No. of Printed Pages : **2**

Sl. No.

BE6-R4 : DATA WAREHOUSE AND DATA MINING

NOTE :

1. Answer question 1 and any FOUR from questions 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) Differentiate between data warehouse and database. How are they similar ?

(b) Why concept hierarchies are useful in data mining ?

(c) Use the two methods below to normalize the following group of data :

200, 300, 400, 600, 1000

- min-max normalization by setting min = 0 and max = 1
- z-score normalization

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

(e) List the various requirements of Cluster Analysis.

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

(g) Define KDD. Identify and describe the phases in KDD process. **(7x4)**

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

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

(c) Give minimum six differences between OLAP and OLTP.

(d) Describe information gain and Gini index. **(5+5+4+4)**

3. (a) Give examples of typical OLAP operations on multidimensional data (Rollup, Drill down, slice and dice and Pivot).

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

(c) How do market basket analysis help in frequent itemset mining and discovery of associations and correlations among items in large transactional or relational data sets ?

(d) Why is tree pruning useful in decision tree induction ? What is a drawback of using a separate set of tuples to evaluate pruning ? **(5+5+4+4)**

4. (a) Define clustering. Briefly describe the portioning and hierarchical clustering methods with example.

(b) How does Backpropagation algorithm work ? Explain the input and output function in Hidden layer of Multilayer Feed-Forward Neural Network.

(c) Consider a database has five transactions. Let min sup = 60% and min conf = 80%.

TID	Items
T100	M, O, N, K, E, Y
T200	D, O, N, K, E, Y
T300	M, A, K, E
T400	M, U, C, K, Y
T500	C, O, O, K, I, E

- Find all frequent item sets using Apriori.
- Find all the frequent rules.

(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.

- What is the mean of the data ? What is the median ?
- What is the mode of the data ? Comment on the data's modality (i.e., bimodal, trimodal, etc.)
- What is the midrange of the data ?
- Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data ?
- Give the five-number summary of the data.

(b) What is association rule mining ? Explain the Apriori algorithm to find the frequent item sets.

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

(6+6+6)

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

(b) Briefly compare the following concepts. You may use an example to explain your point(s).

- Snowflake schema, fact constellation
- Enterprise warehouse, data mart, virtual warehouse

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

(6+6+6)

7. (a) Explain the density-based clustering method based on connected regions with sufficiently high density.

(b) Write short notes on following :

- Bayesian classification
- Genetic Algorithm
- Neural Network Technology
- Lattice of Cuboid

(6+12)

- o O o -