M.Tech. Computer Science & Engineering CBCS Pattern Semester-I

PCSS141 / PCSS14A - Data Warehousing and Data Mining

P. Pages : 2 **GUG/W/23/10944**

Time : Three Hours * 6 0 3 3 * Max. Marks : 70

Notes :
1. Solve **any five** question.
2. All questions carry equal marks.
3. Due credit will be given to neatness and adequate dimensions.
4. Assume suitable data wherever necessary.
5. Diagrams and Chemical equation should be given wherever necessary.

1. a) Write and explain ID3 classifications algorithms? **7**

b) Draw and explain KDD process in Data Mining? **7**

2. a) Write down the application of Data Mining in Machine Learning? **7**

b) What is data modeling? Draw data model for railway system? **7**

3. a) How Data mart help to build Data Warehouse? **7**

b) What are the different OLAP operations? Explain any four operation with example? **7**

4. a) Suppose that a data warehouse for Big University consists of the four dimensions student, course, semester, and instructor and two measures count and avg grade. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. (i) Draw a snowflake schema diagram for the data warehouse. (ii) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each Big University student? **7**

b) How DW Environment is created? Explain with example? **7**

5. a) What are the different application of data mining? **7**

b) How data mining is different from data warehousing? **7**

6. a) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (i) What is the mean of the data? What is the median? (ii) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). (iii) What is the midrange of the data? (iv) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data? (v) Give the five-number summary of the data. (vi) Show a boxplot of the data. (vii) How is a quantile-quantile plot different from a quantile plot? **7**

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-depth) partitioning (ii) equal-width partitioning (iii) clustering. **7**

7. a) A database has five transactions. Let min sup = 60% and min conf = 80%. **7**

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}

i) Find all frequent item-sets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

ii) List all the strong association rules (with support s and confidence c) matching the following meta-rule, where X is a variable representing customers, and item i denotes variables representing items (e. g., "A", "B",): $\forall x \in$ transaction, buys (X, item 1) $\wedge$ buys(X, item 2) $\Rightarrow$ buys(X, item 3) [s, c].

b) Suppose that we want to select between two prediction models, M1 and M2. We have performed 10 rounds of 10-fold cross-validation on each model, where the same data partitioning in round i is used for both M1 and M2. The error rates obtained for M1 are 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5, 26.0. The error rates for M2 are 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0. comment on whether one model is significantly better than the other considering a significance level of 1%. **7**

8. a) The support vector machine is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with a large set of data tuples. Discuss how to overcome this difficulty and develop a scalable SVM algorithm for efficient SVM classification in large data sets? **7**

b) What are neural Networks? Where to use these Networks? **7**

****