IV B.Tech I Semester Examinations – November 2016

**DATAWAREHOUSING AND DATAMINING**

Time: **3** hours (CSE / ECM / IT) Max. Marks: **60**

# SECTION – A

(Short Answer Questions)

**Answer all ten questions 10×1M=10M**

1. A----------------is a department subset of the data warehouse that focuses on selected subjects, and thus its scope is department wide.
2. --------------------process detects errors in the data and rectifies them when possible.
3. Give an example for single and multi dimensional association rule.
4. In, --------------------each value in a bin is replaced by the mean value of the bin.
5. The -------------------- operation performs aggregation on a data cube.
6. All nonempty subsets of a -------------------- must also be frequent.
7. In -------------------- the class label of each training tuple *is* provided.
8. The topmost node in a decision tree is called the --------------------.
9. Clustering is used in -------------------- applications.
10. A tree structure called a --------------------is commonly used to represent the process of hierarchical clustering**.**

**SECTION – B**

**Answer all five questions 5×2M= 10M**

1. Write about uniform minimum support count in multi level association rules.
2. Define support and confidence.
3. Differentiate classification and prediction.
4. Describe the features of partition based clustering algorithms.
5. What is dimensionality reduction?

**SECTION – C**

**Answer all four questions 4×5M = 20M**

1. Define each of the following data mining functionalities: characterization, discrimination, association, correlation analysis and classification

**(OR)**

1. Describe the steps in the process of knowledge discovery.
2. How data mining systems can be classified.

**(OR)**

1. How is a data warehousedifferent from a *database*? How are they similar?
2. Describe the general strategies used for Data Cube Computation

**(OR)**

1. Describe the Apriori Algorithm for finding the frequent itemsets.
2. Write short notes on K-means clustering.

**(OR)**

1. Explain with an example density based clustering methods.

**SECTION – D**

**Answer all two questions 2×10M= 20M**

1. 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.

(a) What is the *mean* of the data? What is the *median*?

(b) What is the *mode* of the data? Comment on the data’s modality (i.e., bimodal, trimodal).

(c) What is the *midrange* of the data?

(d) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of the data?

(e) Give the *five-number summary* of the data.

**(OR)**

25. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distancebetween the two objects.

(b) Compute the Manhattan distance between the two objects.

(c) Compute the Minkowski distancebetween the two objects, using *q =*3.

26. The following table consists of training data from an employee database.

The data have been generalized. For example, “31… 35” for *age* represents the age range of 31to 35.



**Let status be the class label.**

Given a test data sample with values “systems,” “26. . . 30” and “46–50K” for the attributes department, age, and salary, respectively, what would be the class label (status) for the test sample using naive Bayesian classification?

**(OR)**

27. Describe the classification of linearly separable data using Support Vector Machines.