B.TECH/IT/6TH SEM/INFO 3201(BACKLOG)/2021

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
<u>any 5 (five)</u> from Group B to E, taking <u>at least one</u> 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 point in a feature space is known as a
(a) Feature array (b) Feature Vector
(c) Convex Hull (d) None of the above

(ii) Hierarchical agglomerative based clustering is a
(a) Bottom up Approach (b) Top down Approach
(c) Both (a) and (b) (d) None of the above

(iii) ROCK clustering means
(a) Robust hierarchical clustering
(b) Robust hierarchical clustering with links
(c) Runtime hierarchical clustering with links
(d) Robust clustering with kmedoids

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

(v) In Fuzzy C Means, C represents
(a) Number of data points (b) Number of clusters
(c) Number of border points (d) Number of neighbours

(vi) The advantage of FP-tree Growth Algorithm is
(a) it counts the support values of the itemsets in the dashed structure as it moves along from one stop point to another.
(b) it avoids the generation of large numbers of candidate sets
(c) to update the association rules when the database discover the set of frequent item sets.
(d) none of the above.

(vii) A density based clustering algorithm is
(a) PAM (b) STIRR (c) ROCK (d) DBSCAN

(viii) 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) DIC (d) FP-Tree growth

(ix) Association rules are always defined on________.
(a) binary attribute (b) single attribute
(c) relational database (d) multidimensional attributes

(x) The algorithms based on partitioning paradigm
(a) K-means (b) STIRR
(c) Both (a) & (b) (d) None of the above

Group – B

2. (a) Explain the term OLAP cube. Discuss the fact constellation schema of a Data warehouse.

(b) Suppose a data warehouse consists of three dimensions : doctor, time and patient. It also consists of two measures: count and charge where charge is the consultation fees for a patient visit to a doctor.

(c) Draw the Star schema diagram for the above data warehouse.

2 + 4 + 6 =12

3. (a) State the difference between OLTP and OLAP. Describe the types of OLAP operations supported by OLAP tools.

(b) Explain the different characteristics of a data warehouse.

(4 + 3) + 5 = 12

Group – C

4. (a) Design all Frequent Itemsets using apriori algorithm from the following transaction data given minimum support = 30%. In addition design all association rules from the above Frequent Sets at min Confidence 60%.

Transaction Id	Data Items
1	A ,B , C , E
2	B , D , E
3	B , C
4	A , B ,D
5	A , C
6	B , C
7	A , C, E
8	A , B , C , E
9	A , B , C
10	C , D, E

(b) What are the shortcomings of apriori algorithm.

10 + 2 = 12

5. (a) Discuss on Dynamic itemset counting.

(b) Discuss the different phases of Fuzzy C-Means clustering algorithm. Discuss the limitations of this algorithm.

6 + (5 + 1) = 12

Group – D

6. (a) Construct a Decision Tree using the weekend spending data, as given in the following Table.

Week End	Weather type	Humidity	Money Expended	Decision
Week1	Hot	High	500	Stay In
Week2	Cold	Low	2000	Shopping
Week3	Rainy	Low	1500	Restaurant
Week4	Rainy	High	500	Stay In
Week5	Hot	Low	2000	Restaurant
Week6	Cold	High	1500	Shopping
Week7	Hot	Low	2000	Shopping
Week8	Cold	Low	500	Restaurant
Week9	Cold	High	2000	shopping
Week10	Rainy	High	500	Stay In

(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

8 + 4 =12

7. (a) Explain the working principle of Naive Bayesian Classification. In addition, find the Class(X) using Naïve Bayes on the following Dataset, where X= (Age=30; Income=high ; Student=No ; Credit Rating= Fair)

age	income	student	credit_rating	buys_computer
< = 30	high	no	fair	no
< = 30	high	no	excellent	no
31 .. 40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31 .. 40	low	yes	excellent	yes

< = 30	medium	no	fair	no
< = 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
< = 30	medium	yes	excellent	yes
31 .. 40	medium	no	excellent	yes
31 .. 40	high	yes	fair	yes
> 40	medium	no	excellent	no

(b) Group 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(2,10); A2(2, 5); A3(8,4); B1(5, 8); B2(7, 5); B3(6,4); C1(1,2); C2(4,9).

7 + 5 = 12

Group – E

8. (a) Explain how parallelism is encountered in Map Reduce paradigm.

(b) Using the Map Reduce paradigm compute the number of words starting with vowel and number of words starting with consonant in the following text.
“ There is a Workshop in HIT. The workshop is on Big Data Analytics. Heritage is in Kolkata.”

2 + 10 = 12

9. (a) Describe with the help of a diagram the architecture of Hadoop Distributed File System.

(b) Discuss on PageRank algorithm with respect to Web structure Mining.

6 + 6 = 12

Department & Section	**Submission Link**
IT	https://classroom.google.com/c/MzY5MTUwODk0ODky/a/MzY5MTUwODk0OTEy/details