BE6-R4: DATA WAREHOUSING AND DATA MINING

NOTE:

1. **Answer question 1 and any FOUR from questions 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1.

a) Explain the difference between discrete and continuous data and give three examples of each.

b) Show the architecture of Multilayered Feed-Forward Neural networks with the help of a neat and labelled diagram.

c) What is the sequence of tasks required for loading and refreshing data in a data warehouse?

d) What is the difference between the ROLAP and MOLAP server architectures?

e) What is k-fold cross-validation?

f) What is partition based clustering?

g) What is Entropy? How is it related to Information gain?

(7x4)

2.

a) List and explain four common operations used in OLAP.

b) A data-warehouse for a university consists of four dimensions – student, course, semester and instructor. Two measure are maintained – count and average-grade. Average grade is the actual grade of a student for a course, semester, instructor at the lowest level, and average for a combination.

i) Draw a snowflake schema for the data-warehouse.

ii) If each dimension has five levels, how many cuboids will the cube contain?

(8+7+3)

3.

a) What is a data mart? How is it related to data warehouse?

b) What is a data cube?

c) Given the set F = {a(100), b (75), c (50), d (25), bc (50), bd (20), bcd (5)} of frequent item-sets with the support counts in bracket. List all association rules with one item on LHS that can be generated from F. Compute confidence of each rule.

(4+4+10)

4.

a) What is Naive Bayes classifier? What is the weakness of the assumption in the method?

b) How would you compute the dissimilarity/distance between objects with?

i) asymmetric Binary variables

ii) nominal variables

iii) interval scaled variables.

(9+9)

5.

a) What is normalization? Given the following set of numbers, normalize using min-max normalization.

23, 3, 67, 10, 38, 10, 45, 92, 56

b) Given the following 2-D points, find Euclidean distance between each pair and show in form of a distance matrix

P1(2, 5), P2(3, 2), P3(7,2), P4(6,2), P5(1, 1)

(8+10)

6.

a) Write an algorithm for K-Nearest Neighbour classification or Decision tree induction.

b) Formally describe the association rule mining problem with notation. How are support, confidence and lift computed?

(9+9)

7. Write short notes on **any three** of the following:

a) Text Mining

b) Spatio-Temporal Mining

c) Web usage mining

d) K-means clustering Algorithm

e) Decision Trees for Classification

(6+6+6)