C5-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) Define Data, Information and Knowledge?
b) What are differences between OLTP and OLAP?
c) In real-world data, tuples (instances) with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.
d) Describe the differences between Association rule mining and Classification in terms of data used and results of learning.
e) What is Data Mining? What is the relation between data mining and KDD?
f) Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, prediction, clustering, and evolution analysis.
g) What are the differences between visual data mining and data visualization?

(7x4)

2. 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.

a) Enumerate three classes of schemas that are popularly used for modeling data warehouses.
b) Draw a star schema diagram for the above data warehouse.
c) 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 2016?
d) 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).

(2+6+6+4)

3.

a) When computing a cube of high dimensionality, we encounter the inherent curse of dimensionality problem: there exist a huge number of subsets of combinations of dimensions.
   i) Suppose that there are only two base cells, f(a1; a2; a3; : : : ; a100), (a1; a2; b3; : : : ; b100), in a 100-dimensional base cuboid. Compute the number of nonempty aggregate cells. Comment on the storage space and time required to compute these cells.
   ii) Suppose we are to compute an iceberg cube from the above. If the minimum support count in the iceberg condition is two, how many aggregate cells will there be in the iceberg cube? Show the cells.
b) What are the different types of OLAP?

(12+6)

4.

a) Compare the advantages and disadvantages of eager classification (e.g., decision trees, Bayesian, etc) versus lazy classification (e.g., k-nearest neighbours).
b) Given the data points (0, 0/1), (1, 0/1), (1, 1/1), (4, 3/2), (3, 4/2), (1, 4/2) where (x, y/c) indicates the x and y attributes and c is the class label. Use k-NN to find the class label for (2, 2/?) when k=1, and when k=5. Use Euclidean distance as the distance measure.

(10+8)

5. A database has five transactions. Let min sup = 60% and min conf = 80%.

TID	items bought
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

a) Find all frequent item sets using Apriori.
b) Find all the frequent rules.

(9+9)

6.

a) Explain K-nearest Neighbour Classifier.
b) Explain usage of Genetic Algorithm for Data Mining Application.

(9+9)

7.

a) Apply the bottom-up hierarchical algorithm for the following 1-dimensional points for (k=2):1; 2; 3; 4; 6; 7; 8; 9. Draw the dendrogram. When there is a tie, choose clusters with “smaller means” before choosing clusters with larger means.
b) 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
i) The three cluster centers after the first round of execution, and
ii) The final three clusters.

(9+9)