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

K 3515

M.C.A. DEGREE EXAMINATION, NOVEMBER/DECEMBER 2007.

Elective

MC 1630 — DATA WAREHOUSING AND DATA MINING

(Regulation 2005)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. In the context of data mining state what hypothesis testing attempts to find?
2. List the three important issues that have to be addressed during data integration.
3. What is data cleaning?
4. What is rule induction?
5. What is classification?
6. Justify the need for pruning phase in decision tree construction.
7. What data model a data warehouse is based on?
8. What is an apex cuboid?
9. What is time series analysis?
10. What is a multi dimensional database?

PART B — (5 × 16 = 80 marks)

11. (a) (i) Data mining is a inter disciplinary field. Discuss. (6)

(ii) What is outlier mining? State its applications. (5)

(iii) What is Evolution analysis? Discuss. (5)

Or

(b) (i) State and discuss the preprocessing steps that may be applied to the data in order to help in improving the accuracy, efficiency, and scalability of the classification and prediction process. (6)

(ii) List and discuss the basic Data Mining tasks. (10)

12. (a) (i) State and discuss the strategies for data reduction. (10)

(ii) List and briefly discuss the basic well known statistical concepts that provide an abstraction and summarization of the data as a whole. (6)

Or

(b) Apply the Apriori algorithm to the following data set. Also state and discuss each step in the Apriori algorithm. (16)

Solution :

Trans ID	Items Purchased
101	milk, bread, eggs
102	milk, juice
103	juice, butter
104	milk, bread, eggs
105	coffee, eggs
106	coffee
107	coffee, juice
108	milk, bread, cookies, eggs
109	cookies, butter
110	milk, bread

The set of items is { milk, bread, cookies, eggs, butter, coffee, juice }. Use 0.2 for the minimum support value.

13. (a) (i) Discuss the issues that have to be addressed to perform Classification using Decision Trees. (12)

(ii) Define entropy. What does entropy measure? Discuss. (4)

Or

(b) (i) What is Cluster analysis? Discuss. (5)

(ii) Discuss the major differences between classification and clustering techniques. (5)

(iii) What is a dendogram? Give relevant example and discuss the same. (6)

14. (a) (i) Explain why a data warehouse is well equipped for providing the data for data mining. (6)

(ii) List and discuss the characteristics and main functions performed by the components of a data warehouse. (10)

Or

(b) Suppose that a data warehouse for a group of Universities consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg. grade. When 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. (8)

(ii) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Hamen-University student. (5)

(iii) If each dimension has five levels (including all), such as "student < major < status < university < all", how many cuboids will this cube contain (including the base and apex cuboids)? (3)

15. (a) (i) Discuss the various techniques used for examining the use of data mining on the World Wide Web. (12)

(ii) What is web personalization? Discuss. (4)

Or

(b) (i) List and discuss the specialized comparison operations spatial selection may involve. (8)

(ii) With relevant examples discuss special queries. (8)
