DATA WAREHOUSING AND DATA MINING
(INFO 3201)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) A data warehouse is an integrated collection of data because it is a
(a) collection of data of different datatypes
(b) relational database
(c) summarized data
(d) data collected from multiple sources.

(ii) _____ is a hierarchical based clustering algorithm.
(a) Fuzzy C-means (b) ROCK
(c) K-means (d) DBSCAN.

(iii) In apriori algorithm border sets are infrequent itemsets whose
(a) all proper subsets are frequent (b) all supersets are frequent
(c) all subsets are also infrequent (d) none of the above.

(iv) Perceptron is not able to implement
(a) OR gate (b) AND gate
(c) XOR gate (c) NOT gate.

(v) The task of correcting and preprocessing data is called _____.
(a) data analysis (b) data processing
(c) data mining (d) data cleansing.

(vi) A schema with a fact table, multiple dimensional table and foreign keys from the fact table to the dimension table is called a _____.
(a) snowflake schema (b) star schema
(c) sub schema (d) logical schema.

(vii) In DBSCAN algorithm points having atleast minimum required neighbours are
(a) core points (b) border points
(c) both (a) and (b) (d) none of the above.

(viii) A density based clustering algorithm is
(a) K-means (b) Agglomerative approach
(c) ROCK (d) DBSCAN.

(ix) In Fuzzy C-Means C represents
(a) number of data points (b) number of clusters
(c) number of border points (d) number of neighbors.

(x) The full form of KDD is _____
(a) Knowledge Database
(b) Knowledge Discovery Database
(c) Knowledge Data House
(d) Knowledge Data Definition.

Group – B

2. (a) What is a data warehouse? Explain the different characteristics of a data warehouse.

(b) Explain about data mart and state its different types.

(c) Compare the modelling paradigm, star schema and snowflake schema, of a data warehouse.

(2 + 3) + 2 + 5 = 12

3. (a) What is OLAP cube? Explain with example the different operations applied on a data cube.

(b) Explain the different steps used in Knowledge Discovery in Database.

(c) Suppose a data warehouse consists of three dimensions doctor, time, patient, and two measures count and charge, where charge is the fee a doctor charges a patient for a visit. Starting with the base cuboid [day, doctor, patient] what specific OLAP operations (e.g., Slice for Time = Year) should be performed in order to list the total fee collected by each doctor in the year 2016?

(2 + 4) + 3 + 3 = 12

Group – C

4. (a) What is the difference between maximal frequent set and border set with respect to association rule mining?

(b) Consider the five transactions given below. If minimum support is 35% and minimum confidence is 86%, determine the frequent itemsets and association rules using the apriori algorithm.

Transaction	Items
T1	Bread, Butter, Honey
T2	Jelly, Cheese, Honey
T3	Jelly, Apple, Cheese
T4	Bread, Butter
T5	Coke, Apple

4 + (5 + 3) = 12

5. (a) Write the algorithm of dynamic itemset counting algorithm, for finding frequent itemsets.

   (b) Find all frequent itemsets or frequent patterns in the following database using dynamic itemset counting algorithm. Take minimum support as 20%.

A1	A2	A3	A4	A5
1	0	0	1	0
0	1	1	1	0
1	1	1	1	1
0	0	0	0	1
1	0	0	1	0

5 + 7 = 12

Group – D

6. (a) Discuss the limitations of k-means algorithm. Cluster the following data points using k-means clustering technique, where k=3 and each data point represented in the form of (x_coordinate, y_coordinate). Consider A1, B1, C1 as the initial cluster centers.

   Data Points: A1(3, 12); A2(6, 3); A3(9, 7); B1(18, 21); B2(45, 56); B3(6, 4); C1(3, 3); C2(19, 20).

   (b) With respect to DBSCAN algorithm, what are border point, core point and noise point?

   (c) In case of categorical clustering, how can we define similarity functions and link between the clusters?

(2 + 6)+ 2 + 2 = 12

7. (a) Explain how the membership matrix in fuzzy C-means clustering algorithm gets updated.

   (b) Write the back propagation based neural network algorithm. Also show the working of the same for designing the two-input one-output XOR network.

3 + (5 + 4) = 12

Group – E

8. (a) Construct the decision tree model from the following training dataset, using gain ratio indices.

Name	Gives birth	Skin Cover	Aquatic Creature	Has Legs	Class
Human	Yes	Hair	No	Yes	Mammals
Python	No	Scales	No	No	Reptiles
Salmon	No	Scales	Yes	No	Fish
Whales	Yes	Hair	Yes	No	Mammals
Pigeon	No	Feathers	No	Yes	Birds

   (b) How is the web usage mining different from web structure mining and web content mining?

8 + 4 = 12

9. Write short notes (any two): **(6 × 2) = 12**
   (i) HDFS architecture
   (ii) Parallel programming in Hadoop MapReduce
   (iii) Web content mining
   (iv) Temporal data mining.