

2056
B.E. (Computer Science and Engineering)
Sixth Semester
CS-605C: Data Mining and Analysis

Time allowed: 3 Hours

Max. Marks: 50

NOTE: Attempt five questions in all, including Question No. 1 (Section-A) which is compulsory and selecting two questions each from Section B-C.

x-x-x

Section -A

- Q 1(a) What are the different steps in knowledge discovery process. (10)
- (b) Why data warehouses are non-volatile.
- (c) Why support and confidence are not sufficient parameters to measure the success of Frequent Pattern mining?
- (d) How difference between two nominal variables is measured?
- (e) What are spatial databases?

Section -B

- Q2 (a) Differentiate between ROLAP and MOLAP. How HOLAP overcomes the limitations of both (5)
- (b) What do you mean by full materialization of data cubes? What are its disadvantages? How partial materialization is performed? (5)
- Q3 (a) Why do we use aggregates in data warehouse? Explain the aggregate fact table using an example. (5)
- (b) What is the use of correlation analysis of data? Explain the use of Quantile plots. (5)
- Q 4 (a) How redundancy is handled during data integration? Describe the different steps of Data Integration? (5)
- (b) Describe different OLAP operations that are performed on data warehouses? (5)

Section -C

- Q5 What are the different ways to improve the Apriori algorithm? Explain how the time complexity can be improved. Is there any effect on outcome? (10)
- Q6 (a) What is Gini Impurity? Explain the Gini Index method to select the attributes. (5)
- (b) How CLARA method improve the partitioning around medoids? How is it better than k-means algorithm? (5)
- Q7 (a) What are time series databases? Explain the mining process in time series databases. (5)
- (b) What is dissimilarity matrix? How to calculate dissimilarity between integer, Boolean, ordinal and exponential variables. (5)

x-x-x