(Please write your Exam Roll No.) Exam Roll No. 02140044 14

# END TERM EXAMINATION

**FOURTH SEMESTER [MCA] MAY-JUNE 2016**

| Paper Code: MCA204 | Subject: Data Warehousing and Data Mining |
|---|---|
| Time: 3 Hours | Maximum Marks: 60 |

**Note: Attempt any five questions including Q no.1 which is compulsory. Select one question from each unit.**

Q1 Briefly explain 'data granularity' with the help of example. **(2x10=20)**

(i) What is ETL (Extraction/Transformation/Loading) process? Discuss in brief. c.
(ii) "Every data in Data warehouse is time stamped." Discuss.
(iii) How is "Data mining" different from the "OLAP"?
(iv) Briefly outline the major steps in decision tree classification.
(v) Differentiate between 'Operational' and 'Decision Support' systems.
(vi) Define Manhattan distance and Euclidean distance.
(vii) Explain the difference between supervised and unsupervised learning with the help of a real world example.
(viii) Discuss the different types of OLAP operations.
(ix) Discuss the measure support and confidence used in association rule mining.
(x) List the major benefits of Data mining.

## UNIT-I

Q2 (a) Discuss in details three main reasons why data warehouse modeling requires modeling techniques other than OLTP database modeling. **(7)**
(b) Every data structure in a data warehouse contains the time element. Why? **(3)**

Q3 (a) What is the difference between three main types of data warehouse usage: information processing, analytical processing, and data mining? **(6)**
(b) Discuss the motivation behind OLAP mining (OLAM). With the help of a clean diagram discuss architecture of OLAM. **(4)**

## UNIT-II

Q4 (a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe any five methods for handling this problem. **(5)**
(b) What is Apriori property? Why it is used? Discuss the Apriori algorithm for discovering frequent itemsets for mining Boolean association rules. **(5)**

Q5 (a) What are the fields in which clustering techniques are used? Mention any four fields. Discuss basic requirements of cluster analysis. **(3)**
(b) Why is outlier mining important? Briefly describe the different approaches behind statistical based outlier detection and distanced based outlier detection. **(7)**

**P.T.O.**

MCA-204

Q6 (a) Why is decision tree induction popular? Discuss over-fitting of an induced tree and two approaches to avoid over-fitting using suitable example/diagrams. **(5)**

(b) How can you use the Web as a data source for your data warehouse? What types of information can you get from the Web? **(5)**

Q7 (a) name the major phases of a data mining operation. Out of these phases, pick two and describe the types of activities in these two phases. **(5)**

(b) Explain data granularity and how it is applicable to the dta warehouse. **(5)**

**UNIT-IV**

Q8 (a) Apple any hierarchical clustering algorithm for clustering the following eight points. Determine the clusters with their elements. The distance function is Euclidean distance. **(5)**

$A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $A_4(5,8)$, $A_5(7,5)$, $A_6(6,4)$, $A_7(1,2)$, $A_8(4,9)$.

(b) Suppose we have the following points:- **(5)**

(1,1)
(2,4)
(3,4)
(5,8)
(6,2)
(7,8)

Use k-Means algorithm (k=2) to find two clusters. The distance function is Euclidean distance. Find 2 clusters using k-means clustering algorithm. Use (1, 1) and (2, 4) to form the initial clusters.

Q9 (a) Discuss naïve Bayesian classification. Why is it called "naïve" **(5)**

(b) Write short notes on **(Any Two)**:- **(2x2.5=5)**

(i) Outlier Analysis
(ii) Decision Support System
(iii) Data Marts

**********

MCA-204