Reg. No.															

SRM Institute of Science and Technology

College of Engineering and Technology

School of Computing

Batch -SET A

DEPARTMENT OF COMPUTING TECHNOLOGIES

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu

Academic Year: 2023-2024 (ODD)

Test: CLAT-3 **Date:** 09.11.2023

Course Code & Title: 18CSE355T - Data Mining And Analytics **Duration:** 2 Periods

Year & Sem: III & IV Year & 05th & 07th Semester **Max. Marks:** 50 Marks

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

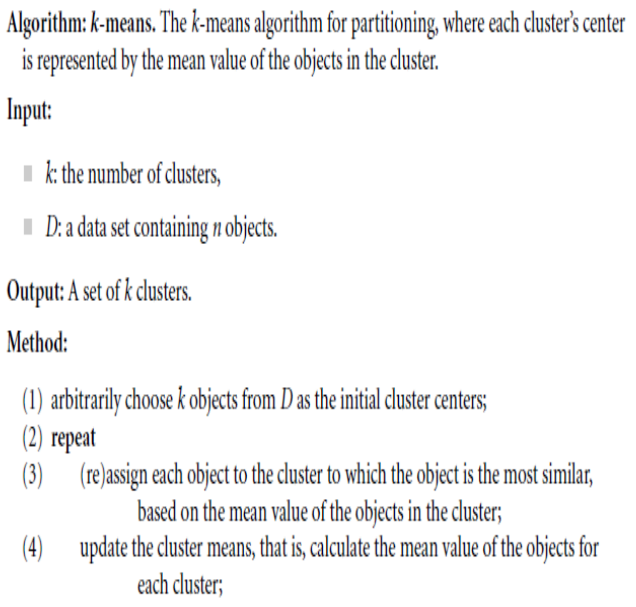
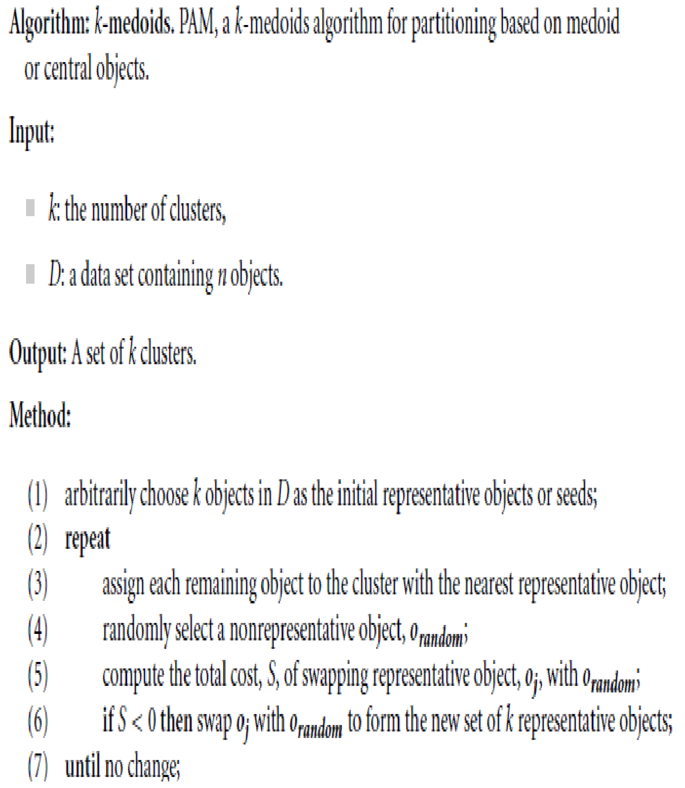
(10 x 1 = 10 Marks)

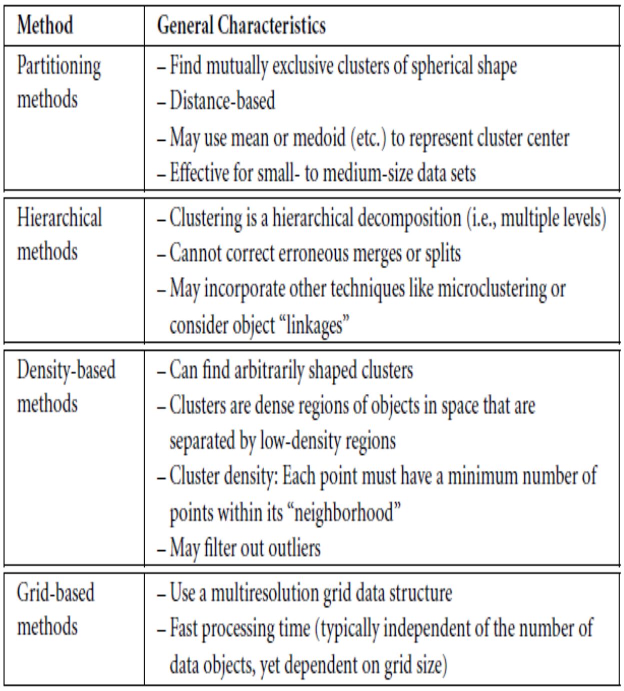
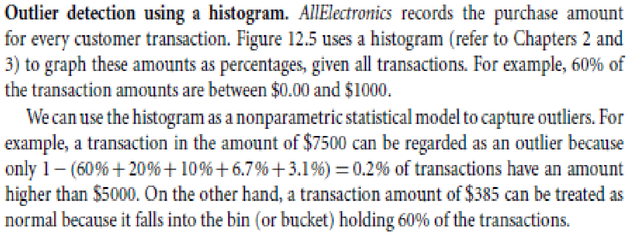
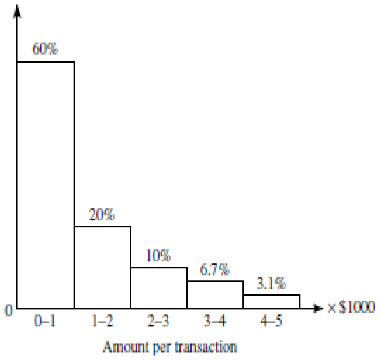
Answer all questions. The duration for answering the part A is 15 minutes (MCQ Answer sheet will be collected after 15 minutes)

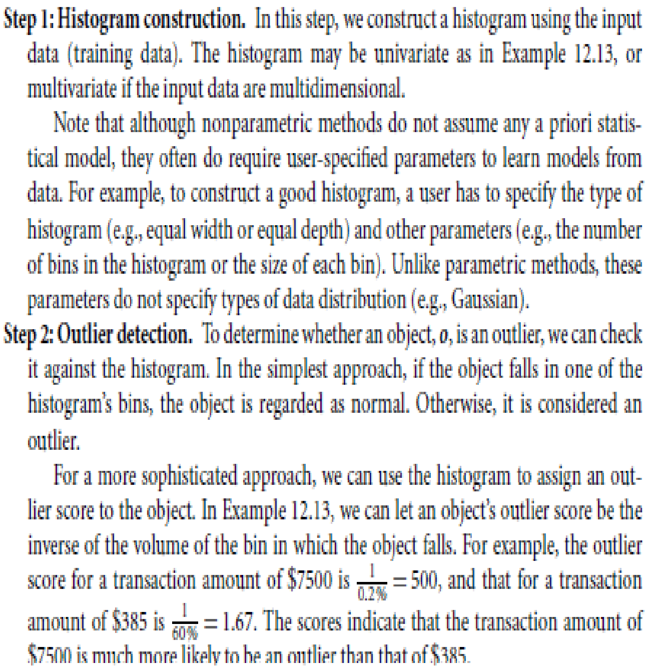
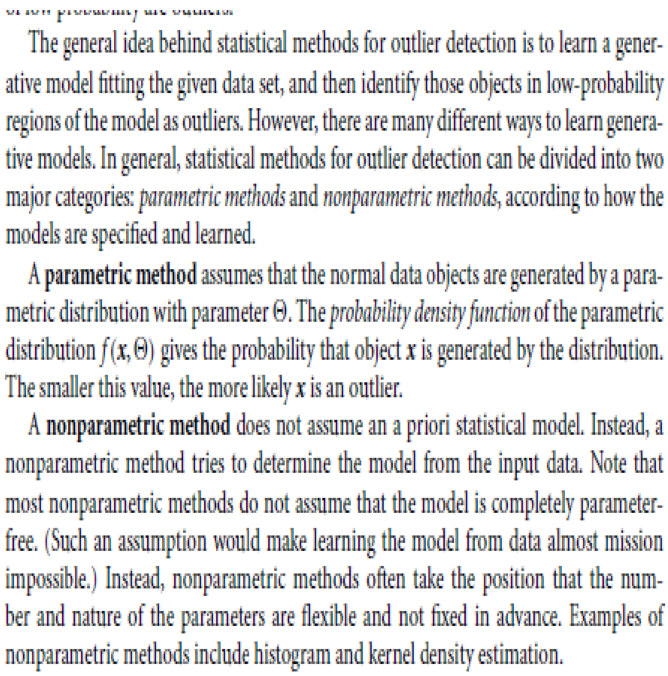
Q. No	Question	Marks	BL	CO	PO	PI Code
1	**Which of the below one is not an typical requirements of clustering in data mining**. a. Scalability b. Ability to deal with different types of attributes c. Ability to deal with noisy data d. Decremental clustering and insensitivity to input order	1	6	4	1	1.7.1
2	**Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k ≤ n. How many groups in this conditions?** a. k groups b. n groups c. Kn d. n^K	1	6	4	2	1.7.1
3	**Which of the below method is used for the bottom-up approach.** a. K mean method b. Agglomerative method c. Divisive method d. Partitioning method	1	6	4	1	1.7.1

4	**What is the time complexity of K means clustering algorithm.** a. O(K) b. O(n-1) c. O(n) d. O(nkt)	1	4	4	2	1.7.1
5	**--------- uses the notions of clustering feature to summarize a cluster, and clustering** feature tree (CF-tree) to represent a cluster hierarchy. a. BIRCH b. DBSCAN c. STING d. CLIQUE	1	6	4	1	1.7.1
6	**Which of the below is one of the challenges of outlier detection.** a. Noises are occurred rarely when the outlier process start b. Duplicate are allowed for outlier detection c. This is most sensitive processing d. The border between data normality and abnormality (outliers) is often not clear cut	1	6	5	1	2.5.2
7	**Where we can apply the Grubb's test?** a. Multivariate outlier detection b. Bi variate outlier detection c. Univariate outlier detection d. Regression based detection	1	6	5	1	2.5.2
8	**Which will play the major role in proximity-based outlier detection?** a. Density of the neighbourhood b. Mean vector of the neighbourhood c. Radius of the neighbourhood d. Threshold of the neighbourhood	1	6	5	1	2.5.2
9	**which method is used to find the intrusion detection in clustering-based outlier detection?** a. Bootstrap b. Angle-based outlier c. Anomalies detection d. CLIQUE	1	6	5	1	2.5.2
10	**In which case objects are labelled as "normal" or "outlier" are not available?** a. Supervised method b. Unsupervised method c. Semi-supervised method d. Hybrid method	1	6	5	1	2.5.2

Part – B
(4 x 5 = 20 Marks)
Answer any 4 Questions

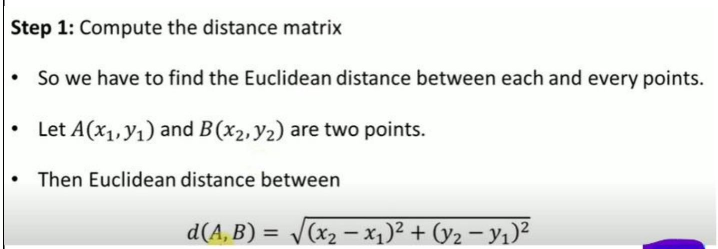
11	Write the algorithm for K means and K medoids.  **Algorithm:** *k*-**means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster. **Input:** - *k*: the number of clusters, - *D*: a data set containing *n* objects. **Output:** A set of *k* clusters. **Method:** (1) arbitrarily choose *k* objects from *D* as the initial cluster centers; (2) **repeat** (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) update the cluster means, that is, calculate the mean value of the objects for each cluster;  **Algorithm:** *k*-**medoids.** PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects. **Input:** - *k*: the number of clusters, - *D*: a data set containing *n* objects. **Output:** A set of *k* clusters. **Method:** (1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds; (2) **repeat** (3) assign each remaining object to the cluster with the nearest representative object; (4) randomly select a nonrepresentative object, o_{random}; (5) compute the total cost, *S*, of swapping representative object, o_j, with o_{random}; (6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of *k* representative objects; (7) **until** no change;	5	3	4	3	8.4.1
12	Discuss the basic characteristic of clustering methods	5	3	4	3	1.7.1

<table>
<tr>
<td></td>
<td><table>
<tr><th>Method</th><th>General Characteristics</th></tr>
<tr><td>Partitioning methods</td><td>– Find mutually exclusive clusters of spherical shape
– Distance-based
– May use mean or medoid (etc.) to represent cluster center
– Effective for small- to medium-size data sets</td></tr>
<tr><td>Hierarchical methods</td><td>– Clustering is a hierarchical decomposition (i.e., multiple levels)
– Cannot correct erroneous merges or splits
– May incorporate other techniques like microclustering or consider object “linkages”</td></tr>
<tr><td>Density-based methods</td><td>– Can find arbitrarily shaped clusters
– Clusters are dense regions of objects in space that are separated by low-density regions
– Cluster density: Each point must have a minimum number of points within its “neighborhood”
– May filter out outliers</td></tr>
<tr><td>Grid-based methods</td><td>– Use a multiresolution grid data structure
– Fast processing time (typically independent of the number of data objects, yet dependent on grid size)</td></tr>
</table></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Discuss the outlier detection using histogram-based approach with example

Outlier detection using a histogram. *AllElectronics* records the purchase amount for every customer transaction. Figure 12.5 uses a histogram (refer to Chapters 2 and 3) to graph these amounts as percentages, given all transactions. For example, 60% of the transaction amounts are between $0.00 and $1000.
We can use the histogram as a nonparametric statistical model to capture outliers. For example, a transaction in the amount of $7500 can be regarded as an outlier because only 1 − (60% + 20% + 10% + 6.7% + 3.1%) = 0.2% of transactions have an amount higher than $5000. On the other hand, a transaction amount of $385 can be treated as normal because it falls into the bin (or bucket) holding 60% of the transactions.
 </td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1.7.1</td>
</tr>
</table>

	 **Step 1: Histogram construction.** In this step, we construct a histogram using the input data (training data). The histogram may be univariate as in Example 12.13, or multivariate if the input data are multidimensional. Note that although nonparametric methods do not assume any a priori statistical model, they often do require user-specified parameters to learn models from data. For example, to construct a good histogram, a user has to specify the type of histogram (e.g., equal width or equal depth) and other parameters (e.g., the number of bins in the histogram or the size of each bin). Unlike parametric methods, these parameters do not specify types of data distribution (e.g., Gaussian). **Step 2: Outlier detection.** To determine whether an object, ***o***, is an outlier, we can check it against the histogram. In the simplest approach, if the object falls in one of the histogram's bins, the object is regarded as normal. Otherwise, it is considered an outlier. For a more sophisticated approach, we can use the histogram to assign an outlier score to the object. In Example 12.13, we can let an object's outlier score be the inverse of the volume of the bin in which the object falls. For example, the outlier score for a transaction amount of \$7500 is $\frac{1}{0.2\%} = 500$, and that for a transaction amount of \$385 is $\frac{1}{60\%} = 1.67$. The scores indicate that the transaction amount of \$7500 is much more likely to be an outlier than that of \$385.					
14	Explain the various steps involved in statistical based outlier detection with suitable diagram  of low probability are outliers. The general idea behind statistical methods for outlier detection is to learn a generative model fitting the given data set, and then identify those objects in low-probability regions of the model as outliers. However, there are many different ways to learn generative models. In general, statistical methods for outlier detection can be divided into two major categories: *parametric methods* and *nonparametric methods*, according to how the models are specified and learned. A **parametric method** assumes that the normal data objects are generated by a parametric distribution with parameter Θ. The *probability density function* of the parametric distribution $f(\boldsymbol{x}, \Theta)$ gives the probability that object $\boldsymbol{x}$ is generated by the distribution. The smaller this value, the more likely $\boldsymbol{x}$ is an outlier. A **nonparametric method** does not assume an a priori statistical model. Instead, a nonparametric method tries to determine the model from the input data. Note that most nonparametric methods do not assume that the model is completely parameter-free. (Such an assumption would make learning the model from data almost mission impossible.) Instead, nonparametric methods often take the position that the number and nature of the parameters are flexible and not fixed in advance. Examples of nonparametric methods include histogram and kernel density estimation.	5	3	5	4	2.6.4

<table>
<tr>
<td>15</td>
<td>
Solve the single link technique using below table

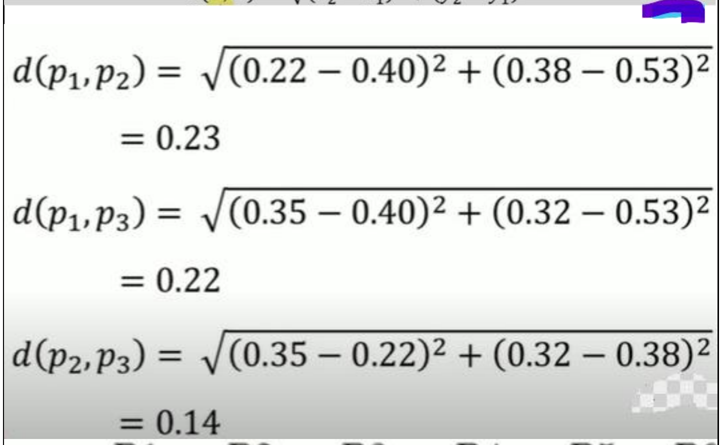
<table>
<tr><td>Sample no</td><td>X</td><td>y</td></tr>
<tr><td>P1</td><td>0.40</td><td>0.53</td></tr>
<tr><td>P2</td><td>0.22</td><td>0.38</td></tr>
<tr><td>P3</td><td>0.35</td><td>0.32</td></tr>
<tr><td>P4</td><td>0.26</td><td>0.19</td></tr>
<tr><td>= (
5</td><td>0.08</td><td>0.41</td></tr>
<tr><td>P6</td><td>0.45</td><td>0.30</td></tr>
</table>



Step 1: Compute the distance matrix

- So we have to find the Euclidean distance between each and every points.
- Let $A(x_1, y_1)$ and $B(x_2, y_2)$ are two points.
- Then Euclidean distance between

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



$$d(p_1, p_2) = \sqrt{(0.22 - 0.40)^2 + (0.38 - 0.53)^2}$$

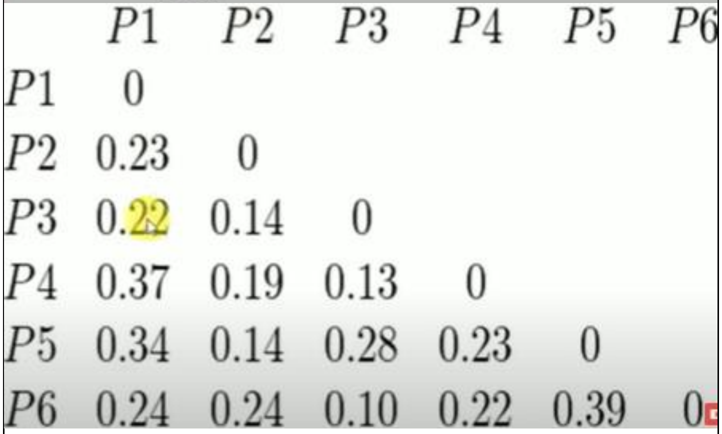
$$= 0.23$$

$$d(p_1, p_3) = \sqrt{(0.35 - 0.40)^2 + (0.32 - 0.53)^2}$$

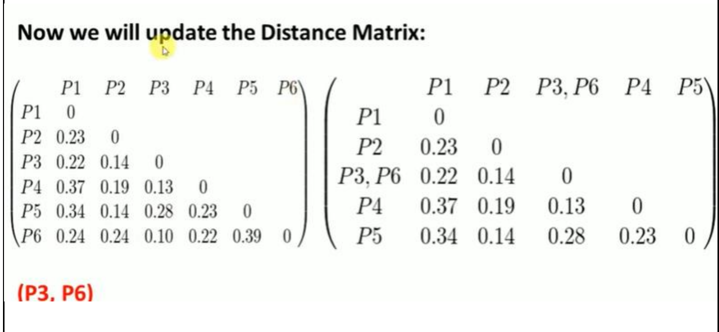
$$= 0.22$$

$$d(p_2, p_3) = \sqrt{(0.35 - 0.22)^2 + (0.32 - 0.38)^2}$$

$$= 0.14$$



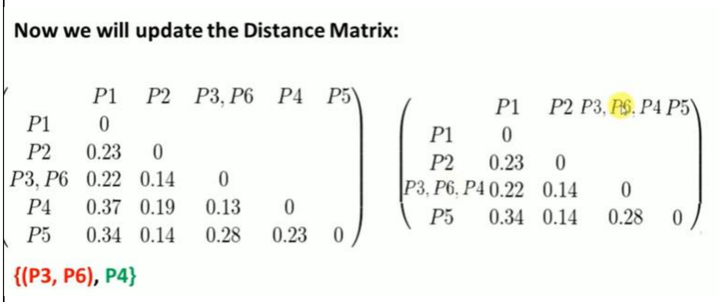
$$\begin{array}{ccccccc} & P1 & P2 & P3 & P4 & P5 & P6 \\ P1 & 0 & & & & & \\ P2 & 0.23 & 0 & & & & \\ P3 & 0.22 & 0.14 & 0 & & & \\ P4 & 0.37 & 0.19 & 0.13 & 0 & & \\ P5 & 0.34 & 0.14 & 0.28 & 0.23 & 0 & \\ P6 & 0.24 & 0.24 & 0.10 & 0.22 & 0.39 & 0 \end{array}$$
</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>2.6.4</td>
</tr>
</table>



Now we will update the Distance Matrix:

$$\left(\begin{array}{ccccccc} & P1 & P2 & P3 & P4 & P5 & P6 \\ P1 & 0 & & & & & \\ P2 & 0.23 & 0 & & & & \\ P3 & 0.22 & 0.14 & 0 & & & \\ P4 & 0.37 & 0.19 & 0.13 & 0 & & \\ P5 & 0.34 & 0.14 & 0.28 & 0.23 & 0 & \\ P6 & 0.24 & 0.24 & 0.10 & 0.22 & 0.39 & 0 \end{array}\right) \left(\begin{array}{cccccc} & P1 & P2 & P3, P6 & P4 & P5 \\ P1 & 0 & & & & \\ P2 & 0.23 & 0 & & & \\ P3, P6 & 0.22 & 0.14 & 0 & & \\ P4 & 0.37 & 0.19 & 0.13 & 0 & \\ P5 & 0.34 & 0.14 & 0.28 & 0.23 & 0 \end{array}\right)$$

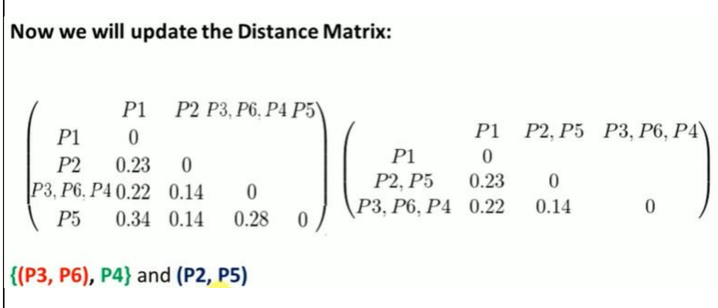
(P3. P6)



Now we will update the Distance Matrix:

$$\left(\begin{array}{cccccc} & P1 & P2 & P3, P6 & P4 & P5 \\ P1 & 0 & & & & \\ P2 & 0.23 & 0 & & & \\ P3, P6 & 0.22 & 0.14 & 0 & & \\ P4 & 0.37 & 0.19 & 0.13 & 0 & \\ P5 & 0.34 & 0.14 & 0.28 & 0.23 & 0 \end{array}\right) \left(\begin{array}{ccccc} & P1 & P2 & P3, P6, P4 & P5 \\ P1 & 0 & & & \\ P2 & 0.23 & 0 & & \\ P3, P6, P4 & 0.22 & 0.14 & 0 & \\ P5 & 0.34 & 0.14 & 0.28 & 0 \end{array}\right)$$

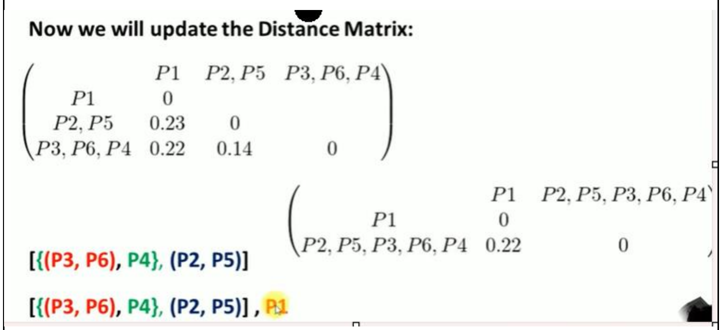
{(P3, P6), P4}



Now we will update the Distance Matrix:

$$\left(\begin{array}{ccccc} & P1 & P2 & P3, P6, P4 & P5 \\ P1 & 0 & & & \\ P2 & 0.23 & 0 & & \\ P3, P6, P4 & 0.22 & 0.14 & 0 & \\ P5 & 0.34 & 0.14 & 0.28 & 0 \end{array}\right) \left(\begin{array}{cccc} & P1 & P2, P5 & P3, P6, P4 \\ P1 & 0 & & \\ P2, P5 & 0.23 & 0 & \\ P3, P6, P4 & 0.22 & 0.14 & 0 \end{array}\right)$$

{(P3, P6), P4} and **(P2, P5)**



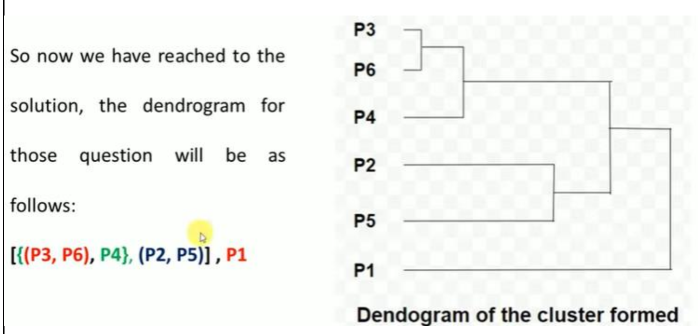
Now we will update the Distance Matrix:

$$\left(\begin{array}{cccc} & P1 & P2, P5 & P3, P6, P4 \\ P1 & 0 & & \\ P2, P5 & 0.23 & 0 & \\ P3, P6, P4 & 0.22 & 0.14 & 0 \end{array}\right)$$

$$\left(\begin{array}{ccc} & P1 & P2, P5, P3, P6, P4 \\ P1 & 0 & \\ P2, P5, P3, P6, P4 & 0.22 & 0 \end{array}\right)$$

[{(P3, P6), P4}, (P2, P5)]

[{(P3, P6), P4}, (P2, P5)] , P1

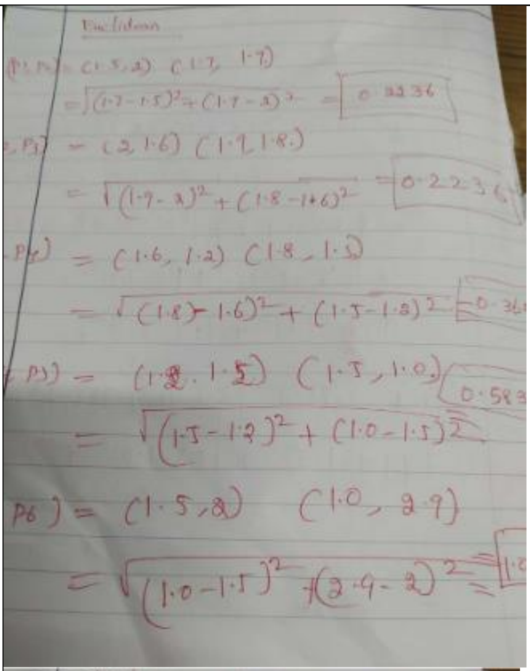
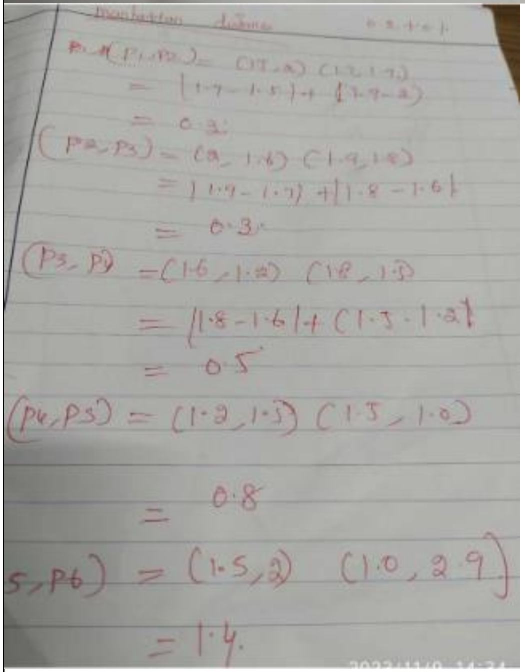
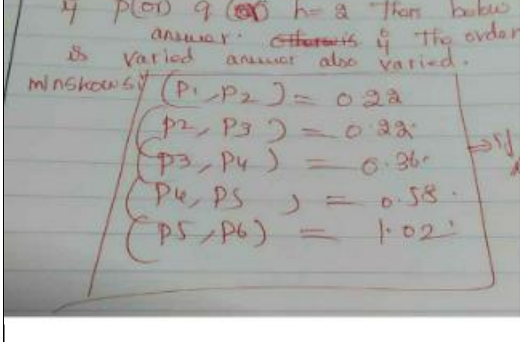


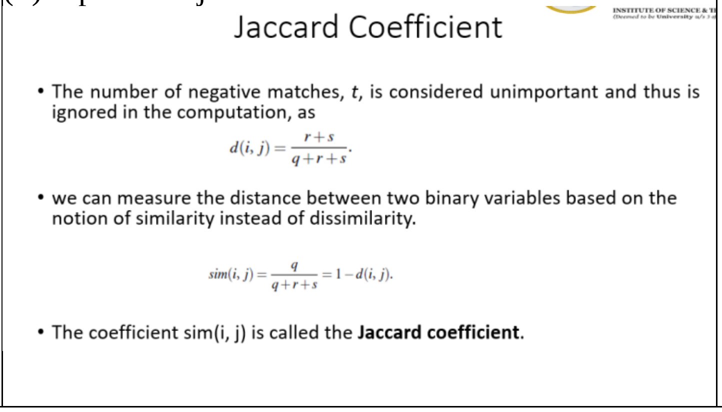
So now we have reached to the solution, the dendrogram for those question will be as follows:

[{(P3, P6), P4}, (P2, P5)] , P1

Dendogram of the cluster formed

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="7">Part – B
(2 x 10 = 20 Marks)</td></tr>
<tr><td>16</td><td>(i)Find the Euclidean distance ,Manhattan distance , Minkowski distance for below table mark split up(3+3+2)
<table>
<tr><td></td><td>X1</td><td>Y1</td><td></td><td></td></tr>
<tr><td>P1</td><td>1.5</td><td>1.7</td><td></td><td></td></tr>
<tr><td>P2</td><td>2</td><td>1.9</td><td></td><td></td></tr>
<tr><td>P3</td><td>1.6</td><td>1.8</td><td></td><td></td></tr>
<tr><td>P4</td><td>1.2</td><td>1.5</td><td></td><td></td></tr>
<tr><td>P5</td><td>1.5</td><td>1.0</td><td></td><td></td></tr>
<tr><td>P6</td><td>2</td><td>2.9</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
</table></td><td>8
2</td><td>6</td><td>5</td><td>2</td><td>1.7.1</td></tr>
</table>

	 Euclidean $(P_1, P_2) = (1.5, 2)\ (1.7, 1.7)$ $= \sqrt{(1.7-1.5)^2 + (1.7-2)^2} = \boxed{0.2236}$ $P_3) = (2, 1.6)\ (1.7, 1.8)$ $= \sqrt{(1.7-2)^2 + (1.8-1.6)^2} = \boxed{0.2236}$ $P_4) = (1.6, 1.2)\ (1.8, 1.5)$ $= \sqrt{(1.8-1.6)^2 + (1.5-1.2)^2} = \boxed{0.36}$ $P_5) = (1.2, 1.5)\ (1.5, 1.0)\ \boxed{0.583}$ $= \sqrt{(1.5-1.2)^2 + (1.0-1.5)^2}$ $P_6) = (1.5, 2)\ (1.0, 2.9)$ $= \sqrt{(1.0-1.5)^2 + (2.9-2)^2} = \boxed{1.0}$  manhattan distance $(P_1, P_2) = (1.5, 2)\ (1.7, 1.7)$ $= \lvert 1.7-1.5 \rvert + \lvert 1.7-2 \rvert$ $= 0.2$ $(P_2, P_3) = (2, 1.6)\ (1.7, 1.8)$ $= \lvert 1.7-1.7 \rvert + \lvert 1.8-1.6 \rvert$ $= 0.3$ $(P_3, P_4) = (1.6, 1.2)\ (1.8, 1.5)$ $= \lvert 1.8-1.6 \rvert + \lvert 1.5-1.2 \rvert$ $= 0.5$ $(P_4, P_5) = (1.2, 1.5)\ (1.5, 1.0)$ $= 0.8$ $5, P_6) = (1.5, 2)\ (1.0, 2.9)$ $= 1.4$ Minkowski changes or order answer you can check here= https://www.redcrab-software.com/en/Calculator/Distance/Minkowski  if p(or) q (or) h = 2 then below answer. ~~Otherwise~~ if the order is varied answer also varied. minskowsky $(P_1, P_2) = 0.22$ $(P_2, P_3) = 0.22$ $(P_3, P_4) = 0.36$ $(P_4, P_5) = 0.58$ $(P_5, P_6) = 1.02$					

	(ii)Explain the jaccard coefficient with formula  Jaccard Coefficient • The number of negative matches, t, is considered unimportant and thus is ignored in the computation, as $d(i, j) = \frac{r+s}{q+r+s}$. • we can measure the distance between two binary variables based on the notion of similarity instead of dissimilarity. $sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j)$. • The coefficient sim(i, j) is called the **Jaccard coefficient**.					
[OR]						
17	Explain the density-based clustering method with pseudo code, suitable diagram with formulas	10	6	6	1	8.4.1

10.4 Density-Based Methods

Partitioning and hierarchical methods are designed to find spherical-shaped clusters. They have difficulty finding clusters of arbitrary shape such as the "S" shape and oval clusters in Figure 10.13. Given such data, they would likely inaccurately identify convex regions, where noise or outliers are included in the clusters.

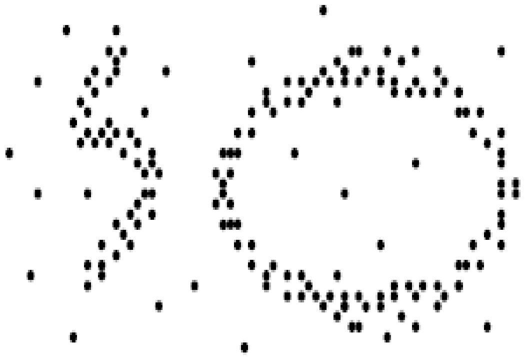
To find clusters of arbitrary shape, alternatively, we can model clusters as dense regions in the data space, separated by sparse regions. This is the main strategy behind *density-based clustering methods*, which can discover clusters of nonspherical shape. In this section, you will learn the basic techniques of density-based clustering by studying three representative methods, namely, DBSCAN (Section 10.4.1), OPTICS (Section 10.4.2), and DENCLUE (Section 10.4.3).

10.4.1 DBSCAN: Density-Based Clustering Based on Connected Regions with High Density

"How can we find dense regions in density-based clustering?" The *density* of an object o can be measured by the number of objects close to o. **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) finds *core objects*, that is, objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters.

"How does DBSCAN *quantify the neighborhood of an object?"* A user-specified parameter $\epsilon > 0$ is used to specify the radius of a neighborhood we consider for every object. The **ϵ-neighborhood** of an object o is the space within a radius ϵ centered at o.

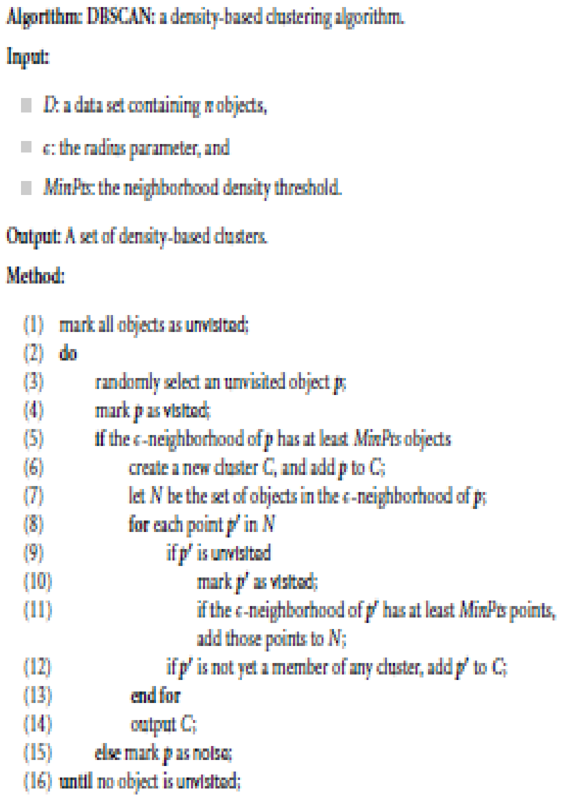
Due to the fixed neighborhood size parameterized by ϵ, the **density of a neighborhood** can be measured simply by the number of objects in the neighborhood. To determine whether a neighborhood is dense or not, DBSCAN uses another user-specified parameter, *MinPts*, which specifies the density threshold of dense regions. An object is a **core object** if the ϵ-neighborhood of the object contains at least *MinPts* objects. Core objects are the pillars of dense regions.



Given a set, D, of objects, we can identify all core objects with respect to the given parameters, ϵ and *MinPts*. The clustering task is therein reduced to using core objects and their neighborhoods to form dense regions, where the dense regions are clusters. For a core object q and an object p, we say that p is **directly density-reachable** from q (with respect to ϵ and *MinPts*) if p is within the ϵ-neighborhood of q. Clearly, an object p is directly density-reachable from another object q if and only if q is a core object and p is in the ϵ-neighborhood of q. Using the directly density-reachable relation, a core object can "bring" all objects from its ϵ-neighborhood into a dense region.

"How can we assemble a large dense region using small dense regions centered by core objects?" In DBSCAN, p is **density-reachable** from q (with respect to ϵ and *MinPts* in D) if there is a chain of objects $p_1, \ldots, p_n$, such that $p_1 = q$, $p_n = p$, and p_{i+1} is directly density-reachable from p_i with respect to ϵ and *MinPts*, for $1 \leq i \leq n$, $p_i \in D$. Note that density-reachability is not an equivalence relation because it is not symmetric. If both o_1 and o_2 are core objects and o_1 is density-reachable from o_2, then o_2 is density-reachable from o_1. However, if o_2 is a core object but o_1 is not, then o_1 may be density-reachable from o_2, but not vice versa.

To connect core objects as well as their neighbors in a dense region, **DBSCAN** uses the notion of density-connectedness. Two objects $p_1, p_2 \in D$ are **density-connected** with respect to ϵ and *MinPts* if there is an object $q \in D$ such that both p_1 and p_2 are density-reachable from q with respect to ϵ and *MinPts*. Unlike density-reachability, density-connectedness is an equivalence relation. It is easy to show that, for objects o_1, o_2, and o_3, if o_1 and o_2 are density-connected, and o_2 and o_3 are density-connected, then so are o_1 and o_3.



Algorithm: DBSCAN: a density-based clustering algorithm.

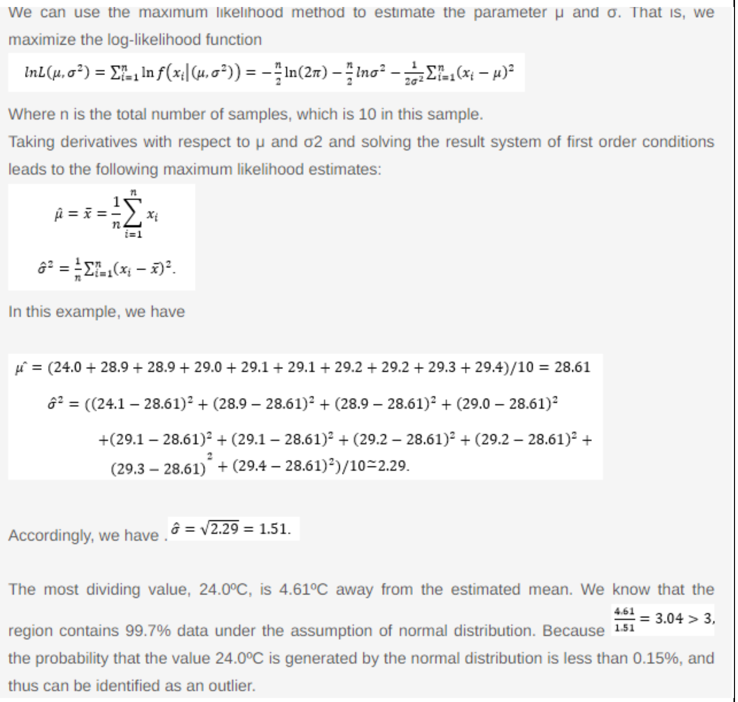
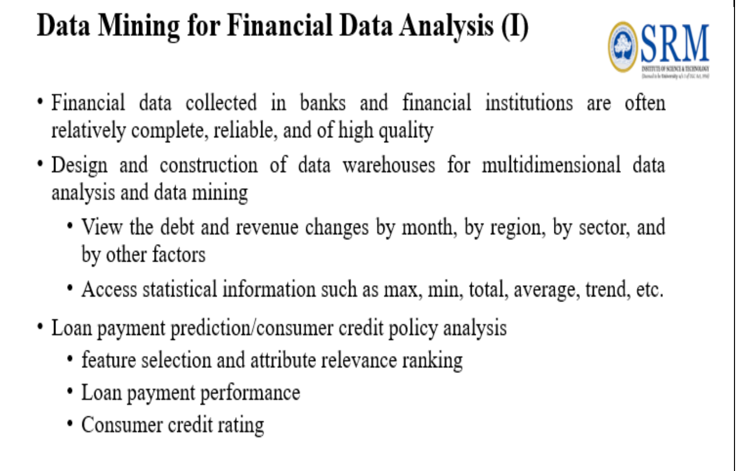
Input:

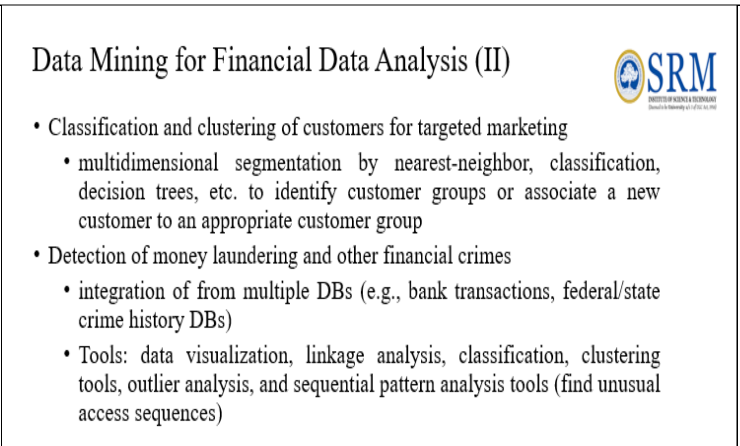
- D: a data set containing n objects,
- ϵ: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

```
mark all objects as unvisited;
do
    randomly select an unvisited object $p$;
    mark $p$ as visited;
    if the $\epsilon$-neighborhood of $p$ has at least $MinPts$ objects
        create a new cluster $C$, and add $p$ to $C$;
        let $N$ be the set of objects in the $\epsilon$-neighborhood of $p$;
        for each point $p'$ in $N$
            if $p'$ is unvisited
                mark $p'$ as visited;
                if the $\epsilon$-neighborhood of $p'$ has at least $MinPts$ points,
                add those points to $N$;
            if $p'$ is not yet a member of any cluster, add $p'$ to $C$;
        end for
        output $C$;
    else mark $p$ as noise;
until no object is unvisited;
```

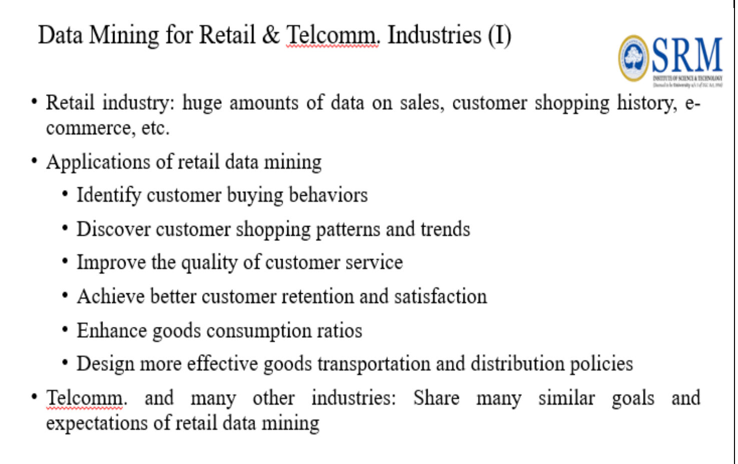
18	A city’s average temperature values in July in the last 10 years are, in value-ascending order, 24.0°C, 28.9°C, 28.9°C, 29.0°C, 29.1°C, 29.1°C, 29.2°C, 29.2°C, 29.3°C and 29.4°C. Let’s assume that the average temperature follows a normal distribution, which is determined by two parameters: the mean, μ, and the standard deviation, σ.Use the maximum likelihood method to estimate the parameter μ and σ  We can use the maximum likelihood method to estimate the parameter μ and σ. That is, we maximize the log-likelihood function $lnL(\mu,\sigma^2) = \sum_{i=1}^{n} \ln f\left(x_i \middle| (\mu,\sigma^2)\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2} ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$ Where n is the total number of samples, which is 10 in this sample. Taking derivatives with respect to μ and σ2 and solving the result system of first order conditions leads to the following maximum likelihood estimates: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$ In this example, we have $\hat{\mu} = (24.0 + 28.9 + 28.9 + 29.0 + 29.1 + 29.1 + 29.2 + 29.2 + 29.3 + 29.4)/10 = 28.61$ $\hat{\sigma}^2 = ((24.1 - 28.61)^2 + (28.9 - 28.61)^2 + (28.9 - 28.61)^2 + (29.0 - 28.61)^2$ $+(29.1 - 28.61)^2 + (29.1 - 28.61)^2 + (29.2 - 28.61)^2 + (29.2 - 28.61)^2 +$ $(29.3 - 28.61)^2 + (29.4 - 28.61)^2)/10 \simeq 2.29.$ Accordingly, we have . $\hat{\sigma} = \sqrt{2.29} = 1.51.$ The most dividing value, 24.0ºC, is 4.61ºC away from the estimated mean. We know that the region contains 99.7% data under the assumption of normal distribution. Because $\frac{4.61}{1.51} = 3.04 > 3,$ the probability that the value 24.0ºC is generated by the normal distribution is less than 0.15%, and thus can be identified as an outlier.	10	4	5	2	8.4.1
	[OR]					
19	How data mining involved in the below application. (i)Data Mining for Financial data analysis  **Data Mining for Financial Data Analysis (I)** • Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality • Design and construction of data warehouses for multidimensional data analysis and data mining • View the debt and revenue changes by month, by region, by sector, and by other factors • Access statistical information such as max, min, total, average, trend, etc. • Loan payment prediction/consumer credit policy analysis • feature selection and attribute relevance ranking • Loan payment performance • Consumer credit rating	2 2 2 2 2	6	6	4	8.4.1



Data Mining for Financial Data Analysis (II)

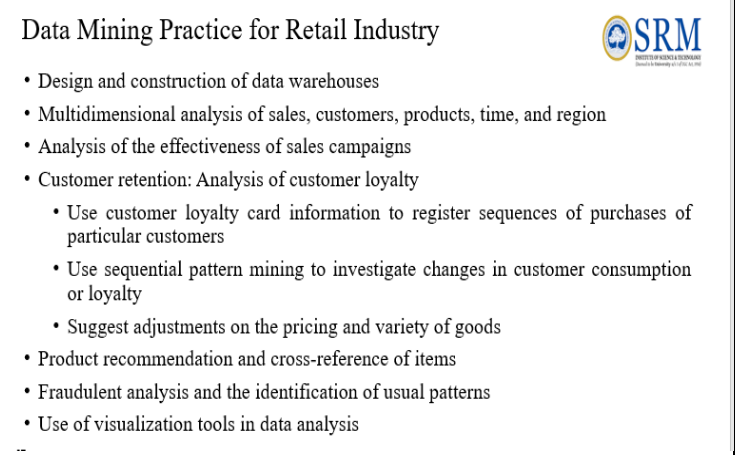
- Classification and clustering of customers for targeted marketing
  - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
  - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
  - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

(ii)Data Mining for Retail and Telecommunication Industries



Data Mining for Retail & Telcomm. Industries (I)

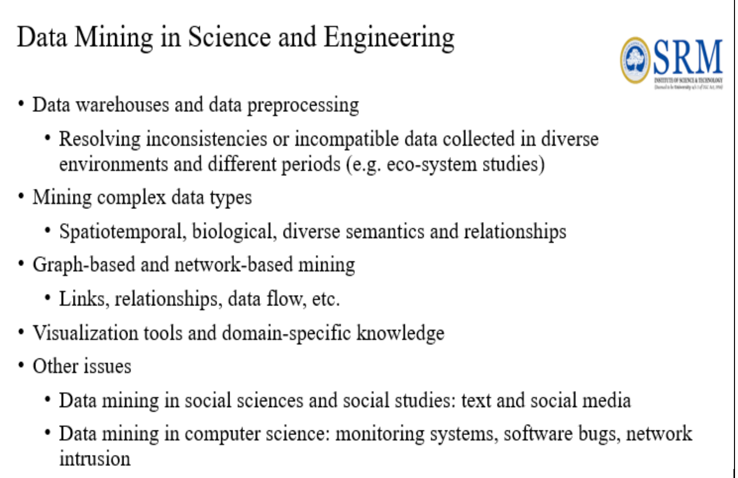
- Retail industry: huge amounts of data on sales, customer shopping history, e-commerce, etc.
- Applications of retail data mining
  - Identify customer buying behaviors
  - Discover customer shopping patterns and trends
  - Improve the quality of customer service
  - Achieve better customer retention and satisfaction
  - Enhance goods consumption ratios
  - Design more effective goods transportation and distribution policies
- Telcomm. and many other industries: Share many similar goals and expectations of retail data mining



Data Mining Practice for Retail Industry

- Design and construction of data warehouses
- Multidimensional analysis of sales, customers, products, time, and region
- Analysis of the effectiveness of sales campaigns
- Customer retention: Analysis of customer loyalty
  - Use customer loyalty card information to register sequences of purchases of particular customers
  - Use sequential pattern mining to investigate changes in customer consumption or loyalty
  - Suggest adjustments on the pricing and variety of goods
- Product recommendation and cross-reference of items
- Fraudulent analysis and the identification of usual patterns
- Use of visualization tools in data analysis

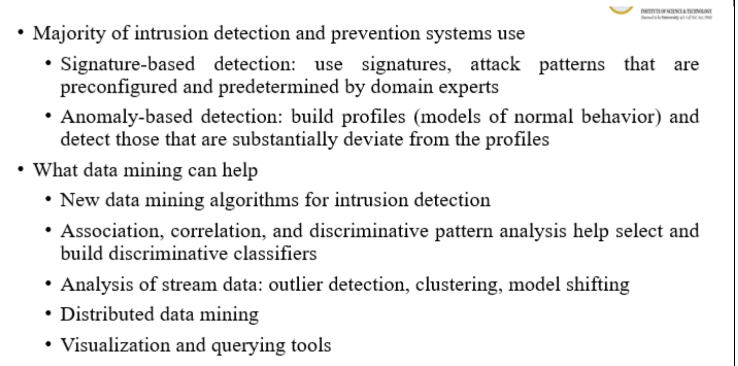
(iii)Data Mining in Science and Engineering



Data Mining in Science and Engineering

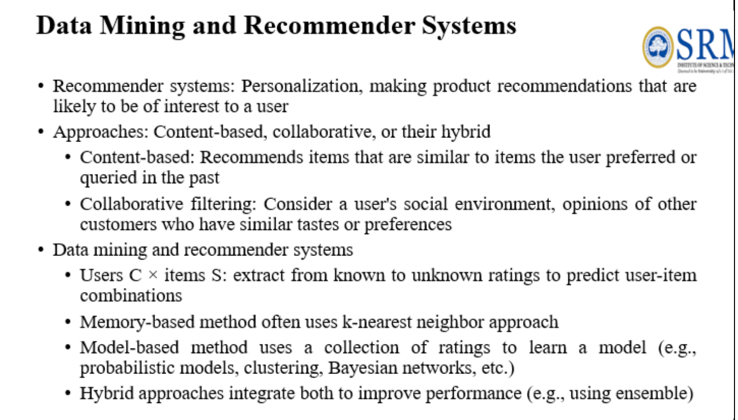
- Data warehouses and data preprocessing
  - Resolving inconsistencies or incompatible data collected in diverse environments and different periods (e.g. eco-system studies)
- Mining complex data types
  - Spatiotemporal, biological, diverse semantics and relationships
- Graph-based and network-based mining
  - Links, relationships, data flow, etc.
- Visualization tools and domain-specific knowledge
- Other issues
  - Data mining in social sciences and social studies: text and social media
  - Data mining in computer science: monitoring systems, software bugs, network intrusion

(iv)Data Mining for Intrusion Detection and Prevention



- Majority of intrusion detection and prevention systems use
  - Signature-based detection: use signatures, attack patterns that are preconfigured and predetermined by domain experts
  - Anomaly-based detection: build profiles (models of normal behavior) and detect those that are substantially deviate from the profiles
- What data mining can help
  - New data mining algorithms for intrusion detection
  - Association, correlation, and discriminative pattern analysis help select and build discriminative classifiers
  - Analysis of stream data: outlier detection, clustering, model shifting
  - Distributed data mining
  - Visualization and querying tools

(v)Data Mining and Recommender Systems

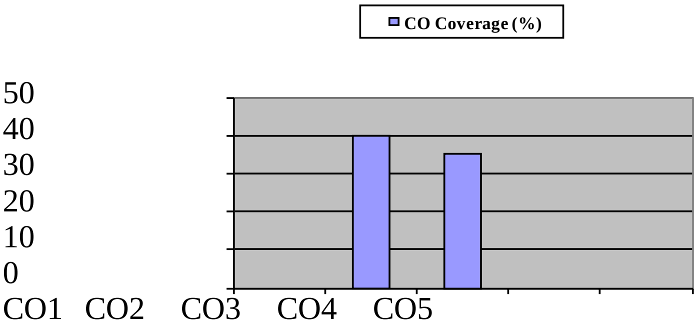


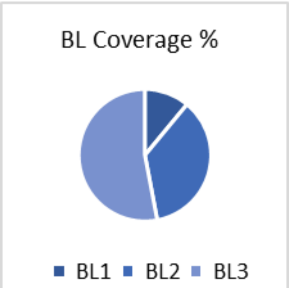
Data Mining and Recommender Systems

- Recommender systems: Personalization, making product recommendations that are likely to be of interest to a user
- Approaches: Content-based, collaborative, or their hybrid
  - Content-based: Recommends items that are similar to items the user preferred or queried in the past
  - Collaborative filtering: Consider a user's social environment, opinions of other customers who have similar tastes or preferences
- Data mining and recommender systems
  - Users C × items S: extract from known to unknown ratings to predict user-item combinations
  - Memory-based method often uses k-nearest neighbor approach
  - Model-based method uses a collection of ratings to learn a model (e.g., probabilistic models, clustering, Bayesian networks, etc.)
  - Hybrid approaches integrate both to improve performance (e.g., using ensemble)

*Performance Indicators are available separately for Computer Science and Engineering in AICTE examination reforms policy.

Course Outcome (CO) and Bloom's level (BL) Coverage in Questions





Approved by the Audit Professor/Course Coordinator

SET B

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)

Test: CLA-T3 | **Date & Session:** 09.11.2023

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: *(to be placed)*

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

Part – A (10 x 1 = 10 Marks)

Answer all questions.
The duration for answering the Part A is 20 minutes
(MCQ Answer sheet will be collected after 20 minutes)

Q. No	Question	Mark	BL	CO	PO	PI Code
1	Which is conclusively produced by HierarchicalClustering? a. Final estimation of cluster centroid **b. Tree showing how nearby things are to each other** c. Assignment of each point to clusters d. Assignment of each centroid to clusters	1	1	4	1	1.7.1
2	Which of the following algorithm is most sensitive to outliers? a. **K-means clustering algorithm** b. K-medians clustering algorithm c. K-modes clustering algorithm d. K-medoids clustering algorithm	1	3	5	4	2.5.2
3	Which clustering technique requires a merging approach? a. Partitional b. **Hierarchical** c. Naive Bayes d. None of the mentioned	1	2	4	4	2.5.2
4	Which method removes sparse clusters as outliers andgroups dense clusters in to larger ones. a. DIANA b. **BIRCH** c. STING	1	2	4	1	1.7.1

	d. DBSCAN					
5	Which is not part of the categories of clustering methods? a. Hierarchical methods b. Density based methods c. Portioning methods d. **Rule-based methods**	1	1	4	1	1.7.1
6	What are the different ways to classify an Intrusion detectionSystem? a. Zone based b. **Host & Network based** c. Network & Zone based d. Level based	1	1	5	2	2.7.1
7	Find the outlier in the given data set below.16, 14, 3,2, 15, 17, 22, 15, 52 a. 22 b. 12 c. **52** d. 3	1	2	5	2	2.7.1
8	The learning algorithms that can deal with both minimal labelled dataset and large unlabeled dataset together is called. a. Supervised b. Unsupervised c. **Semi supervised** d. Reinforcement	1	1	5	2	2.7.1
9	In customer relationship management, we can detect outlier customers using______________. a. Data sparsity b. **Contextual outlier detection** c. Collective outlier detection d. Conventional Outlier Detection	1	1	6	2	2.7.1
10	Internet search engine are tasks related to the area of _______. a. **Information retrieval** b. Information storage c. Information cluster d. Information visualization	1	1	6	8	8.4.2
Part – B **(4 x 5 = 20 Marks)** **Answer any 4 Questions**						
11	**Explain the different types of clustering concepts with appropriate applications.** • Hierarchical Clustering: Forms clusters in a hierarchical manner based on similarity. It can be agglomerative (bottom-up) or divisive (top-down). Used in document clustering, bioinformatics, customer segmentation. • K-Means Clustering: Partitions data into k clusters based on the k cluster centers or centroids. Widely used for customer segmentation, image compression, recommendation systems. • Density-based Clustering: Forms clusters based on density of data points. Can identify arbitrary shapes. Used for spatial data analysis, anomaly detection, biological data analysis. Examples are DBSCAN and OPTICS algorithms. • Distribution-based Clustering: Assumes data follows a statistical distribution and clusters them based on the	5	3	4	1	2.5.2

	distribution parameter(s). Useful when inherent clusters are shaped according to known distributions. Example is the Expectation Maximization (EM) algorithm. • Fuzzy Clustering: Assigns a membership probability to each data point for belonging to each cluster, rather than a hard assignment. Accounts for ambiguities in boundaries. Used in image processing, bioinformatics, customer profiling. Example is Fuzzy C-Means algorithm. • Subspace Clustering: Searches for clusters in different subspaces within a dataset. Useful for high dimensional data where clusters may exist in lower dimensional subspaces. Application is gene expression data analysis. Example is CLIQUE algorithm.					
12	**Give the details of partitioning algorithms and explain with suitable examples.** Partitioning algorithms are also commonly used in data mining for dividing large datasets into smaller chunks to enable more efficient processing and analysis. Here are some key examples: • Horizontal partitioning: Divides a dataset into partitions such that each partition contains a subset of the rows of the original dataset. For example: Input: CustomerID | Name | Age | City 1 | John | 35 | Boston 2 | Sarah | 28 | Austin 3 | David | 42 | Miami 4 | Amy | 24 | Seattle Output (2 partitions): Partition 1: 1 | John | 35 | Boston 3 | David | 42 | Miami Partition 2: 2 | Sarah | 28 | Austin 4 | Amy | 24 | Seattle • Vertical partitioning: Divides a dataset into partitions such that each partition contains a subset of the columns of the original dataset. For example: Input: CustomerID | Name | Age | City | Profession | Income 1 | John | 35 | Boston | Engineer | $80K 2 | Sarah | 28 | Austin | Teacher | $55K Output (2 partitions): Partition 1: CustomerID | Name | Age | City 1 | John | 35 | Boston 2 | Sarah | 28 | Austin Partition 2: CustomerID | Profession | Income 1 | Engineer | $80K 2 | Teacher | $55K • Round robin partitioning: Divides data into partitions by assigning successive tuples to different partitions in a circular order. Often used for parallel processing. • Hash partitioning: Uses a hash function to map each tuple to a partition. Ensures uniform distribution across partitions.	5	3	4	2	2.5.2
13	**Explain in detail about BIRCH technique.** • BIRCH is an incremental clustering algorithm suitable for large datasets. • It constructs a Clustering Feature (CF) tree incrementally to summarize dataset. • Each node represents a subcluster using a CF triple: (n, LS, SS) • n = number of data points in subcluster • LS = linear sum of data points • SS = squared sum of data points • CF vectors capture sufficient stats to incrementally update cluster means and variances. • Data points are inserted recursively down the CF tree. • If entry matches leaf node CF within a threshold, absorb into subcluster. • Else create a new leaf node. • If limit exceeded, apply clustering algorithm and rebuild sub-tree. • BIRCH uses heuristics like diameter threshold to decide emerging subclusters.	5	4	4	2	1.7.1

	• After one scan, apply other clustering algorithms like k-means on leaf nodes to consolidate clusters. Advantages: • Provides clustering of large databases in one scan. • Incrementally absorbs data points using CF vectors. • Minimizes I/O cost. Limitations: • Quality inferior to multiphase algorithms. • Works better for spherical vs arbitrary shapes. Overall, BIRCH provides a good balance between efficiency and quality for large scale clustering.					
14	**Discuss about Outlier detection approaches based on user-labeled examples.** • Supervised outlier detection uses labeled data where normal and anomalous instances are explicitly marked. • Models like SVM, neural networks etc can be trained to distinguish outliers based on predictive features. • Key advantage is performance improvement from availing known ground truth labels. • Drawback is requirement of large labeled dataset which can be expensive. Semi-supervised approaches help address this. • Active learning iteratively selects most informative instances for user labeling. Reduces labeling effort. • Users can provide labels in the form of: • Class labels - Binary normal or outlier classes • Real-valued scores - Reflecting degree of outlierness • Relative constraints - Comparative outlier-ness between instances • Interactive approaches incorporate user feedback to continuously refine the outlier model. • Users can guide model training by validating results, editing rules, marking misclassified instances etc. • Enables user knowledge to augment data patterns for more contextualized outlier detection. • Domain-specific heuristics and rules provided by users can improve model effectiveness.	5	4	5	4	1.7.1
15	**Elaborate LOF in detail with mathematical formulations and working.** LOF is an unsupervised anomaly detection algorithm that identifies outliers by measuring local deviation of density of a data point compared to its neighbors. Steps: 1. Calculate k-nearest neighbors for each point using Euclidean distance. 2. Calculate reachability distance of a point p to a point o as: reachability_dist(p,o) = max(k_dist(o), d(p,o)) where, k_dist(o) = distance between o and its k-th nearest neighbor d(p,o) = distance between points p and o 3. Calculate local reachability density (lrd) of a point p as: lrd(p) = 1 / (sum of reachability_dist(p,o) for all points o in k-neighbors of p) 4. Calculate local outlier factor (LOF) for each point p as: LOF(p) = avg(lrd(o)/lrd(p) for all o in k-neighbors of p) Points with LOF significantly greater than 1 are identified as potential outliers. Intuition: • Reachability distance measures how far p is from o relative to o's typical nearest neighbors distance. • LRD measures density of p based on reachability.	5	4	5	4	2.5.2

	• LOF compares local density of p to its neighbors. • Outliers have lower local density than their neighbors. So LOF captures degree of outlier-ness of each point based on local density deviations. It is an efficient unsupervised technique for anomaly detection.					
Part – C **(2 x 10 = 20 Marks)**						
16	**Explain the K-Medoids algorithm in detail and Apply the K-Medoids algorithm for the following five points (with (x, y) representing locations) into two clusters: P1(2,6), P2(3,8), P3(4,7), P4(6, 2), P5(6, 4), P6(7,3), P7(7,4), P8(8,5), P9(7,6) and P10(3,4).** The K-Medoids algorithm is a clustering technique similar to K-Means, but it chooses actual data points as cluster centers (medoids) instead of taking the mean value. The steps are: 1. Initialize: Randomly select k data points as the initial medoids 2. Associate points: Assign each data point to the closest medoid based on distance 3. Update medoids: For each medoid m, calculate total cost S of swapping m with each non-medoid point o: • Remove m and temporarily replace it with o as the medoid for that cluster • Calculate total distance of all points in the cluster to o • S is the difference between the distance to m and distance to o 4. Select new medoid: Choose the point o with the lowest cost S as the new medoid, replace m 5. Repeat steps 2-4 until medoids stabilize For the given 10 points and k=2 clusters: Initial medoids: P1 and P5 Iteration 1: • P1 medoid: {P2, P3, P4, P5} • P5 medoid: {P1, P6, P7, P8, P9, P10} • Swap P1 with lowest cost P4 as new medoid Iteration 2: • P4 medoid: {P1, P2, P3, P5} • P5 medoid: {P6, P7, P8, P9, P10} • No change in medoids Final clusters: • P4 medoid: {P1, P2, P3, P5}	10	3	4	2	2.5.2

	• P5 medoid: {P6, P7, P8, P9, P10} So the algorithm converged in 2 iterations with P4 and P5 as the final medoid centers.					
(OR)						
17	**Provide a comprehensive explanation of hierarchical clustering methods with appropriate examples.** Hierarchical clustering algorithms create a hierarchy of clusters that can be represented as a dendrogram. There are two main types: Agglomerative clustering (bottom-up approach): • Starts with each data point in its own cluster • Iteratively merges the most similar or closest pairs of clusters • Continues until only a single cluster remains Steps: • Assign each data point to its own cluster • Compute proximity matrix containing distances between all points • Find two closest clusters and merge them • Update proximity matrix to reflect new cluster merges • Repeat steps 3-4 until only one cluster remains Example: In customer segmentation, agglomerative clustering can group customers based on purchasing behavior. Customers with the most similar buying patterns are clustered together iteratively. Divisive clustering (top-down approach): • Starts with all data points in one cluster • Recursively splits clusters into smaller clusters • Stops when each cluster has only one data point Steps: • Place all data points into one cluster • Identify dimensions/attributes to split the cluster • Split cluster into smaller sub-clusters along that dimension • Recursively split sub-clusters further along different dimensions • Stop when each cluster contains only one data point Example: In image segmentation, divisive clustering can recursively divide an image into areas of similar pixel characteristics like color and intensity.	10	2	4	2	2.5.2

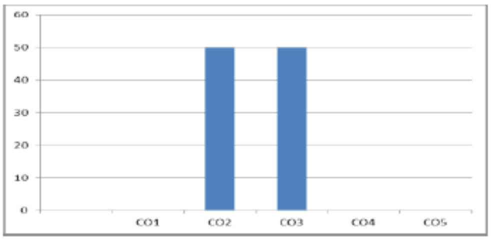
	Key advantages of hierarchical clustering include dendrogram visualization and no need to pre-specify number of clusters. Limitations include inability to correct erroneous merges or splits and lack of scalability. Overall, it provides a multilevel hierarchy well-suited for data exploration.					
18	**Elaborate comprehensively on Statistical and Proximity based methods used for detecting outliers.** Statistical approaches assume that the objects in a data set are generated by a stochastic process (a generative model) Idea: learn a generative model fitting the given data set, and then identify the objects in low probability regions of the model as outliers Methods are divided into two categories: parametric vs. non-parametric **Parametric method** Assumes that the normal data is generated by a parametric distribution with parameter θ The probability density function of the parametric distribution f(x, θ) gives the probability that object x is generated by the distribution The smaller this value, the more likely x is an outlier **Non-parametric method** Not assume an a-priori statistical model and determine the model from the input data ν Not completely parameter free but consider the number and nature of the parameters are flexible and not fixed in advance Examples: histogram and kernel density estimation	10	4	5	4	2.7.1
(OR)						
19	**How can the utilization of data mining algorithms for intrusion detection and prevention be optimized to enhance cybersecurity on a broader scale, considering the evolving nature of cyber threats and the complex network environments of today?** • Employ ensemble models that combine multiple algorithms like SVM, neural networks, etc. This provides more robust and adaptive detection capabilities compared to any single technique. • Leverage both supervised learning for detecting known attack patterns and unsupervised learning to identify anomalies indicative of novel threats. This balances recognition of known and unknown attacks. • Implement incremental learning approaches to continuously	10	4	5	4	1.7.1

	update the models with new data without full retraining. This allows responding faster to evolving threats. • Utilize distributed data mining architectures for scalable processing of huge volumes of security events across large networks. This enables capturing signals from global traffic. • Incorporate contextual information like threat intelligence feeds, asset criticality, vulnerability data etc. to improve threat awareness and prioritization for the specific environment. • Promote sharing of attack data patterns between organizations following privacy guidelines. This helps transfer learning to improve collective detection of emerging threats. • Benchmark systems against adversarial simulation environments to harden effectiveness against evasion attempts by attackers. • Develop standardized frameworks that allow interoperability and integration of different tools and algorithms. This facilitates broader adoption. • Promote research and development of advanced techniques like deep learning and adversarial learning for IDS/IPS.					

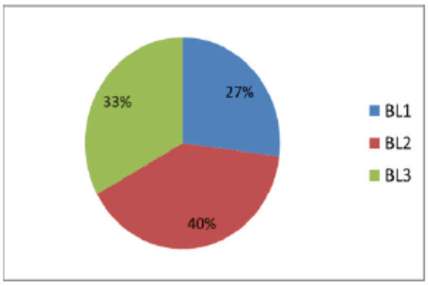
***Program Indicators are available separately for Computer Science and Engineering in AICTE examination reforms policy.**

Course Outcome (CO) and Bloom's level (BL) Coverage in Questions

CO Coverage in %



BL Coverage in %



Approved by the Audit Professor/Course Coordinator

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

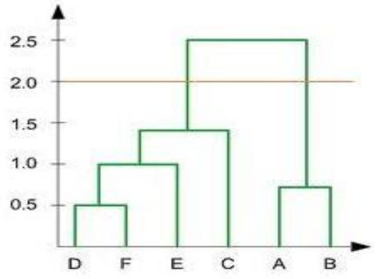
Course Articulation Matrix:

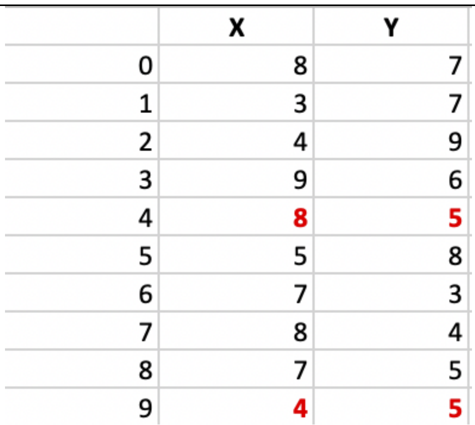
S. No	Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
1	CO1	H	L	M									
2	CO2		H	M	L								
3	CO3		H	M	L								
4	CO4	H	M	L									
5	CO5	H	M	L									
6	CO6	L			M	H							

Part – A (10 x 1 = 10 Marks)

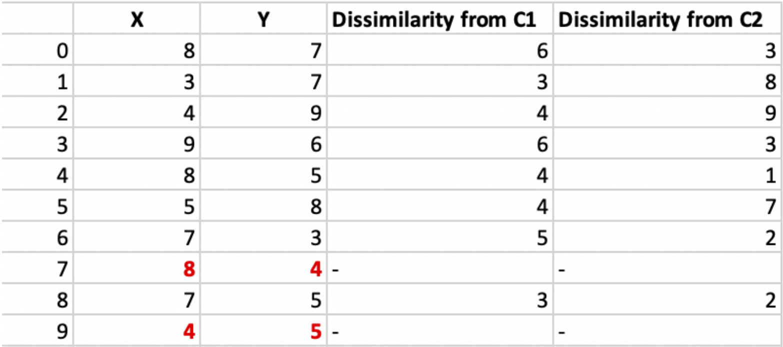
Answer all Questions

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



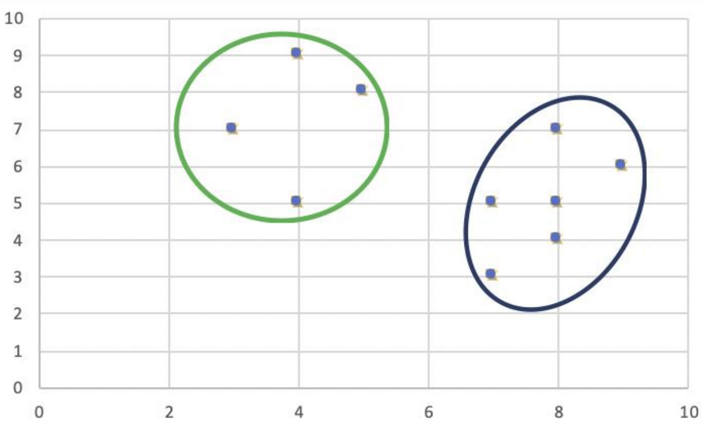
	X	Y
0	8	7
1	3	7
2	4	9
3	9	6
4	8	5
5	5	8
6	7	3
7	8	4
8	7	5
9	4	5



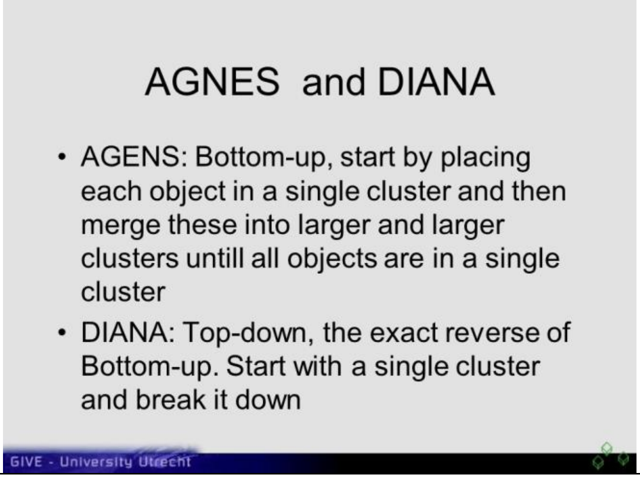
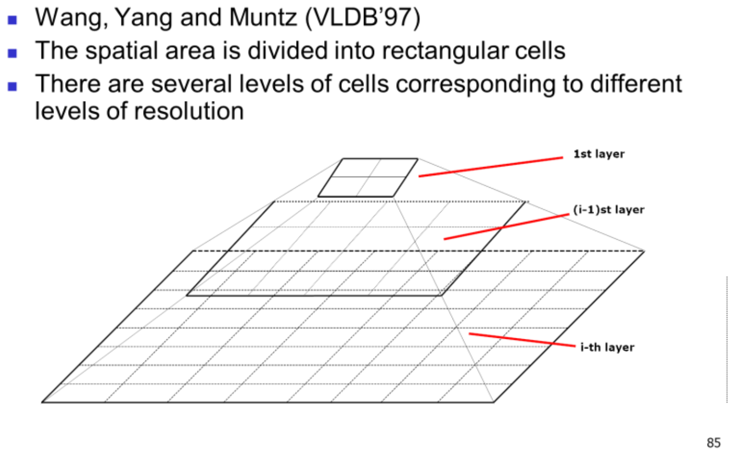
	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	3
1	3	7	3	8
2	4	9	4	9
3	9	6	6	3
4	8	5	4	1
5	5	8	4	7
6	7	3	5	2
7	8	4	-	-
8	7	5	3	2
9	4	5	-	-

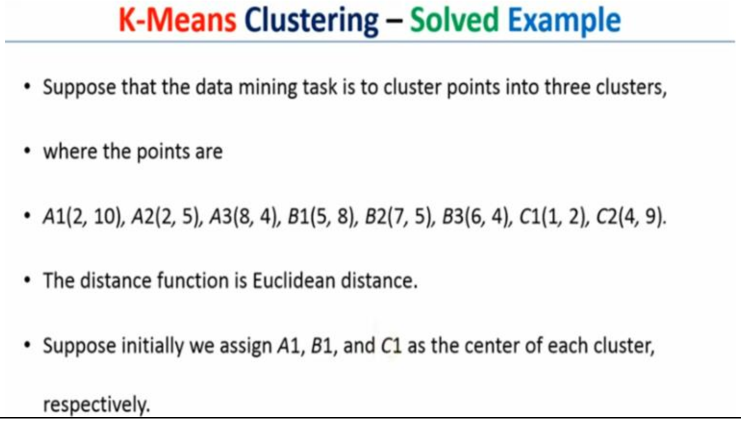
Each point is assigned to that cluster whose dissimilarity is less. So, points 1, 2, and 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2. The New cost = (3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22 Swap Cost = New Cost – Previous Cost = 22 – 20 and **2 >0** As the swap cost is not less than zero, we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

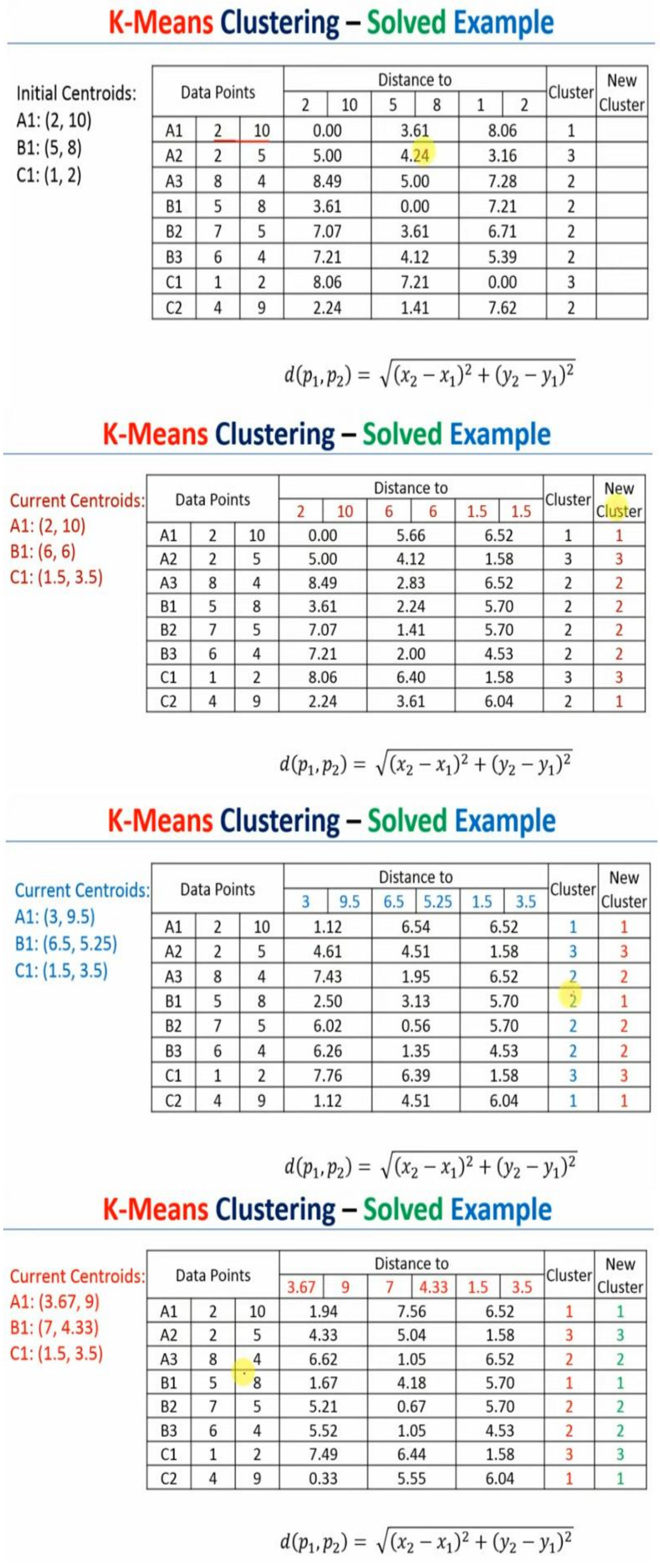
complexity is .



12	Differentiate between AGNES and DIANA algorithms.

	 AGNES and DIANA • AGENS: Bottom-up, start by placing each object in a single cluster and then merge these into larger and larger clusters untill all objects are in a single cluster • DIANA: Top-down, the exact reverse of Bottom-up. Start with a single cluster and break it down GIVE - University Utrecht					
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 • *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.	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

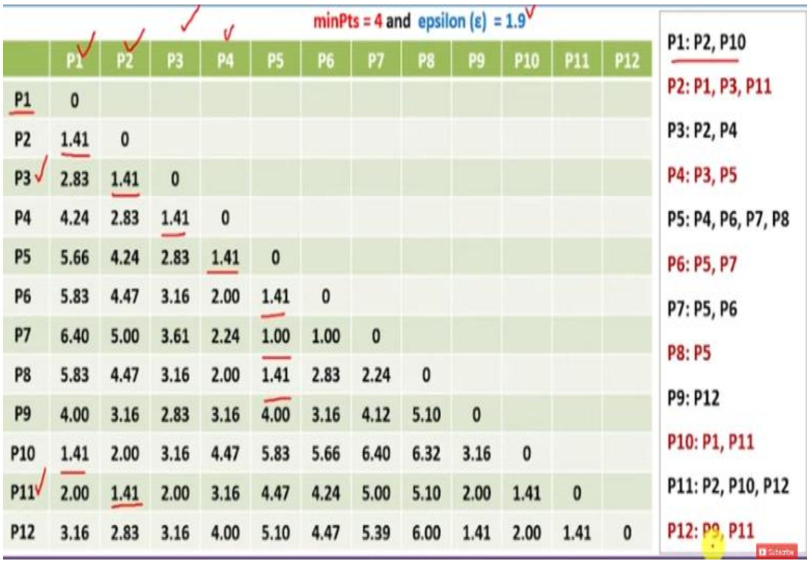
Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

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

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

<table>
<tr>
<td>12</td>
<td>

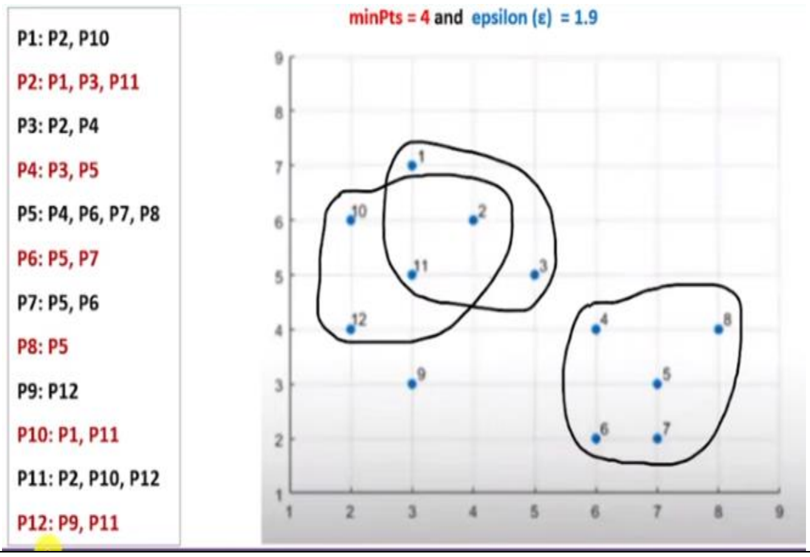
What is a DBSCAN? Apply DBSCAN algorithm to the given data points to create the cluster with minpts = 4, epsilon = 1.9 and p1(3,7),p2(4,6),p3(5,5),p4(6,4),p5(7,3),p6(6,2),p7(7,2),p8(8,4), p9(3,3),p10(2,6),p11(3,5),p12(2,4).



minPts = 4 and epsilon (ε) = 1.9

<table>
<tr><th></th><th>P1</th><th>P2</th><th>P3</th><th>P4</th><th>P5</th><th>P6</th><th>P7</th><th>P8</th><th>P9</th><th>P10</th><th>P11</th><th>P12</th></tr>
<tr><td>P1</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>P2</td><td>1.41</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>P3 ✓</td><td>2.83</td><td>1.41</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>P4</td><td>4.24</td><td>2.83</td><td>1.41</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>P5</td><td>5.66</td><td>4.24</td><td>2.83</td><td>1.41</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>P6</td><td>5.83</td><td>4.47</td><td>3.16</td><td>2.00</td><td>1.41</td><td>0</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>P7</td><td>6.40</td><td>5.00</td><td>3.61</td><td>2.24</td><td>1.00</td><td>1.00</td><td>0</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>P8</td><td>5.83</td><td>4.47</td><td>3.16</td><td>2.00</td><td>1.41</td><td>2.83</td><td>2.24</td><td>0</td><td></td><td></td><td></td><td></td></tr>
<tr><td>P9</td><td>4.00</td><td>3.16</td><td>2.83</td><td>3.16</td><td>4.00</td><td>3.16</td><td>4.12</td><td>5.10</td><td>0</td><td></td><td></td><td></td></tr>
<tr><td>P10</td><td>1.41</td><td>2.00</td><td>3.16</td><td>4.47</td><td>5.83</td><td>5.66</td><td>6.40</td><td>6.32</td><td>3.16</td><td>0</td><td></td><td></td></tr>
<tr><td>P11 ✓</td><td>2.00</td><td>1.41</td><td>2.00</td><td>3.16</td><td>4.47</td><td>4.24</td><td>5.00</td><td>5.10</td><td>2.00</td><td>1.41</td><td>0</td><td></td></tr>
<tr><td>P12</td><td>3.16</td><td>2.83</td><td>3.16</td><td>4.00</td><td>5.10</td><td>4.47</td><td>5.39</td><td>6.00</td><td>1.41</td><td>2.00</td><td>1.41</td><td>0</td></tr>
</table>

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



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

minPts = 4 and epsilon (ε) = 1.9

</td>
<td>10</td>
<td>L2</td>
<td>4</td>
<td>2</td>
<td>2.5.2</td>
</tr>
<tr>
<td>13</td>
<td>

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.

</td>
<td>10</td>
<td>L2</td>
<td>6</td>
<td>4</td>
<td>2.7.1</td>
</tr>
</table>

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

Reg. No.															

SRM Institute of Science and Technology
College of Engineering and Technology
School of Computing
EPARTMENT OF COMPUTING TECHNOLOGIES

Set D

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu
Academic Year: 2023-2024 (ODD)

Test: CLAT-3 **Date:** 09.11.2023
Course Code & Title: 18CSE355T - Data Mining and Analytics **Duration:** 2 Periods
Year & Sem: III Year & 05th Semester **Max. Marks:** 50 Marks

Part – A (10 x 1 = 10 Marks) Answer all questions.		
Q. No	**Question**	**Marks**
1	The process of grouping a set of physical or abstract objects into classes of similar objects is called as ________. a) Prediction b) Association c) Correlation d) Clustering **Answer:** d) Clustering	1
2	Let p1=(3,2) and p2=(4,1) represent two objects, what will be the Euclidean distance? a) 2.449 b) 2.236 **c) 1.414** d) 1.732 **Answer:** c) 1.414	1
3	What is the primary assumption made by the K-means clustering algorithm? a) Clusters have a spherical shape b) Clusters have similar densities c) Clusters have similar sizes d) Clusters are linearly separable **Answer:** a) Clusters have a spherical shape	1
4	_____method removes sparse clusters as outliers and groups dense clusters in to larger ones. a) DIANA b) BIRCH c) STING d) DBSCAN **Answer:** b) BIRCH	1
5	Hierarchical agglomerative clustering is typically visualized using ________. a) Dendrogram b) Binary trees c) Block diagram d) Graph **Answer:** a) Dendrogram	1
6	What does the term 'outlier' mean? a) A score that is left out of the analysis because of missing data b) The arithmetic mean c) A type of variable that cannot be quantified d) An extreme value at either end of a distribution **Answer:** d) An extreme value at either end of a distribution	1
7	In Univariate outlier detection uses the maximum likelihood to estimate which of the following parameter(s)? i) Mean ii) Median iii) Standard deviation iv) User-specified parameter. a) (i) and (ii) b) (i) and (iii) c) (iii) and (iv) d) (ii), (iii) (iv)	1

	Answer: b) (i) and (iii)	
8	The learning algorithms that can deal with both minimal labelled dataset and large unlabeled dataset together is called _____. a) Supervised b) Unsupervised c) Semi supervised d) Reinforcement **Answer:** c) Semi supervised	1
9	Which of the following data mining application is used to allow the retailer to understand the purchase behaviour of a buyer? a) Manufacturing Engineering b) Fraud Detection c) Corporate Surveillance d) Market Basket Analysis **Answer:** d) Market Basket Analysis	1
10	Which of the following is an example of a Content-based Recommendation System? i) Recommending movies to a user based on the movies rated by that user in the past ii) Recommending the most popular movies iii) Randomly recommending some selections from a list of movies a) Only i b) Only ii and iii c) Only i and iii d) Only ii **Answer:** a) Only i	1
Part – B **(4 x 5 = 20 Marks)** **Answer any 4 Questions**		
11	Explain CLIQUE Technique with example. **Answer:** • CLIQUE can be considered as both density- based and grid-based • It partitions each dimension into the samenumber of equal length interval • It partitions an m-dimensional data space intonon-overlapping rectangular units • A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter • A cluster is a maximal set of connected denseunits within a subspace **The Major Steps** • Partition the data space and find the number of points that lie inside each cell of the partition. • Identify the subspaces that contain clustersusing the Apriori principle • Identify clusters • Determine dense units in allsubspaces of interests • Determine connected dense units inall subspaces of interests. • Generate minimal description for the clusters • Determine maximal regions that covera cluster of connected dense units for each cluster • Determination of minimal cover foreach cluster Strength • automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces • *insensitive* to the order of records in input and does not presume some	**5**

	canonical data distribution • scales *linearly* with the size of input and has good scalability as the number of dimensions in the data increases Weakness • The accuracy of the clustering result may be degraded at the expense of simplicity of the method	
12	Discuss the Challenges of clustering Data mining. **Key** • Scalability • Ability to deal with different types of attributes • Constraint-based clustering • Interpretability and usability Above topics with briefing about each.	**5**
13	How to Measure the Clustering Quality? Explain. **Answer:** • Dissimilarity/Similarity metric Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$ • The definitions of distance functions are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables • Weights should be associated with different variables based on applications and data semantics • Quality of clustering: • There is usually a separate "quality" function that measures the "goodness" of a cluster. • It is hard to define "similar enough" or "good enough" • The answer is typically highly subjective	**5**
14	Outline the Challenges of Outlier Detection. **Key Points :** ● Modeling normal objects and outliers properly ○ Hard to enumerate all possible normalbehaviors in an application ○ The border between normal and outlier objects is often a gray area ● Application-specific outlier detection ○ Choice of distance measure among objects and the model of relationship among objects are often application- dependent ○ E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations ● Handling noise in outlier detection ○ Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection ● Understandability ○ Understand why these are outliers: Justification of the detection ○ Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism	**5**
15	**Discuss about any two Major Statistical Data Mining Methods** **Key:** **Explanation of any two of the following methods: 2 * 2.5= 5 marks** • Regression • Generalized Linear Model • Analysis of Variance	**5**

- Mixed-Effect Models
- Factor Analysis
- Discriminant Analysis
- Survival Analysis

Part – C
(2 x 10 = 20 Marks)
Answer All the questions

16 Briefly Explain the K-Means algorithm and Apply the K-means algorithm for the following five points (with (x, y) representing locations) into two clusters: A1(3, 10), A2(7, 5), A3(10, 4), A4(5, 9), A5(8, 5).
Initial cluster centers are: A1(3, 10), and A4(5, 9)
Use K-Means Algorithm to find the two cluster centers after two iterations.
Note: The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- P(a, b) = |x2 – x1| + |y2 – y1| **10**

Key:

K-Means Clustering Algorithm – 5 marks
Problem solution – 5 marks

Answer:

K-Means Clustering Algorithm

K-Means Clustering Algorithm involves the following steps-

Step-01:

Choose the number of clusters K.

Step-02:

Randomly select any K data points as cluster centers.
Select cluster centers in such a way that they are as farther as possible from each other.

Step-03:

Calculate the distance between each data point and each cluster center.
The distance may be calculated by using Euclidean distance formula (or by using given distance function in variants of the algorithm).

Step-04:

Assign each data point to some cluster.
A data point is assigned to that cluster whose center is nearest to that data point.

Step-05:

Re-compute the center of newly formed clusters.
The center of a cluster is computed by taking mean of all the data points contained in that cluster.

Step-06:

Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met:
Center of newly formed clusters do not change.
Data points remain present in the same cluster Maximum number of iterations are reached.

Problem:

Answer

A1(3, 10), A2(7, 5), A3(10, 4), A4(5, 9), A5(8, 5).

Iteration 1

Initial cluster centers are: A1(3, 10), and A4(5, 9)

Given Points	**Distance from**	**Distance from**	**Point belongs**

	center of Cluster-01	center of Cluster-02	to Cluster
A1(3, 10)	0	3	C1
A2(7, 5)	9	6	C2
A3(10, 4)	13	10	C2
A4(5, 9)	3	0	C2
A5(8,5)	10	7	C2

Iteration 2

New Cluster centers are: (3, 10), and (30/4=7.5,23/4=5.75)

Given Points	Distance from center of Cluster-01	Distance from center of Cluster-02	Point belongs to Cluster
A1(3, 10)	0	8.75	C1
A2(7, 5)	9	1.25	C2
A3(10, 4)	13	4.25	C2
A4(5, 9)	3	5.75	C1
A5(8,5)	10	1.25	C2

Iteration 3

New Cluster centers are: (4, 9.5), and (25/3=8.33,14/3=4.67)

Given Points	Distance from center of Cluster-01	Distance from center of Cluster-02	Point belongs to Cluster
A1(3, 10)	1.5	10.66	C1
A2(7, 5)	7.5	3.66	C2
A3(10, 4)	11.5	2.34	C2
A4(5, 9)	1.5	7.66	C1
A5(8,5)	8.5	0.66	C2

Between Iterations 2 and 3 no change.
So, Clusters are:
C1 - A1(3, 10) and A4(5, 9),
C2 - A2(7, 5), A3(10, 4) and A5(8,5).

17 Outline how to compute dissimilarity between objects described by the following types of variables with example. **10**

i. Interval-Scaled Variable, ii. Binary Variable, iii. Categorical Variable

Answer:

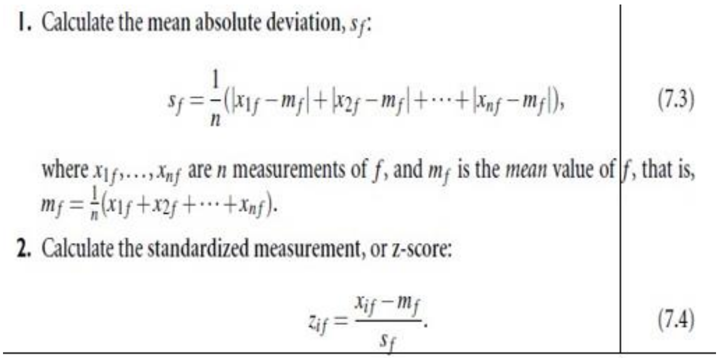
Key Points:

Interval-Scaled Variable

- Interval-scaled variables are continuous measurements of a roughly linear scale. Eg:weight and height
- The measurement unit used can affect theclustering analysis.
- For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead toa very different clustering structure.
- To avoid dependence on the choice ofmeasurement units, the data should bestandardized.
- Standardizing measurements attempts to giveall variables an equal

weight.

Data Normalization



1. Calculate the mean absolute deviation, s_f:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|), \qquad (7.3)$$

where $x_{1f}, \ldots, x_{nf}$ are n measurements of f, and m_f is the *mean* value of f, that is, $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$.

2. Calculate the standardized measurement, or z-score:

$$z_{if} = \frac{x_{if} - m_f}{s_f}. \qquad (7.4)$$

<u>Explanation of Distance measures.</u>

Binary Variable

- The total number of variables is *p*, where *p* = *q*+*r*+*s*+*t*.
- A binary variable is symmetric if both of its states are equally valuable and carry the sameweight.
- There is no preference on which outcomeshould be coded as 0 or 1.
- Dissimilarity that is based on symmetric binary variables is called symmetric binarydissimilarity.
- A binary variable is asymmetric if the outcomes of the states are not equally important, such as the positive and negativeoutcomes of a disease test.
- A binary variable contains two possible outcomes: 1 (positive/present) or 0 (negative/absent). If there is no preference forwhich outcome should be coded as 0 and which as 1, the binary variable is called *symmetric*.
- For example, the binary variable "is evergreen?" for a plant has the possible states"loses leaves in winter" and "does not lose leaves in winter." Both are equally valuable and carry the same weight when a proximity measure is computed.
- If the outcomes of a binary variable are notequally important, the binary variable is called *asymmetric*.
- An example of such a variable is the presenceor absence of a relatively rare attribute, such as "is color-blind" for a human being.
- While you say that two people who are color-blind have something in common, you cannotsay that people who are not color-blind have something in common.

Categorical Variable

- A **categorical variable** is a generalization ofthe binary variable in that it can take on morethan two states.
- For example, *map color* is a categorical variable that may have, say, five states: *red*, *yellow*, *green*, *pink*, and *blue*.
- Let the number of **states** of a categorical variable be **M**. The states can be denoted byletters, symbols, or a set of integers, such as1, 2, : : : , M.
- The dissimilarity between two objects i and jcan be computed based on the ratio of mismatches equation

$$d(i,j) = \frac{p-m}{p},$$

- m - where m is the number of matches (i.e., the number of variables for which i and j are in the same state), and p is the total number of variables.

object identifier	*test-1* (categorical)	*test-2* (ordinal)	*test-3* (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

A sample data table containing variables of mixed type.

Suppose that we have the sample data of Table except that only the object-identifier and the variable(or attribute) test-1 are available, where test-1 is categorical. Let's compute the dissimilarity matrix.

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

Since here we have one categorical variable, *test-1*, we set $p = 1$ in Equation (7.12) so that $d(i,j)$ evaluates to 0 if objects i and j match, and 1 if the objects differ. Thus, we get

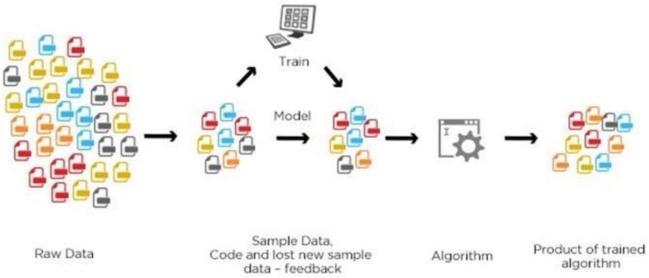
$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

18 Discuss in detail the semi-supervised methods for detecting the outliers. **10**

<u>Answer:</u>

Semi Supervised Algorithm for outliers

- Situation: In many applications, the number of labeled data is often small: Labels could beon outliers only, normal objects only, or both
- Semi-supervised outlier detection: Regardedas applications of semi-supervised learning
- If some labeled normal objects are available
    - Use the labeled examples and the proximate unlabeled objects to train amodel for normal objects
    - Those not fitting the model of normalobjects are detected as outliers
- If only some labeled outliers are available, asmall number of labeled outliers many not cover the possible outliers well
    - To improve the quality of outlier detection, one can get help from models for normal objects learnedfrom unsupervised methods



19 Outline how data mining algorithms can be used for recommender systems and financial data analysis. **10**

<u>Key:</u>

Recommender system – 5 marks

Financial data analysis – 5 marks

Answer:

Data Mining for Recommender systems

- Personalization, making product recommendations that are likely to be of interest to a user
- Approaches: Content-based, collaborative, or their hybrid
- Content-based: Recommends items that are similar to items the user preferred or queried in the past
- Collaborative filtering: Consider a user's social environment, opinions of other customers who have similar tastes or preferences
- Data mining and recommender systems
- Users C × items S: extract from known to unknown ratings to predict user-item combinations
- Memory-based method often uses k-nearest neighbor approach
- Model-based method uses a collection of ratings to learn a model (e.g., probabilistic models, clustering, Bayesian networks, etc.)

Hybrid approaches integrate both to improve performance (e.g., using ensemble)

Data Mining for Financial Data Analysis

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of data warehouses for multidimensional data analysis and data mining
  - View the debt and revenue changes by month, by region, by sector, and by other factors
  - Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
  - feature selection and attribute relevance ranking
  - Loan payment performance
  - Consumer credit rating
- Classification and clustering of customers for targeted marketing
  - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
  - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
  - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)