BE6-R4: DATA WAREHOUSING 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) How is a data warehouse different from a database? How are they similar?

b) Briefly describe the following advanced database systems and their applications: spatial databases and text databases.

c) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

d) Differentiate between Web Content Mining and Web Usage Mining.

e) Discuss issues to consider during data integration.

f) Briefly compare the following concepts with the help of an example: Snowflake schema, fact constellation and starnet query model.

g) What is Hypothesis Testing? How it is generated? Explain with an example.

(7x4)

2.

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

b) Differentiate between Discrete and Continuous Attributes variables.

c) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

(8+5+5)

3.

a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

b) Suppose a group of 12 students age is recorded has been sorted as follows: 5; 10; 11; 13; 15; 35; 50; 55; 72; 92; 204; 215: Partition them into three bins by each of the following methods.

i) equal-frequency partitioning

ii) equal-width partitioning

c) Define the following terms:

i) Data Characterization

ii) Data Discrimination

iii) Data Tube

iv) OLAP

(6+6+6)

4.

a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit and count is number of patient.

i) Draw a STAR schema diagram for the above data warehouse.

ii) Starting with the base cuboid [day; doctor; patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

iii) To obtain the same list, write an SQL query assuming the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

b) Compare the advantages and disadvantages of eager classification (e.g., decision tree, Bayesian, neural network) versus lazy classification (e.g., k-nearest neighbour, case-based reasoning).

(10+8)

5.

a) A database has five transactions. Let min support = 60% and min confidence = 80%.
{M,O,N,K,E,Y},{D,O,N,K,E,Y}, {M,AK,E}, {M,U,C,K,Y}, {C,O,O,K,I,E}

i) Find all frequent itemsets using Apriori algorithm.

ii) List all of the strong association rules (with support s and confidence c) matching the following Meta rule, where X is a variable representing customers and item i denote variables representing items (e.g., “A”,”B”, etc.):

∀x ∈ transaction, buys(X, item1) ∧ buys(X, item2) ⇒ buys(X, item3) [s,c]

b) Discuss the various methods used for assessing the accuracy of a classifier.

(10+8)

6.

a) Suppose that the data mining task is to cluster the following eight points (with (x; y) representing location) into three clusters.
A1(2; 10);A2(2; 5);A3(8; 4);B1(5; 8);B2(7; 5);B3(6; 4);C1(1; 2);C2(4; 9):
The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only.

i) The three cluster centers after the first round of execution and

ii) The final three clusters

b) What is Genetic Algorithm? Discuss the role of various operators used in Genetic Algorithm.

c) Outliers are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Taking fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

(8+6+4)

7.

a) Briefly describe the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data, and constraint-based methods. Give examples in each case.

b) What are Feed Forward Neural networkings? How are they beneficial in Artificial Neural Networks?

c) How to choose an efficient training data in case of Artificial Neural Networks? Explain, what are Self-Organizing Maps (SOM).

(6+6+6)