech
Branch: CSE, IT

AUTUMN END SEMESTER EXAMINATION-2023

5th Semester B.Tech (DE-II & DE-I)

DATA MINING AND DATA WAREHOUSING

IT 3031

(For 2022 (L.E), 2021 & Previous Admitted Batches)

Time: 3 Hours Full Marks: 50

Answer any SIX questions.
Question paper consists of four SECTIONS i.e. A, B, C and D.
Section A is compulsory.
Attempt minimum one question each from Sections B, C, D.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

SECTION-A

1. Answer the following questions. [1 × 10]

(a) Differentiate between data characterization and discrimination?

(b) State why concept hierarchies are useful in data mining.

(c) Consider the 14 data values are given as follows:

2,5,6,6,7,8,8,8,9,9,10,12, 12, 15

Find the five-point summary for the above data.

(d) Consider the following set of frequent 3-itemsets: {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}. Assume that there are only five items in the dataset. List all the candidates 4-itemsets obtained by a candidate generation procedure.

(e) What is entropy? For a series {0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.50} calculate the entropy. How entropy is associated with uncertainty of information?

(f) Find the normalized data after performing min-max normalization by setting min = -1 and max = +1 for the data values 100, 200, 300, 400, 500, 600, 700.

(g) Consider the set of data X = {15, 27, 62, 35, 39, 50, 44, 44, 22, 98}. Do the data preprocessing using smoothing by bin boundary to smooth the data, using a bin of depth 3.

(h) Differentiate between agglomerative and divisive hierarchical clustering.

(i) Distinguish between classification and clustering.

(j) What is the significance of OLAP in data warehouse? Describe OLAP operations with necessary diagram/ example.

SECTION-B

2. (a) Compute the Euclidean distance, Manhattan distance and Minkowski distance for similarity/dissimilarity among the following data. [4]

	A_1	A_2	A_3
X_1	1.5	1.2	2
X_2	7	6.3	4
X_3	3.9	2.8	9
X_4	4.2	7	3

(b) Give an account on the requirements of clustering algorithms. Consider the six points: P1(0.40, 0.53), P2(0.22, 0.38), P3(0.35, 0.32), P4(0.26, 0.19), P5(0.08, 0.41) and P6(0.45, 0.30). Perform the single link hierarchical clustering and show your results by drawing a dendrogram. Compare hierarchical clusteringwith density based methods. [4]

3. (a) Explain the data cube approaches used in data mining with suitable explanation, diagram and example. [4]

(b) Explain the algorithm for constructing a decision tree from training samples. Build a Decision Tree for classification using the training data in the table given below. Divide the Height attribute into 3 ranges as [4]

follows: Less than 1.6, 1.6-1.8, and greater than 1.8. Mention about advantage and disadvantage of decision tree over any other classification techniques.

Gender	Height	Class
F	1.58	Tall
M	1.58	Medium
M	1.7	Medium
F	1.65	Tall
F	1.85	Tall
F	1.4	Short
M	1.4	Short
M	1.7	Medium
F	1.75	Tall
M	1.82	Tall
F	1.6	Tall

SECTION-C

4. (a) State and explain each steps associated in the Apriori algorithm. Apply a priori algorithm to the following data set. Assume the transactions are as follows [4]

Trans ID	Items Purchased
101	Apple, Orange, Litchi, Grapes
102	Apple, Mango
103	Mango, Grapes, Apple
104	Apple, Orange Litchi, Grapes
105	Pears, Litchi
106	Pears
107	Pears, Mango
108	Apple Orange, Strawberry, Litchi, Grapes
109	Strawberry, Grapes
110	Apple, Orange, Grapes

The set of items is {Apple, Orange, Strawberry, Litchi, Grapes, Pears, Mango}. Use 0.3 for the minimum support value. Find all association rules from all frequent itemsets generated.

(b) What are the different types of data warehouse architectures used in data mining? Explain Three-tier architecture of data warehouse with suitable example. [4]

5. (a) Assume the dataset: (2, 2), (4, 4), (5, 5), (6, 6), (9, 9), (0, 4), (4, 0) is given. K-Means algorithm runs with k=3 to cluster the dataset. Manhattan distance[d((x1, x2), (x1', x2'))= |x1-x1'| + |x2-x2'|] is used as the distance function to compute distances between centroids and objects in the dataset. Moreover, K-Mean's initial clusters C1, C2, and C3 are as follows: C1: {(2, 2), (4, 4), (6, 6)}, C2: {(0, 4), (4, 0)}, C3: {(5, 5), (9, 9)}. Now K-means is run for two iterations. Find the new clusters and their centroids in each iteration. Explain how clustering of high dimensional data are done? Compare K-means and K-medoid algorithms. [4]

(b) Describe in brief of OLTP operations performed on multidimensional data model. Differentiate between OLTP and OLAP. Explain the advantages and challenges of OLTP operation. [4]

6. (a) Explain the algorithm for constructionof *k*-nearest neighbor classifier. From the given data, using *k*-nearest neighbor classifier find the class of the data tuple (38,45), where *k* value is 3.Mention about advantage and disadvantage of *k*-nearest neighbor classifier over any other classification techniques. [4]

Sr.	A	B	Class	Sr.	A	B	Class
1	26	30	L	4	36	52	P
2	30	32	L	5	40	62	P
3	36	42	L	6	43	70	P

(b) Explain the significance of activation function in data mining. Give some examples of activation functions. Discuss how activation function is used to train the artificial neural network. Define the significance of learning rate and bias in artificial neural network. [4]

SECTION-D

7. (a) A biologist assumes that there is a linear relationship between the amount of fertilizer supplied to tomato plants and the subsequent yield of tomatoes obtained. [4]

Eight tomato plants, of the same variety, were selected at random and treated, weekly, with a solution in which x grams of fertilizer was dissolved in a fixed quantity of water. The yield, y kilograms, of tomatoes was recorded.

Calculate the equation of the least squares regression line of y on x.

Plant	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

(b) Give a brief description of Naive Bayes classification algorithm. The given data set describes two categorical input variables and a class variable that has two outputs as follows: [4]

Weather	Car	Class		Weather	Car	Class
sunny	working	go-out		sunny	working	go-out
rainy	broken	go-out		rainy	broken	go-out
sunny	working	go-out		sunny	working	go-out
sunny	working	go-out		sunny	working	go-out
sunny	working	go-out		sunny	working	go-out

Predict the class label of second instance using Naive Bayes classifier. Mention about advantage and disadvantage of Naive Bayes classifier over any other classification techniques.

8. (a) Consider an artificial neural network with back propagation model is used to train the truth table of an OR gate. This MLP has 2 units in the input layer (corresponding to two inputs of the OR gate), 2 units in the hidden layer and 2 units in the output layer (one unit represents class 0 and the other unit represents class 1). Consider that the input layer to hidden layer weights are initially set as follows: w11=0.7, w12=0.5, w21=0.7 and w22=0.6. Hidden layer to the output layer weights are initially set as follows: H11=0.8, H12=0.5, H21=0.4 and H22=0.6. Consider that the transfer functions for the hidden layer units as well as the output layer units are sigmoid function($y = 1/(1+e^{-s})$). Assume that the input layer units transfer their inputs without any change and $\eta = 0.9$. Determine the new weights after an input pattern (11) is given as the training data. The expected output is 1. [4]

(b) Construct the confusion matrix for the following actual vs predicted data and compute the following performance matrices (in term of percentages). [4]

a. Accuracy
b. Specificity
c. Precision
d. Recall

Actual	N	N	Y	Y	Y	N	Y	N	Y	N	Y	Y	N	Y
Predicted	Y	Y	N	N	Y	N	Y	N	N	Y	Y	N	N	Y

*****