AUTUMN END SEMESTER EXAMINATION-2019

7th Semester B.Tech & B.Tech Dual Degree

DATA MINING AND DATA WAREHOUSING

IT 4037

(For 2017(L.E), 2016 & 2015 Admitted Batches)

Time: 3 Hours Full Marks: 60

Answer any SIX questions.
Question paper consists of four sections-A, B, C, D.
Section A is compulsory.
Attempt minimum one question each from Sections B, C, D.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

SECTION-A

1. Answer the following questions. [2× 10]

(a) Explain how the evolution of database technology lead to data mining.

(b) Briefly discuss the Data cleaning.

(c) How FP growth tree is better than Apriori?

(d) Use the two methods below to normalize the following group of data: 4. 200, 300, 400, 600, 1000 (a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization

(e) Compare between data warehouse and a traditional database system.

(f) Define support and confidence in Association rule mining.

(g) How to handle noise in data?

(h) Analyse and illustrate in which context k-medoid algorithm is more suitable than the k-means algorithm with suitable example.

(i) What is a confusion matrix and examine its role in evaluation of a classifier's performance?

(j) Compare the eager learners with lazy learners.

SECTION-B

2. (a) Explain KDD process in details. [4]

   (b) Demonstrate the classification process carried out by decision tree induction? List out the methods of attribute selection measures. [4]

3. (a) Illustrate the procedure of spatial mining and text mining. [4]

   (b) Demonstrate the working principle of genetic algorithm with a flow chart. [4]

SECTION-C

4. (a) Analyse the differences between OLAP and OLTP. [4]

   (b) Define the essential quality measures of an information retrieval system and analyse which one is a good measure for this purpose. [4]

5. (a) In a database, there are 2000 patient records out of which 600 are flu patient records. A disease predictor application accesses that database and predicts 150 flu patients correctly out of 200 flu patients and 300 non-flu patients. Find the following performance parameters. a) Sensitivity b) Specificity c) Accuracy d) Precision [4]

   (b) Consider the following transactional database T. Let min sup = 60% and min conf = 80%. [4]

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}

I. Find all frequent itemsets using Apriorialgorithms.

II. Which of the itemsets from (I) are closed? Which of the itemsets from (I) are maximal?

III. Determine strong association rules.

6. (a) The task is to cluster the following 8 points (with (x, y) representing location) into 3 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, C1 as centre of each cluster, respectively. Use k-means algorithm to show only [4]

(i) The three cluster centers after the first round execution.

(ii) The final 3 clusters

(b) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8). Compute Euclidean distance, Manhatton distance and Minkowski distance(q=3) between these two given objects. [4]

SECTION-D

7. (a) Consider the given dataset [4]

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

a) Which attributes are asymmetric and why?

b) Draw the contingency table for (jack , mary) and (jim, mary).

c) Calculate distance d(jack, mary), d(jim, mary) .

(b) Consider the given credit audit database. Predict the "Response" of the customer "David" using 3- Nearest Neighbors algorithm. [4]

Customer	Age	Income	No. Of credit cards	Response
John	35	35K	3	N
Rachel	22	50K	2	Y
Hannah	63	200K	1	N
Tom	59	170K	1	N
Nellie	25	40K	4	Y
David	37	50K	2	?

8. (a) Given below the training dataset, using Bayesian classifier find out the conditional probability that a certain member of the school is a 'Teacher' given that he is a 'Man'. [4]

	Male	Female
Teacher	12	8
Student	48	32

(b) Construct the single link agglomerative hierarchical clustering for the given distance matrix: [4]

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

*****