No. of Printed Pages : **4**

Sl. No.

# BE6-R4 : DATA WAREHOUSING AND DATA MINING

**NOTE :**

1. **Answer question 1 and any FOUR questions from 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

**Total Time : 3 Hours** **Total Marks : 100**

**1.** (a) Compare and Contrast data warehouse and database ?

(b) Describe in brief the role of statistics in data mining ?

(c) Data reduction is an important feature for large size/dimensionality of data. Discuss the important techniques used for reducing data in data mining.

(d) What is KDD in Data Mining and mention the main steps involved in typical KDD process ?

(e) What is Hypothesis Testing ? How it is generated ? Explain with an example

(f) Define metadata, explain why metadata necessary in a data warehouses and give one example of metadata.

(g) Differentiate between Web Content Mining and Web Usage Mining.

**[7x4]**

**2.** (a) What is the difference between Online Transaction (OLTP) and Data Warehousing System ? Explain with example.

(b) 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, 22, 22, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 35, 35, 36, 40, 45, 46, 52, 55.

(i) What is the mean of the data ? What is the median ?

(ii) What is the mode of the data ? Comment on the data's modality.

(iii) What is the midrange of the data ?

(c) What is the difference between data cleaning and data transformation ? Describe the steps used to clean data.

**[6+6+6]**

**3.** (a) Density-based clustering method is a popular algorithm for connected regions with sufficiently high density. Explain the algorithm.

(b) Define OLAP and discuss its applications in brief.

(c) Discuss the following Web mining techniques :

(i) Web Content Mining

(ii) Web Usage Mining

(iii) Web Structure Mining

**[6+6+6]**

**4.** (a) Apply K-Medoids clustering to solve the following problem. Let k = 2 and randomly selected two medoids are: C1 (4, 5) and C2 (8, 5) .

| S. No. | X | Y |
|---|---|---|
| 0 | 8 | 7 |
| 1 | 3 | 7 |
| 2 | 4 | 9 |
| 3 | 9 | 6 |
| 4 | 8 | 5 |
| 5 | 5 | 8 |
| 6 | 7 | 3 |
| 7 | 8 | 4 |
| 8 | 7 | 5 |
| 9 | 4 | 5 |

(b) Apply Agglomerative hierarchical clustering method to draw the dendrogram to cluster the following data points. Use Manhattan distance and single linkage.

A (1, 1), B(1, 3), C(3, 7), D(4, 9) & E(7,8)

**[9+9]**

**5.** (a) What is multiple regression ? How it is different from linear regression ? Discuss some of the applications where multiple linear regression can be useful.

(b) Consider a fictional dataset as given below that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as "Yes" or "No" for playing golf. Apply Naïve Bayes theorem to predict the class of a new data item X = (Sunny, Hot, Normal, False)

| S. No. | Outlook | Temperature | Humidity | Windy | Play Golf |
|---|---|---|---|---|---|
| 0 | Rainy | Hot | High | False | No |
| 1 | Rainy | Hot | High | True | No |
| 2 | Overcast | Hot | High | False | Yes |
| 3 | Sunny | Mild | High | False | Yes |
| 4 | Sunny | Cool | Normal | False | Yes |
| 5 | Sunny | Cool | Normal | True | No |
| 6 | Overcast | Cool | Normal | True | Yes |
| 7 | Rainy | Mild | High | False | No |
| 8 | Rainy | Cool | Normal | False | Yes |
| 9 | Sunny | Mild | Normal | False | Yes |
| 10 | Rainy | Mild | Normal | True | Yes |
| 11 | Overcast | Mild | High | True | Yes |
| 12 | Overcast | Hot | Normal | False | Yes |
| 13 | Sunny | Mild | High | True | No |

**[9+9]**

**6.** (a) Consider the following dataset. Use Apriori algorithm to find frequent item sets, and generate association rules for them. Minimum support count is two and minimum confidence is 50%.

| TID | Items |
|---|---|
| T1 | I1, I2, I5 |
| T2 | I2, I4 |
| T3 | I2, I3 |
| T4 | I1, I2, I4 |
| T5 | I1, I3 |
| T6 | I2, I3 |
| T7 | I1, I3 |
| T8 | I1, I2, I3, I5 |
| T9 | I1, I2, I3 |

(b) Discuss Frequent Pattern Growth Algorithm. How it is different than Apriori algorithm ?

**[10+8]**

**7.** (a) Discuss the following two approaches of data generalization :

(i) Data cube approach

(ii) Attribute oriented induction

(b) Discuss the procedures are adopted for Class discrimination or comparison mines characterization.

(c) Discuss in brief the pros and cons of using Genetic Algorithms in datamining.

**[9+6+3]**

**- o O o -**