No. of Printed Pages : **2**

Sl. No.

BE6-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

1. **Answer question no. 1 and any FOUR questions from 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

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

1. (a) Define KDD. Identify and describe the phases in KDD process with neat and clean figure.

(b) Differentiate between OLAP and OLTP using suitable examples.

(c) What is spatial database ? Explain the methods of mining spatial databases.

(d) What is the role of statistics in data mining ?

(e) Briefly discuss the schemas for multidimensional databases.

(f) Why is naïve Bayesian classification called "naïve" ?

(g) What is an outlier detection ? How it is useful in data mining ? **(7x4)**

2. (a) Define Data warehouse. Differentiate between operational database systems and data warehouses.

(b) Write an algorithm for k-nearest-neighbour classification given k and n, the number of attributes describing each tuple.

(c) How market basket analysis helps in frequent item set mining and discovery of associations and correlations among items in large transactional or relational data sets ? **(6+6+6)**

3. (a) Explain various OLAP operations on multidimensional data such as rollup, drill down, slice and dice and pivot. Give examples.

(b) Briefly explain three-tier data warehousing architecture.

(c) Give the difference between Boolean association rule and Quantitative Association rule. **(6+6+6)**

4. (a) Explain with example Regression Analysis, Histogram and Standard deviation.

(b) Explain data discretization techniques with suitable example. How concept hierarchies for numerical attributes constructed based on data discretization ?

(c) How does Back propagation algorithm work ? Explain the input and output function in Hidden layer of Multilayer Feed-Forward Neural Network. **(6+6+6)**

5. (a) Suppose that the data for analysis includes the attribute age. The age values for the 10 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, 36, 40, 45, 46, 52, 70.

(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 (i.e., bimodal, trimodal, etc.)

(iii) What is the midrange of the data ?

(iv) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data ?

(v) Give the five-number summary of the data.

(b) Define and describe the basic similarities and differences between ROLAP, MOLAP and HOLAP.

(c) Define hierarchical clustering method. Explain the advantages of hierarchical clustering. **(6+6+6)**

6. (a) What is Association rule mining ? Explain the Apriori algorithm to find the frequent item sets using your own example.

(b) Define clustering. Explain the density-based clustering method based on connected regions with sufficiently high density. **(9+9)**

7. (a) Suppose that a data warehouse consists of the four dimensions-*date, spectator*, *location*, and *game*, and the two measures - *count* and *charge*, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(i) Draw a star schema diagram for the data warehouse.

(ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators ?

(iii) Bitmap indexing is useful in data warehousing. Taking this cube as an example, briefly discuss advantages and problems of using a bitmap index structure.

(b) Differentiate between the following :

(i) Regression and Classification

(ii) Supervised Learning and Unsupervised Learning

(iii) Data cleaning and Data transformation

(c) Compare the advantages and disadvantages of eager classification and lazy classification methods. **(5+9+4)**

- o O o -