B.TECH/IT/6TH SEM/INFO 3201/2017

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) ---------- is a repository of information gathered from multiple sources stored under a unified schema at a single site.
(a) Data mining (b) Web server (c) Data Warehouse (d) None.

(ii) A point in a feature space is known as a
(a) Feature array (b) Feature Vector
(c) Convex Hull (d) None of the above.

(iii) 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.

(iv) In Fuzzy C Means, C represents
(a)Number of data points (b) Number of clusters
(c) Number of border points (d) Number of neighbors.

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

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

(vii) _____ is the technique which is used for discovering patterns in dataset at the beginning of data mining process.
(a) Kohenon map (b) Visualization (c) OLAP (d) SQL.

(viii) _____ is the application of data mining techniques to discover patterns from the Web.
(a) Text Mining (b) Multimedia Mining
(c) Web Mining (d) Link Mining.

(ix) _____analysis divides data into groups that are meaningful, useful, or both.
(a) Cluster (b) Association (c) Classification (d) Relation.

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

Group – B

2. (a) How is a data warehouse different from database? State the different characteristics of the same.

(b) With an example explain what is Meta data? What is data mart?

(c) Compare and contrast star schema and snowflake schema

(2 + 3) + (2 + 2) + 3 = 12

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

(b) Differentiate between OLAP and OLTP.

(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 + 5) + 3 + 2 = 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%

(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. Write short notes (any two)
a) Hadoop
b) Web Structure Mining vs. Web Content Mining
c) Text Mining
d) Page Rank Algorithm

(2 × 6) = 12

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) How is data mining different from KDD?

(b) Find all frequent itemsets or frequent patterns in the following database using Dynamic Itemset counting algorithm and FP-tree growth algorithm. Take minimum support as 20%.

Tid	Items
1	A, E, F, H
2	B, D, H
3	D, E, G
4	C
5	E, F, G
6	B, C, D
7	B, F, G, I
8	E
9	H
10	C, E, G
11	C, E, G
12	E, F, H
13	B, D, F, G
14	A, C, E, G
15	B, C, I

3 + (4 + 5) = 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