Reg. No. : | | | | | | | | | | | | |

Question Paper Code : 20215

B.E./B.Tech. DEGREE EXAMINATION, MAY/JUNE 2012.

Sixth Semester

Information Technology

CS 2032/CS 701 — DATA WAREHOUSING AND DATA MINING

(Common to Seventh Semester Computer Science and Engineering)

(Regulation 2008)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. How is a data warehouse different from a database? How are they similar?

2. Why data transformation is essential in the process of knowledge discovery?

3. Define how the complex aggregation at multiple granularities is achieved using multi-feature cubes?

4. What is time series analysis?

5. List the primitives that specify a data mining task.

6. Mention the steps involved in the class comparison procedure.

7. What is correlation analysis?

8. What is naive Bayesian Classification? How is it differing from Bayesian Classification?

9. Give the categorization of major Clustering methods.

10. Distinguish between Classification and Clustering.

PART B — (5 × 16 = 80 marks)

11. (a) (i) Explain a three-tier Data Warehouse architecture with a neat sketch. (8)

(ii) Illustrate the different schemas for Multidimensional databases. (8)

Or

(b) (i) Describe the steps involved in the design and Construction of Data Warehouses. (8)

(ii) Suppose that the data for analysis include the attribute 'age'. The age values for the data tuples are : 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.

(1) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data. (4)

(2) How might you determine outliers in the data? (4)

12. (a) (i) Compare the concepts: discovery-driven cube, multi-feature cube and virtual warehouse. (8)

(ii) Suppose that a data warehouse consists of the four dimensions date, spectator, location, and game, and the two measures count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(1) Draw a star schema diagram for the data warehouse. (4)

(2) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM_Place in 2000? (4)

Or

(b) (i) What are the differences between the three main types of data warehouse usage : Information processing, Analytical processing, and Data mining? (8)

(ii) Consider the following multi-feature cube query: Grouping by all subsets of {item, region, month}, find the minimum shelf life in 2000 for each group, and the fraction of the total sales due to tuples whose price is less than $100, and whose shelf life is within 25% of the minimum shelf life, and within 50% of the minimum shelf life.

(1) Draw a multi-feature cube graph for the query. (3)

(2) Express the query in extended SQL. (3)

(3) Is this a distributive multi-feature cube? Justify. (2)

13. (a) (i) Describe the various descriptive statistical measures for data mining. (8)

(ii) What are the major issues in data mining? Explain. (8)

Or

(b) (i) What is attribute-oriented induction? Describe how this is implemented. (8)

(ii) Discuss the various issues that have to be addressed during data integration. (8)

14. (a) (i) Giving a concrete example, explain a method that performs frequent itemset mining by using the prior knowledge of frequent itemset properties. (10)

(ii) Discuss in detail the constraint based association mining. (6)

Or

(b) (i) Explain how the Bayesian Belief Networks are trained to perform classification. (8)

(ii) What is Decision Tree? Explain how classification is done using Decision Tree Induction. (8)

3 **20215**

15. (a) What is k-means algorithm? Suppose that the data mining task is to cluster the following eight points (locations) into three clusters.

$A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$.

The distance function is Euclidean distance. Suppose initially A_1, B_1, and C_1 are assigned as the center of each cluster, respectively. Use k-means algorithm to show only,

(i) the three cluster centers after the first round execution, and (8)

(ii) the final three clusters. (8)

Or

(b) Why is outlier mining important? Briefly describe the different approaches behind statistical-based outlier detection, distance-based outlier detection, and deviation based outlier detection. (16)
