RegNo															

SRM Institute of Science and Technology
College of Engineering and Technology
Department of Computing Technologies, School of Computing
SRM Nagar, Kattankulathur – 603203
Academic Year: 2023-24 (ODD)
SET - C

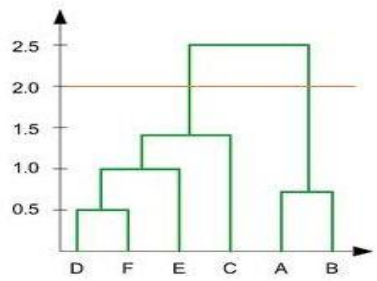
Test: CLA-T3 **Date & Session:**
Course Code & Title: 18CSE355T & Data Mining and Analytics **Duration:** 1 Hr 40 minutes
Year & Sem: III/IV Year & V/VII Sem **Max. Marks:** 50

Course Articulation Matrix:

S. No	Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
1	CO1	H	L	M									
2	CO2		H	M	L								
3	CO3		H	M	L								
4	CO4	H	M	L									
5	CO5	H	M	L									
6	CO6	L			M	H							

Part – A (10 x 1 = 10 Marks)
Answer all Questions

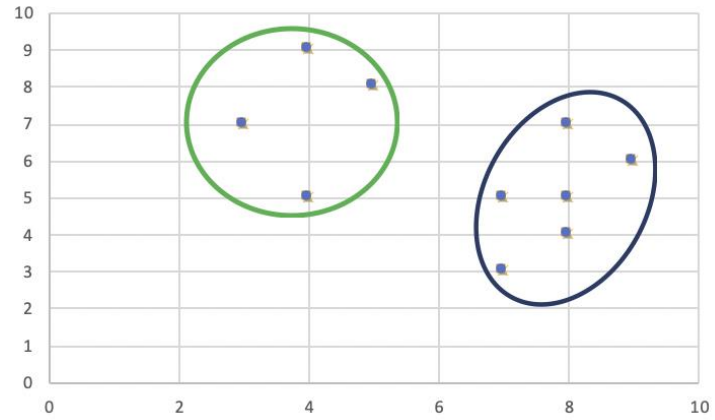
Q. No	Question	Mark	BL	CO	PO	PI Code
1	Which of the following is not true in detecting outliers? a. Proximity-Base Approaches b. Clustering-Base Approaches **c. Time-Base Approaches** d. Classification Approaches	1	L2	5	4	1.7.1
2	Let p1=(1,2) and p2=(3,5) represent two objects,what will be the Euclidean distance? a) 5 **b) 3.61** c) 6.31 d) 2	1	2	4	1	1.7.1
3	Which of the following is cluster analysis? a) Simple segmentation **b) Grouping similar objects** c) Label classification d) Query results grouping	1	1	4	1	1.7.1
4	Which one of the following statements about the K-means clustering is incorrect? a. The goal of the k-means clustering is to partition (n) observation into (k) clusters b. K-means clustering can be defined as the method of quantization **c. The nearest neighbour is the same as the K-means** d. All of the above	1	L2	4	4	2.5.2
5	Which clustering technique requires a merging approach? a. Partitional b. **Hierarchical** c. Naive Bayes d. None of the mentioned	1	L2	4	4	1.7.1
6	Which one of the following can be defined as the data object which does not comply with the general behaviour (or the model of available data)? a. Evaluation Analysis **b. Outliner Analysis** c. Classification d. Prediction	1	L2	6	2	1.7.1

7	Euclidean distance measure is can also defined as _____ a. The process of finding a solution for a problem simply by enumerating all possible solutions according to some predefined order and then testing them **b. The distance between two points as calculated using the Pythagoras theorem** c. A stage of the KDD process in which new data is added to the existing selection. d. It is a kind of process of executing implicit, previously unknown and potentially useful information from data	1	L2	5	1	5.4.1
8	The analysis performed to uncover the interesting statistical correlation between associated -attributes value pairs are known as the _____. a. Mining of association **b. Mining of correlation** c. Mining of clusters d. Mining of Prediction	1	L2	4	4	2.5.2
9	The K means clustering algorithm fails to give good results in ____ **a. When the dataset contains outliers.** b. When the data points follow a non-convex shape. c. When the data points follow a convex shape. d. Both a and b	1	L2	4	4	2.5.2
10	In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?  a. 1 **b. 2** c. 3 d.4	1	L1	5	2	2.5.2
	Part – B (4 x 5 = 20 Marks)					
Answer All the Questions						
11	Write K-Medoids clustering algorithm with an example. **K-Medoids** (also called Partitioning Around Medoid) algorithm was proposed in 1987 by Kaufman and Rousseeuw. A medoid can be defined as a point in the cluster, whose dissimilarities with all the other points in the cluster are minimum. The dissimilarity of the medoid(Ci) and object(Pi) is calculated by using E = |Pi – Ci| **Algorithm:** 1. Initialize: select *k* random points out of the *n* data points as the medoids. 2. Associate each data point to the closest medoid by using any common distance metric methods. 3. While the cost decreases: For each medoid m, for each data o point which is not a medoid: • Swap m and o, associate each data point to the closest medoid, and recompute the cost. • If the total cost is more than that in the previous step, undo the swap.	5	L3	4	2	2.5.2

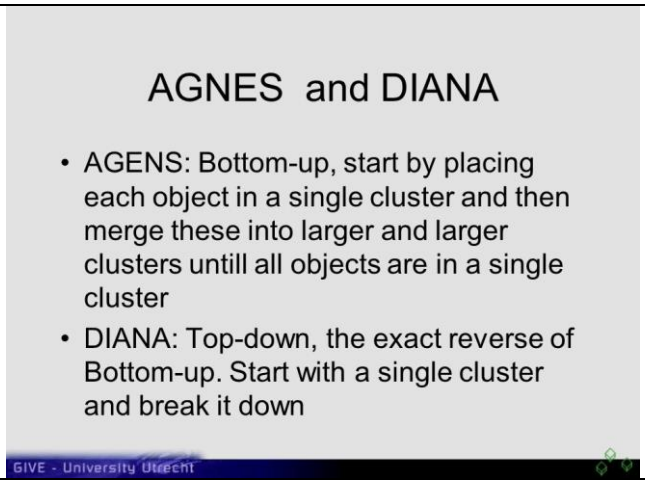
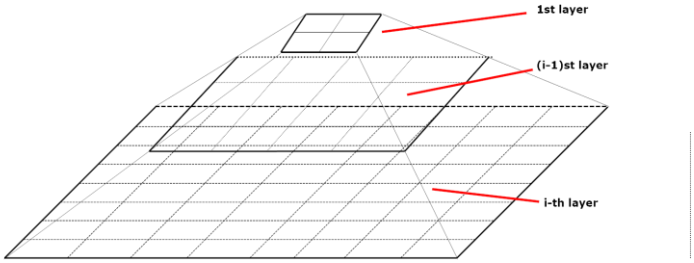
	X	Y
0	8	7
1	3	7
2	4	9
3	9	6
4	8	5
5	5	8
6	7	3
7	8	4
8	7	5
9	4	5

	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	3
1	3	7	3	8
2	4	9	4	9
3	9	6	6	3
4	8	5	4	1
5	5	8	4	7
6	7	3	5	2
7	8	4	-	-
8	7	5	3	2
9	4	5	-	-

Each point is assigned to that cluster whose dissimilarity is less. So, points 1, 2, and 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2. The New cost = (3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22 Swap Cost = New Cost – Previous Cost = 22 – 20 and **2 >0** As the swap cost is not less than zero, we undo the swap. Hence (4, 5) and (8, 5) are the final medoids. The clustering would be in the following way The **time complexity** is .



12	Differentiate between AGNES and DIANA algorithms.					

						
13	Discuss about STING method from grid based clustering algorithm. ▪ Wang, Yang and Muntz (VLDB'97) ▪ The spatial area is divided into rectangular cells ▪ There are several levels of cells corresponding to different levels of resolution  85 ■ Each cell at a high level is partitioned into a number of smaller cells in the next lower level ■ Statistical info of each cell is calculated and stored before hand and is used to answer queries ■ Parameters of higher level cells can be easily calculated from parameters of lower level cell ■ *count*, *mean*, *sd*, *min*, *max* ■ type of distribution—*normal*, *uniform*, etc. ■ Use a top-down approach to answer spatial data queries ■ Start from a pre-selected layer—typically with a small number of cells ■ For each cell in the current level compute the confidence interval ■ Remove the irrelevant cells from further consideration ■ When finish examining the current layer, proceed to the next lower level ■ Repeat this process until the bottom layer is reached All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected					
14	Explain different types of outlier. ■ Three kinds: *global*, *contextual* and *collective* outliers **1. Global outlier** (or point anomaly) ■ Object is O_g if it significantly deviates from the rest of the data set ■ Ex. Intrusion detection in computer networks ■ Issue: Find an appropriate measurement of deviation ■ **2. Collective Outliers**					

	■ A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers ■ Applications: E.g., *intrusion detection*: ■ When a number of computers keep sending denial-of-service packages to each other **3. Contextual outlier** (or *conditional outlier*) Object is O_c if it deviates significantly based on a selected context					
15.	Explain Data Mining for Financial data analysis Data Mining is a quite strong field to execute advanced examination of data as well as it carries off techniques and mechanisms from statistics and machine learning. Business intelligence and advanced analytics applications use the information which is generated by it which involves the analysis of verified data. Financial analysis of data is very important in order to analyze whether the business is stable and profitable to make a capital investment. Financial analysts focus their analysis on the balance sheet, cash flow statement, and income statement. Data mining techniques have been used to extract hidden patterns and predict future trends and behaviors in financial markets. Advanced statistical, mathematical and artificial intelligence techniques are typically required for mining such data, especially the high-frequency financial data					
	Part – C (2 x 10 = 20 Marks)					
11	Consider the Following data points to compute Cluster Values when K=3 using K-Means Clustering Algorithm: K= {X1(2,10),X2(2,5),X3(8,4),X4(5,8),X5(7,5),X6(6.4),X7(1,2),X8(4,9)}. **K-Means Clustering – Solved Example** • Suppose that the data mining task is to cluster points into three clusters, • where the points are • *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*, and *C1* as the center of each cluster, respectively.	10	L3	4	2	2.5.2

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10	0.00		3.61		8.06		1	
A2	2	5	5.00		4.24		3.16		3	
A3	8	4	8.49		5.00		7.28		2	
B1	5	8	3.61		0.00		7.21		2	
B2	7	5	7.07		3.61		6.71		2	
B3	6	4	7.21		4.12		5.39		2	
C1	1	2	8.06		7.21		0.00		3	
C2	4	9	2.24		1.41		7.62		2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A1	2	10	1.12		6.54		6.52		1	1
A2	2	5	4.61		4.51		1.58		3	3
A3	8	4	7.43		1.95		6.52		2	2
B1	5	8	2.50		3.13		5.70		2	1
B2	7	5	6.02		0.56		5.70		2	2
B3	6	4	6.26		1.35		4.53		2	2
C1	1	2	7.76		6.39		1.58		3	3
C2	4	9	1.12		4.51		6.04		1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

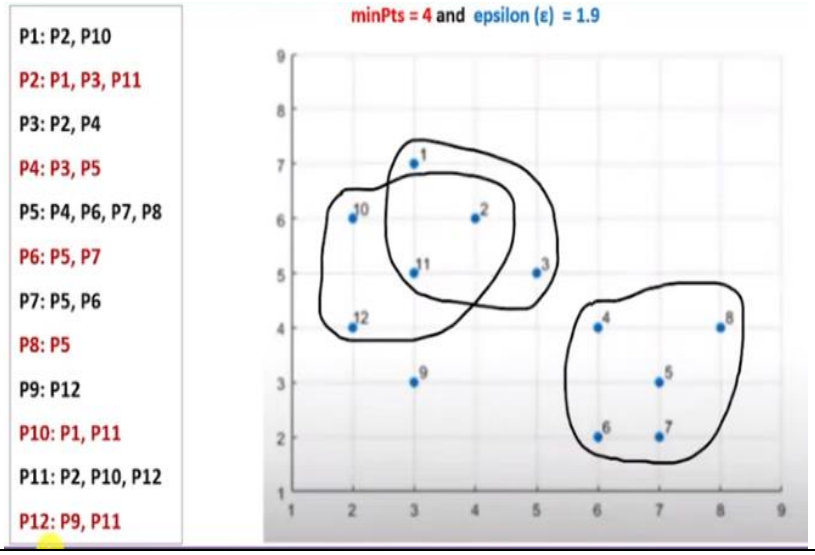
K-Means Clustering – Solved Example

Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	1
A2	2	5	4.33		5.04		1.58		3	3
A3	8	4	6.62		1.05		6.52		2	2
B1	5	8	1.67		4.18		5.70		1	1
B2	7	5	5.21		0.67		5.70		2	2
B3	6	4	5.52		1.05		4.53		2	2
C1	1	2	7.49		6.44		1.58		3	3
C2	4	9	0.33		5.55		6.04		1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

RegNo															

12	What is a DBSCAN? Apply DBSCAN algorithm to the given data points to create the cluster with minpts = 4, epsilon = 1.9 and p1(3,7),p2(4,6),p3(5,5),p4(6,4),p5(7,3),p6(6,2),p7(7,2),p8(8,4), p9(3,3),p10(2,6),p11(3,5),p12(2,4). minPts = 4 and epsilon (ε) = 1.9 		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
---	---	---	---	---	---	---	---	---	---	---	---	---
P1	0											
P2	1.41	0										
P3	2.83	1.41	0									
P4	4.24	2.83	1.41	0								
P5	5.66	4.24	2.83	1.41	0							
P6	5.83	4.47	3.16	2.00	1.41	0						
P7	6.40	5.00	3.61	2.24	1.00	1.00	0					
P8	5.83	4.47	3.16	2.00	1.41	2.83	2.24	0				
P9	4.00	3.16	2.83	3.16	4.00	3.16	4.12	5.10	0			
P10	1.41	2.00	3.16	4.47	5.83	5.66	6.40	6.32	3.16	0		
P11	2.00	1.41	2.00	3.16	4.47	4.24	5.00	5.10	2.00	1.41	0	
P12	3.16	2.83	3.16	4.00	5.10	4.47	5.39	6.00	1.41	2.00	1.41	0 | P1: P2, P10 P2: P1, P3, P11 P3: P2, P4 P4: P3, P5 P5: P4, P6, P7, P8 P6: P5, P7 P7: P5, P6 P8: P5 P9: P12 P10: P1, P11 P11: P2, P10, P12 P12: P9, P11 	10	L2	4	2	2.5.2
13	Discuss about attributes of healthcare recommendation system using Data mining approach with example. Healthcare recommender systems are meant to provide accurate and relevant predictions to the patients. It is very difficult for people to explore various online sources to find some useful recommendations as per their medical conditions. Patients are categorized into different groups based on their profiles and then rules predicting the medical condition of each group are mined. The proposed approach is unique in the way that it provides accurate treatments to the patients in the form of recommendations based on content based matching. It also considers the preferences of the patient, which are stored in the system as mined rules or estimated from the medical history of patient. The results of experimental setup also demonstrate that the proposed system provides more accurate outcomes over other healthcare recommendation systems.	10	L2	6	4	2.7.1

14	Interpret the supervised method for detecting the outlier.
	
	■ Modeling outlier detection as a classification problem
	  ■ Samples examined by domain experts used for training & testing
	■ Methods for Learning a classifier for outlier detection effectively:
	  ■ Model normal objects & report those not matching the model as outliers, or
	  ■ Model outliers and treat those not matching the model as normal
	■ Challenges
	  ■ Imbalanced classes, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers
	  ■ Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers)
		10	L2	5	4	1.7.1