USN | | | | | | | | | | |

202MC009

# Ph.D./M.S.(Engg.) by Research Degree Examination, April/May 2024
# Data Mining and Data Warehousing

Time: 3 hrs. Max. Marks: 100

**Note:** ***Answer any FIVE full questions, choosing ONE full question from each module.***

Important Note : 1. On completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages.
2. Any revealing of identification, appeal to evaluator and /or equations written eg, 42+8 = 50, will be treated as malpractice.

## Module-1

1 a. Define Data preprocessing and explain in detail sampling of data. **(10 Marks)**

b. Explain with a help of a neat diagram the steps involved in knowledge discovery process. **(10 Marks)**

**OR**

2 a. In real – world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. **(10 Marks)**

b. Suppose a group of 12 sales price records has been sorted as follows :
5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215
Partition them into three bins by each of the following methods :
i) Equal – frequency (equal – depta) partitioning ii) Equal – width partitioning
ii) Clustering **(06 Marks)**

c. Use a flowchart to summarize the procedure stepwise forward selection for attribute subset selection. **(04 Marks)**

## Module-2

3 a. What is Data warehouse? With the help of a neat diagram explain the architecture of a data warehouse. **(10 Marks)**

b. Illustrate with examples stars, snowflakes and fact constellations schemas for multidimensional data models. **(10 Marks)**

**OR**

4 a. Illustrate with the help of examples the OLAP operations. **(10 Marks)**

b. With the help of example illustrate the difference between operational Data Base systems and Data warehouse. **(10 Marks)**

## Module-3

5 a. Why is tree pruning useful in decision induction? What is drawback of using a separate set of tuples to evaluate pricing? **(05 Marks)**

b. The following table consists of training data from an employee database. The data have been generalized for e.g "31 . . . . 35" for age represent the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age and salary given in that row.

| Department | Status | Age | Salary | Count |
|---|---|---|---|---|
| Sales | Senior | 31 . . . 35 | 46K . . . 50K | 30 |
| Sales | Junior | 26 . . . 30 | 26K . . . 30K | 40 |
| Sales | Junior | 31. . . 35 | 31K . . . 35K | 40 |
| Systems | Junior | 21. . . 25 | 46K . . . 50K | 20 |
| Systems | Senior | 31. . . 35 | 66K . . . 70K | 5 |
| Systems | Junior | 26. . . 30 | 46K . . . 50K | 3 |
| Systems | Senior | 41. . . 45 | 66K . . . 70K | 3 |
| Marketing | Senior | 36. . . 40 | 46K . . . 50K | 10 |
| Marketing | Junior | 31. . . 35 | 41K . . . 45K | 4 |
| Secretary | Senior | 46. . . 50 | 36K . . . 40K | 4 |
| Secretary | Junior | 26. . . 30 | 26K . . . 30K | 6 |

VTUEXAMS@LIBRARY

Let status be the class label attribute

i) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple

ii) Use you algorithm to construct a decision tree from the given data

iii) Given a data tuple having the values "system", "26 . . . . 30" and "46 . . . 50K" for the attributes, department, age and salary, respectively, what would a naïve Bayesian classification of the status for tuple be? **(10 Marks)**

c. What is Boosting? State why it may improve the accuracy of decision tree induction. **(05 Marks)**

**OR**

**6** a. Consider the data tuples in the below table it is sorted by decreasing probability value, as returned by a classifier. For each tuple, compute the values for the member of truce positive (TP) False Positives (FP), True Negative (TN) and False Negative (FN). Compute the true positive rate (TPR) and False Positive Rate (FPR), plot the ROC curve for the data :

| Tuple | Class | Probability |
|---|---|---|
| 1 | P | 0.95 |
| 2 | N | 0.85 |
| 3 | P | 0.78 |
| 4 | P | 0.66 |
| 5 | N | 0.60 |
| 6 | P | 0.55 |
| 7 | N | 0.53 |
| 8 | N | 0.52 |
| 9 | N | 0.51 |
| 10 | P | 0.40 |

**(10 Marks)**

b. List and explain the different technique for improving the classification accurancy. **(10 Marks)**

## Module-4

**7** a. Analyze the requirements of clustering in data mining. **(10 Marks)**

b. How can you evaluate whether the clustering results are good. **(05 Marks)**

c. Write the pseudo-code of DBSCAN algorithm. **(05 Marks)**

**OR**

**8** a. Explain K-means algorithm for clustering and analyze its drawbacks. **(08 Marks)**

b. Create a Dandrogram representation for data objects {a, b, c, d, e}. **(05 Marks)**

c. Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are $A_1 = (2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1 = (7, 5)$, $B_2 = (7, 5)$, $B_3 = (6, 4)$, $C_1(1, 2)$, $C_2 = (4, 9)$.

The distance function is Euclidean distance. Suppose initially we assign $A_1$, $B_1$ and $C_1$ as the centre of each cluster, respectively. Use K-means algorithm to show only

i) The three cluster center centers after the first round of execution

ii) The final three clusters. **(07 Marks)**

## Module-5

**9** a. Discuss in brief the complex data types for mining. **(06 Marks)**

b. What are the four major components to characterize time series data? **(04 Marks)**

c. Explain the major themes in mining graph and networks. **(10 Marks)**

**OR**

**10** a. Discuss in detail any two domains of Data mining applications with the help of example. **(10 Marks)**

b. Explain ubiquitous and invisible data mining. **(05 Marks)**

c. Discuss the views on Data Mining Foundation. **(05 Marks)**