B.TECH/IT/6TH SEM/INFO 3201/2019
DATA WAREHOUSING & 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 schema with a fact table, multiple dimensional tables 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.

(ii) A data warehouse is said to contain a 'time-varying' collection of data because
(a) its contents vary automatically with time
(b) its life span is very limited
(c) it contains historical data
(d) its content has explicit time-stamp.

(iii) The algorithm which uses the concept of a train running over data to find associations of items in data mining is known as
(a) apriori (b) partition
(c) dynamic item-set counting (d) FP-tree growth.

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

(v) One of the limitation of Fuzzy C- means is requirement for
(a) number of data points (b) number of clusters
(c) number of border points (d) number of neighbours.

(vi) ------------ is a partitioning based clustering algorithm.
(a) Fuzzy C-means (b) ROCK
(c) DBSCAN (d) none of this.

(vii) DIC stands for
(a) Dynamic Itemset Counting (b) Data Informative Counting
(c) Dynamic Informative Count (d) None of this.

(viii) Sequence of jobs to load data in to warehouse is as follows:
(a) First load data into fact tables then dimension tables, then aggregates if any
(b) First load data into dimension tables, then fact tables, then aggregates if any
(c) First aggregates then load data into dimension tables, then fact tables
(d) Does not matter if we load either of fact, dimensions, or aggregates.

(ix) Naive Bayes and Decision Tree are used for
(a) supervised learning (b) unsupervised learning
(c) both of this (d) none of this.

(x) The programming paradigm used in Hadoop ecosystem is
(a) Mapper (b) Reducer
(c) MapReduce (d) none of this.

Group – B

2. (a) Explain what an OLAP cube is. Discuss the fact constellation schema of a data warehouse.

(b) Suppose a data warehouse consists of three dimensions : doctor, time, patient and two measures count and charge where charge is the fee of a doctor charges for a patient for a visit.
Draw the star schema diagram for the above data warehouse.

(2 + 4) + 6 = 12

3. (a) State the difference between OLTP and OLAP. What is metadata? How is it different from external data?

(b) Explain the two different characteristics of a data mart.

(c) Explain with example star schema & snowflake schema.

(2+ 2+ 2) + 2+ (2 + 2) =12

Group – C

4. (a) Given a database of books with given 10 transactions and the support of all 1- itemsets are given in the table. Considering minimum support as 35%, find out all the frequent itemsets and design five association rules that exists (if any) having confidence as 80%.

{Let us C}	6
{Numerical Methods}	2
{Datastructure and Algorithms}	7
{Digital Electronics}	6
{Object Oriented Programming}	2
{Database Theory}	7
{Data Analytics}	3
{Pattern Recognition}	4
{Ecommerce}	1
{Web Technology}	3

(b) Discuss the limitation of apriori algorithm.

(8 + 2) + 2= 12

5.(a) Apply **DIC** algorithm for the following transactional database to generate the frequent item sets. Also design atleast five association rules (if possible) with confidence of 55%. The minimum support is 35%. The stop number M=5. Also calculate number of passes required to find all the frequent itemsets.
T1--->P,Q,R,T, T2---> Q,S,T, T3--> Q,R, T4--->P,Q,S, T5---> P, R, T6--> Q,R, T7--> P,R,T, T8--> P,Q,R,T, T9-->P,Q,R, T10-->Q,R

(b) State the downward closure property and upward closure property with respect to association rule mining.

9 + 3 =12

Group – D

6. (a) Cluster the following data points using k-means clustering technique, where k=2 and each data point represented in the form of (x_coordinate, y_coordinate). Consider A1, B1, C1 as the initial cluster centroids.

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) Consider the transactional database below. Using the concept of ROCK clustering, find out the neighbors of each object and also find the link between object 1 and 3, considering the threshold =1/3.

Transaction Id	Items Bought
T1	A,C,D
T2	D,F,G,R
T3	A,C,D
T4	A,G,R,C,F

6 + 6 = 12

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

Name	Hair	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned
Ana	blonde	tall	average	yes	none
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned

(b) Write the back propagation based neural network algorithm. Also show the working of the same for designing the two i/p one o/p XOR network.

7 + 5 = 12

Group – E

8. (a) Explain the Hadoop HDFS architecture.

(b) Explain in details the working principle of Hadoop MapReduce architecture.

5 +7 = 12

9. Write short notes on (any two)
(i) Web mining
(ii) PCA
(iii) Link Analysis Text Mining
(iv) Map Reduce Technique

2 × 6 = 12