8. Discuss the following clustering algorithm using examples : 14

(a) DBSCAN

(b) Web mining

(c) Temporal mining

(d) BIRCH algorithm

CS+IT

COPYRIGHT RESERVED UL(7) — DWH & DM

CS+IT
VII Sem

2020(A)

Time : 3 hours

Full Marks : 70

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

The figures in the margin indicate full marks.

Answer any five questions.

1. (a) Demonstrate the applications of data mining for financial analysis. 7

   (b) How classification of data mining system is done ? Explain them with example. 7

2. Examine the steps involved for the design and construction of data warehouse. 14

3. (a) Explain FP tree algorithm with an example. 7

   (b) Explain K-means algorithm with example. 7

Scanned with CamScanner

4. (a) Explain the algorithm for constructing a decision tree from training samples. 7

(b) Explain the data structures and schema that support multidimensional data in data warehouse with suitable illustration. 7

5. Suppose that a data warehouse for big university consist of the following four dimensions : Student, Course, Semester and Instructor, and two measures count and avg_grade. When at the lowest conceptual level (Ex. For a given student, course, semester and instructor combination). the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

(a) Draw a snowflake schema diagram for the data warehouse. 4

(b) Starting with the base cuboid [Student, Course, Semester, Instructor], what specific OLAP operations (eg - Roll-up from Semester to year) should one perform in order to list the average grade of CS for each big- university students. 5

(c) If each dimension has five levels (including all) such as Student<major<status< university<all, how many cuboids will this cube contains (including the base and Apex cuboids). 5

6. (a) What is a spatial database ? Explain the methods of mining spatial databases ? 7

(b) Describe how OLAP technology helps in discovery driven exploration of data cubes. 7

7. Write an algorithm for constructing a decision tree. Construct a decision tree for the following data set using information gain. Predict the class label for a data point with values<Female, 2, standard, high : 14

Gender	Car ownership	Travel cost	Income level	Transport mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

Scanned with CamScanner

6. (a) Write down the difference between Artificial Neural Network and Biological Neural Network. 7

(b) What are the different learning law in ANN, explain in brief ? 7

7. (a) What is linearly inseparable problem ? Show that Ex-OR and Ex-NOR are linearly inseparable. 7

(b) Explain Genetic Algorithm. Illustrate with a simple example. 7

8. (a) What is the significance of ensemble learning in machine learning ? Explain with suitable example. 7

(b) Explain logistic regression in machine learning. Explain with example. 7

Scanned with CamScanner

8. (*a*) Explain the application of the data warehousing and data mining in Government. 7

(*b*) Describe three-tier data warehouse architecture. 7

VII Sem. CS+IT

2019 (A)

Full Marks : 70

Time : 3 hours

DEC-19

Answer any five questions.

The figures in the right-hand margin indicate marks.

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

1. (*a*) Differentiate between OLTP and OLAP systems. 7

(*b*) Describe star schema, snowflake schema and fact constellation schema with example. 7

2. (*a*) What are the different ways to handle missing values in data mining ? 7

(*b*) Differentiate between database and data warehouse. 7

3. (*a*) Define box plot. Draw a box plot for the data given below :

2, 51, 53, 54, 43, 51, 62, 49, 50, 63, 60. 5

Scanned with CamScanner

(*b*) What is min-max normalization ? Use the min-max normalization method to normalize the following group of data : 200, 300, 400, 600, 1000 by setting min value = 0 and max value = 1. 5

(*c*) What is Noisy Data ? Remove the noisy data by smoothing techniques for given data 4, 8, 15, 21, 21, 24, 25, 28, 34. 4

4. A database has nine transactions. Consider min_support as 22.22% and min_confidence as 70%.

TID	List of Item_IDs
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

Find all frequent item sets using :

(*a*) Apriori algorithm 7

(*b*) FP-growth algorithm. 7

5. (*a*) Describe major steps for constructing a decision tree from the training dataset. 7

(*b*) Describe information gain, gain ratio and gini index. 7

6. (*a*) Describe K-means clustering. 4

(*b*) Suppose that the data mining task is to cluster points [with (x, y) representing location] into two clusters, where the points are A_1 (2, 10), A_2 (2, 5), A_3 (8, 4), A_4 (5, 8), B_1 (7, 5), B_2 (6, 4), B_3 (1, 2), B_4 (4, 9). The distance function is Euclidean distance. Suppose initially we assign A_1 and B_1 as the centre of each cluster, respectively. Use the k-means algorithm to find the two cluster centers after the second round of iteration. 10

7. Write the short notes on the following : 14

(*a*) Web content mining

(*b*) Text mining

(*c*) DBSCAN

(*d*) BIRCH

Scanned with CamScanner

(*b*) What is Artificial Neural Network ? Give two examples of ANN in detail. 7

8. Write short notes on (any *two*) : 7+7

(*i*) Bayesian Network

(*ii*) Fuzzy Logic

(*iii*) Frames

Scanned with CamScanner

7. Draw decision tree for the following data sets. Use entropy as a node selection mechanism. 14

RID	age	income	student	credit–rating	Class: buys-computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

8. (*a*) Explain the working principle of DBSCAN with example. 7

(*b*) Explain the applications of the data warehousing and data mining in Government. 7

2018

Full Marks : 70

Time : 3 hours

Answer any **five** questions.

The figures in the right-hand margin indicate marks.

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

1. (*a*) How is a data warehouse different from a database ? How are they similar ? 7

(*b*) Describe the steps involved in data mining when viewed as a process of knowledge discovery. 7

2. Briefly compare the following concepts. You may use an example to explain your point(s).

(*a*) Snowflake schema, fact constellation, starnet query model 7

(*b*) Data cleaning, data transformation, refresh 7

Scanned with CamScanner

3. (*a*) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 7

(*b*) Consider the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. 7

(*i*) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

(*ii*) How might you determine outliers in the data ?

(*iii*) What other methods are there for data smoothing ?

4. A database has five transactions. Let min sup D 60% and min conf D 80%.

TID	items-bought
T 100	{M, O, N, K, E, Y}
T 200	{D, O, N, K, E, Y}
T 300	{M, A, K, E}
T 400	{M, U, C, K, Y}
T 500	{C, O, O, K, I, E}

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes. 14

5. (*a*) Briefly outline the major steps of decision tree classification. 7

(*b*) What is web content mining ? How is it different from web structure mining ? 7

6. (*a*) Briefly describe and give examples of each of the following approaches to clustering: partitioning methods, hierarchical methods, density-based methods, and grid-based methods. 7

(*b*) Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are

$A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$.

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the *k*-means algorithm to show only the three cluster centers after the first round of execution. 7

Scanned with CamScanner