UNIT 15 CLUSTERING

Structure

15.1 INTRODUCTION

Clustering or cluster analysis is a method for dividing a group of data objects into subgroups based on a single observation. Here, each cluster is a subset of the data, where objects with resemblance or similar properties are grouped. Also, they are similar to each other in one cluster but differ from those objects which are in different clusters. In simple terms, the effort of dividing a population into multiple groups so that data points of one group can easily be compared with data points of another group. Thus, to separate the groups with identical features and assign them into clusters. This process is performed by machines, not by humans and is known as unsupervised learning because clustering is a form of learning by observation. Clustering is often confused with classification in data analysis, where separation of data happens based on class labels, while in clustering partitioning of large data sets occurs in groups occurs based on similarity.

Let's understand this with an example. Suppose, you are the head of a rental store and wish to understand the preferences of your customers to scale up your business. Is it possible for you to look at the details of each customer and devise a unique business strategy for each one of them? Not. But, what you can do is to cluster all of your customers into say 10 groups based on their purchasing habits and use a separate strategy for customers in each of these 10 groups. And this is what we call **clustering.**

"Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data."

Data clustering is a prominent technique in data analysis and is applied in various research areas including mining of data, data statistics, area of machine learning for any kind of analysis. It is also applied in the world of financial services, health information systems web mining, financial sectors and many more. Cluster analysis is the most recent area of research in data analysis due

to the massive volumes of data produced in databases. An example of clustering is outlier detection where credit card fraud and criminal activities are monitored. Clustering can be used in image recognition to find clusters or patterns in image or text recognition systems. Clustering has a lot of uses in web search as well. Due to the enormous quantity of online pages, a keyword search may frequently produce a huge number of hits (i.e., pages relevant to the search). Thus, clustering is a very promising machine learning process and is proved to be one of the most pragmatic data mining tools. In this unit, you will learn about the various types of clustering techniques and algorithms.

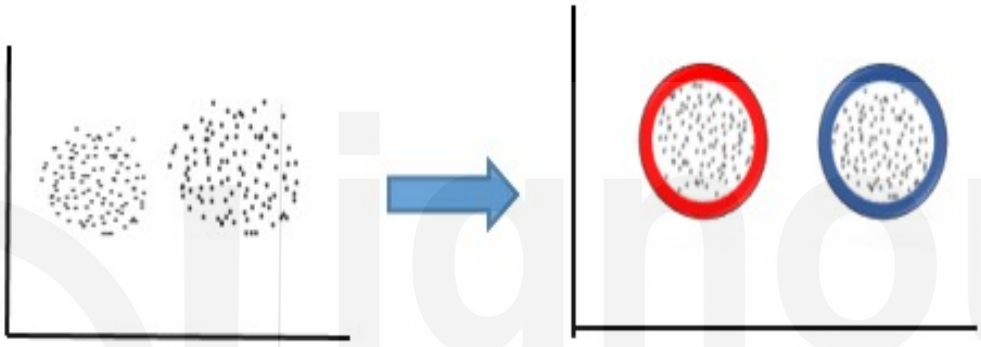


Fig 1. Clustering

Real World Example:

A bank can potentially have millions of customers. Would it make sense to look at the details of each customer separately and then decide? Certainly not! It is a manual process and will take a huge amount of time.

So, what can the bank do? Clustering comes to the rescue in these situations where the banks can group the customers based on their income, as shown:

Applications of Clustering in Real-World Scenarios

Clustering is a widely used technique in the industry. It is being used in almost every domain, ranging from banking to recommendation engines, document clustering to image segmentation.

- Customer Segmentation

Customer segmentation is one of the most common applications of clustering. And it isn't just limited to banking. This strategy is across functions, including telecom, e-commerce, sports, advertising, sales, etc.

- Document Clustering

Document Clustering is another common application of clustering. Let's assume that you have multiple documents, and you need to cluster similar documents together. Clustering helps us group these documents such that similar documents are in the same clusters.

- Image Segmentation

We can also use clustering to perform image segmentation. Here, we try to club similar pixels in the image together. We can apply clustering to create clusters having similar pixels in the same group.

15.2 Types of Clustering

There are many clustering algorithms in the literature. Traditionally, documents are grouped based on how similar they are to other documents. Similarity-based algorithms define a function for computing document similarity and use it as the basis for assigning documents to clusters. Each cluster should have data that are comparable to one another but different from those in other clusters. Clustering algorithms fall into different categories based on the underlying methodology of the algorithm (agglomerative or partition), the structure of the final solution (flat or hierarchical), or the density based All the above-mentioned clustering types are discussed in detail in the rest of this chapter.

Check Your Progress - 1

Qn1. What do you understand by the term "Clustering"?

Qn2. Where is Clustering used in present day scenario?

15.3 Partition Based Clustering

Partition algorithm is one of the most applied clustering algorithms. It has been widely used in many applications due to its simple structure and easy implementation as compared to other clustering algorithms. This clustering method classifies the information into multiple groups based on the characteristics and similarities of the data. In the partitioning method when database(D) contains multiple(N) objects then the partitioning approach divides the data into user-specified(K) partitions, each of which represents a cluster and a specific region. That is, the data is divided into K groups. That is, it divides the data into K groups or partitions in such a manner that each group should have at least one object from the existing data. To put it another way, it splits the data items into non-overlapping subsets (clusters) so that each data object fits perfectly into one of them.

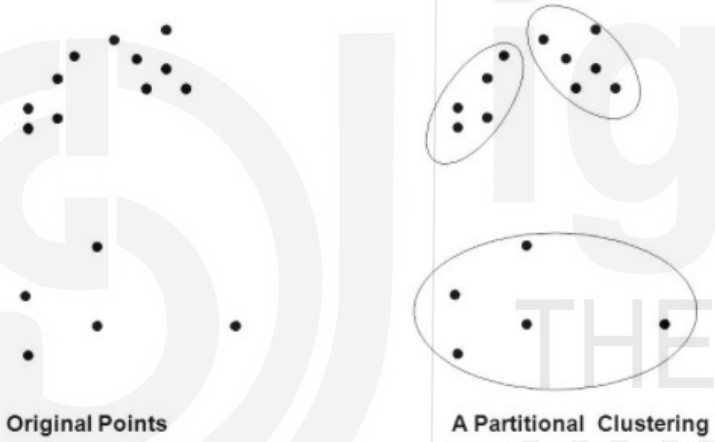


Fig 2. Partition based clustering

Partitioning approaches require a set of starting seeds (or clusters), which are then enhanced iteratively by transferring objects from one group to another to improve partitioning.Objects in the same cluster are "near" or related to one other, whereas objects in other clusters are "far apart" or significantly distinct, according to the most prevalent approach to good partitioning.

Many algorithmsthat come under partitioning method some of the popular ones are K-Mean, PAM (K-Mediods), and CLARA algorithm (Clustering Large Applications) etc.

CHECK YOUR PROGRESS

Qn1. Briefly explain how Partition Based Clustering works.

Qn2. Name three Partition Based Clustering methods.

15.4 Hierarchical BasedClustering

Hierarchical Clustering analysis is an algorithm that groups the data points with similar properties and these groups are termed "clusters". As a result of hierarchical clustering, we get a set of clusters, and these clusters are always different from each other. Clustering of this data into clusters is classified as:

- Agglomerative Clustering (involving decomposition of cluster using bottom-up strategy)
- Divisive Clustering (involving decomposition of cluster using top-down strategy)

Hierarchical clustering helps in creating clusters in the proper order (or hierarchy).

Example:

The most common everyday example we see is how we order our files and folders in our computer by proper hierarchy.

As mentioned, Hierarchical clustering is classified into two types i.e., Agglomerative clustering (AGNES) and Divisive Clustering (DIANA)

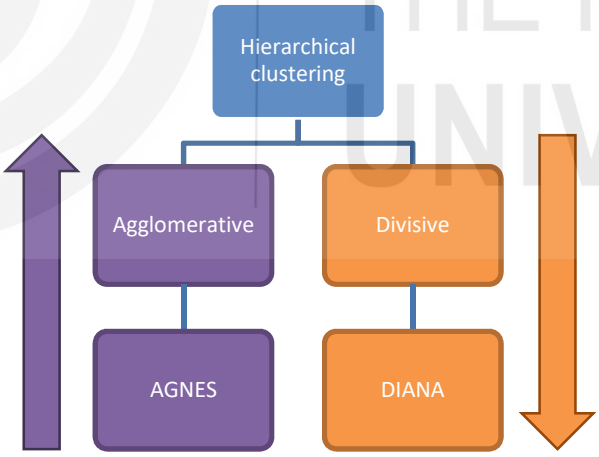


Fig3. Showing hierarchical clustering AGNES vs DIANA

Hierarchical clustering Technique in terms of space and time complexity:

- **Space complexity:** When the number of data points is large, the space required for the Hierarchical Clustering Technique is large since the similarity matrix must be stored in RAM. The space complexity is measured by the order of the square of n.

  Space complexity = $O(n^2)$ where n is the number of data points.

- **Time complexity:** The time complexity is also very high because we have to execute n iterations and update and restore the similarity matrix in each iteration. The order of the cube of n is the time complexity.

  Time complexity = $O(n^3)$ where n is the number of data points.

Limitations of HierarchicalClustering Technique:

1. Hierarchical clustering does not have a mathematical goal.
2. All the methodologies applied for calculating the similarity index between clusters does not apply fully in every situation, each technique has its own merits and demerits.
3. Due to high space and time complexity. This clustering algorithm is not applicable for huge data.

Check Your Progress - 2

Qn1. Differentiate between Hierarchical Clustering and Partition Based Clustering.

Qn2. What are sub-types of Hierarchical Clustering and what is the difference between them?

Qn3. What is the space and time complexity of Hierarchical clustering?

Qn4. List some of the limitations of Hierarchical clustering

15.5 Density Based Clustering Technique

Clusters are formed in the Density Depending Clustering Technique based on the density of the data points represented in the data space.Those locations that become dense as a result of the large amount of data points that reside there are termed as clusters.

How it works:

1. It starts with a random unvisited starting data point. All points within a distance 'Epsilon' – Ɛ classify as neighborhood points.

2. We need a minimum number of points within the neighborhood to start the clustering process. In this scenario, the current point of data turns into the first point in the cluster. Else, the point is regardedas 'Noise.' In either case, the current point becomes a visited point.

3. All points within the distance(Ɛ) become part of the same cluster. This process is repeated for all the newly added data points in the cluster group.

4. Continue with the process until you visit and label each point within the 'Ɛ' neighborhood of the cluster.

5. On completion of the process, start again with a new unvisited point thereby leading to the discovery of more clusters or noise. At the end of the process, you ensure that you mark each point as either cluster or noise.

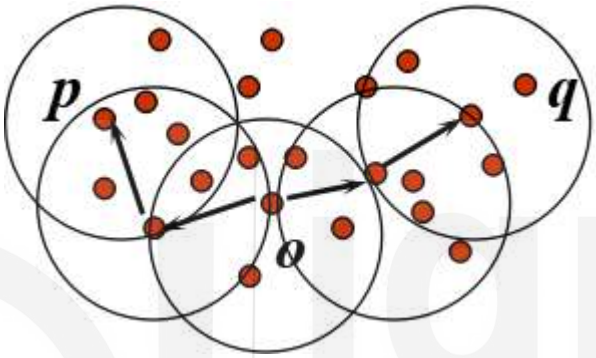


Fig 4. Showing Density connected points

Check Your Progress - 3

Qn1. What do you understand by the term "Density Based Clustering"?

Qn2. How is "Density Based Clustering" performed?

15.6 Clustering Algorithms

a) K-Means

Among clustering algorithms, is an algorithm that tries to minimize the distance of the points in a cluster with their *centroid* – the k-means clustering technique.

K-means is a centroid-based algorithm. It can also be called asa distance-based technique where distances between points are calculated to allocate a point to a cluster. Each cluster in K-Means is paired with a centroid.

The K-Means algorithm's main goal is to reduce the sum of distances between points and their corresponding cluster centroid.

Real World Example:

Let's have a look at an example. Assume you went to a bookstore to purchase some books. There are several types of books that can be found there. One thing you'll notice is that the books are sorted into groups based on their category. All of the literature books will be kept in one location, while science books will be organized by type. The K-Means algorithm's main goal is to reduce the sum of distances between points and their corresponding cluster centroid.

Now we will understand this with the help of these figures.

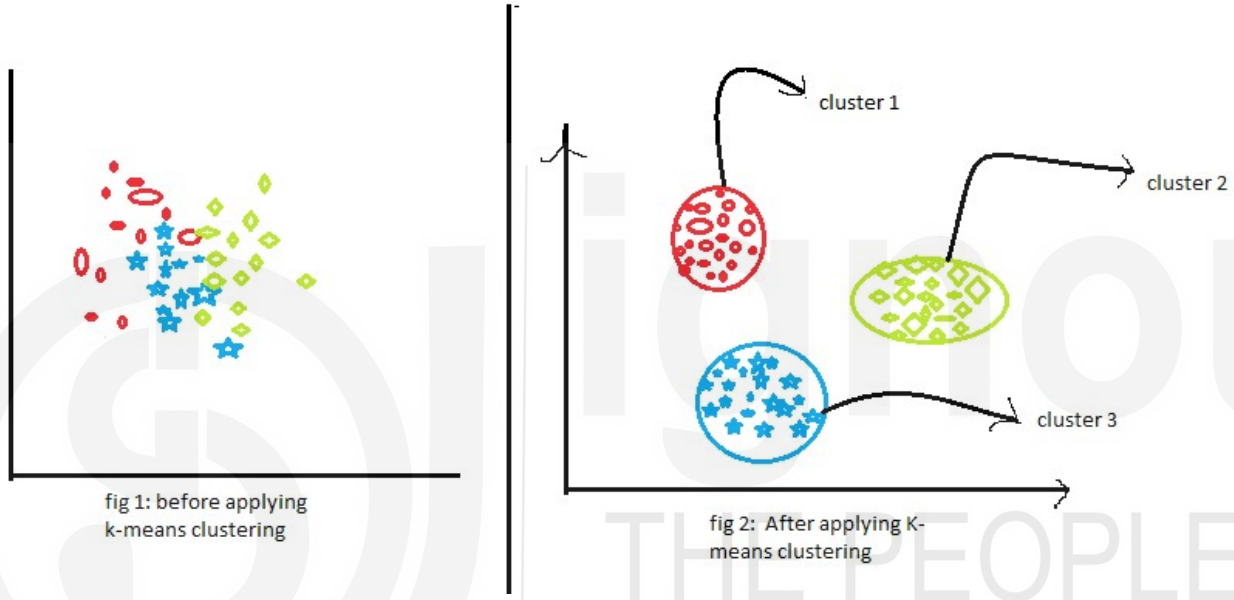


Fig 5. Before and after K-means clustering

In the figure 5 data is presented into two stages. The first figure shows the data in raw stage which is not clustered by k-means. Here all types of data are clubbed together thus becomes impossible to differentiate them into their original category.

The second figure shows three clusters of three different colors red, green, and blue. These clusters are formed after applying k-means clustering. The second figure shows data into three different categories which are called clusters.

Working of K-means clustering algorithm .

k-means clustering technique is an immense clustering algorithm to group similar types of data in groups which are so called clusters. It is the simplest and commonly used technique in machine learning extensively used for data analysis. It can easily locate the similarity points between the different data items and can group them into the clusters. The working of K-means clustering algorithm is shown in the following three steps. Let’s see what these three steps are.

1. Selection of k values.
2. Initialization of the centroids.
3. Selection of the cluster and finding the average.

Let us understand the above steps with the help of the Fig6:

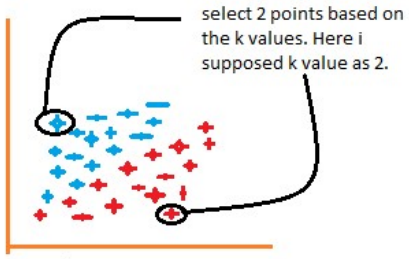


Figure 1

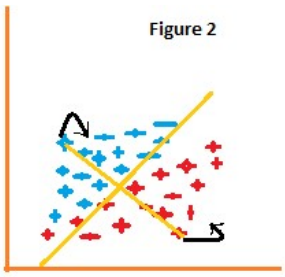
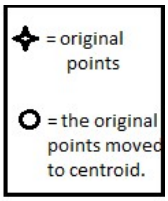
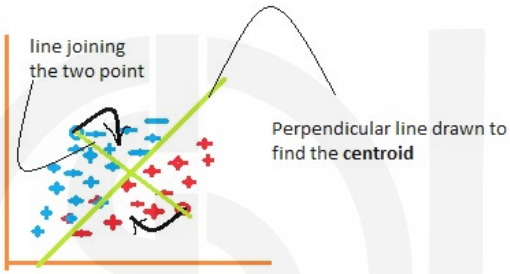


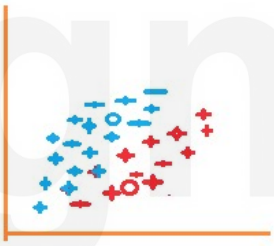
Figure 2

F2: Find the average of all the blue points and red points and move the selected points to **centroid.**





F3: Some of the red points changed to blue points, that means they belong to the group blue now. Again the repeat the same process.



F4: The same process has been applied here. This process will be continued until we get the **two complete different cluster.**

Below is a Practice Problem Based on K-Means Clustering Algorithm:

Problem-01: The following eight points (with (x, y) denoting places) should be grouped into three clusters:

A1(2, 11), A2(2, 15), A3(8, 5), A4(6, 8), A5(7, 9), A6(6, 3), A7(1, 4), A8(4, 8)

The first cluster centers will be: A1(2, 11), A4(6, 8) and A7(1,4).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$$P(a, b) = |x2 – x1| + |y2 – y1|$$

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Solution: We follow the above discussed K-Means Clustering Algorithm-

Iteration-01:

- We measure the distance between each location and the center of each of the three clusters.
- The specified distance function is used to determine the distance.

The calculation of distance between point A1(2, 11) and each of the center of the three clusters-

Calculating Distance Between A1(2, 11) and C1(2, 11)-

P(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |11 – 11|

= 0

Calculating Distance Between A1(2, 10) and C2(6, 8)-

P (A1, C2)

= |x2 – x1| + |y2 – y1|

= |6 – 2| + |8 – 10|

= 4 + 2

= 6

Calculating Distance Between A1(2, 10) and C3(1, 4)-

P(A1, C3)

= |x2 – x1| + |y2 – y1|

= |1 – 2| + |4 – 10|

= 1 + 6

= 7

Problem-02 (Self-Test)

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Assume that the first seeds (cluster centers) are A1, A4, and A7. Only use the k-means method for one epoch. Show: a) The new clusters (i.e., the cases belonging to each cluster) b) The new clusters' centers at the end of this epoch c) Draw a 10 by 10 space with all 8 points and illustrate the clusters and new centroids after the first epoch. d) How many more iterations will it take to reach convergence? For each period, draw the result.

b) Agglomerative clustering

In this case of clustering, the hierarchical decomposition is done with the help of bottom-up strategy where it starts by creating atomic (small) clusters by adding one data object at a time and then merges them together to form a big cluster at the end, where this cluster meets all the termination conditions. This procedure is iterative until all the data points are brought under one single big cluster.

Basic algorithm of agglomerative clustering

1. Determine the proximity matrix.

2. Assume that each data point belongs to a cluster.

3. Do it again.

4. Combine the two groups that are the closest together.

5. Make changes to the proximity matrix.

6. Continue until just one cluster remains.

AGNES (Agglomerative Nesting) is a type of agglomerative clustering that combines the data objects into a cluster based on similarity. The result of this algorithm is a tree-based structure called Dendrogram. Here it uses the distance metrics to decide which data points should be combined with which cluster. Basically, it constructs a distance matrix and checks for the pair of clusters with the smallest distance and combines then. The figure 7. given below shows dendrogram.

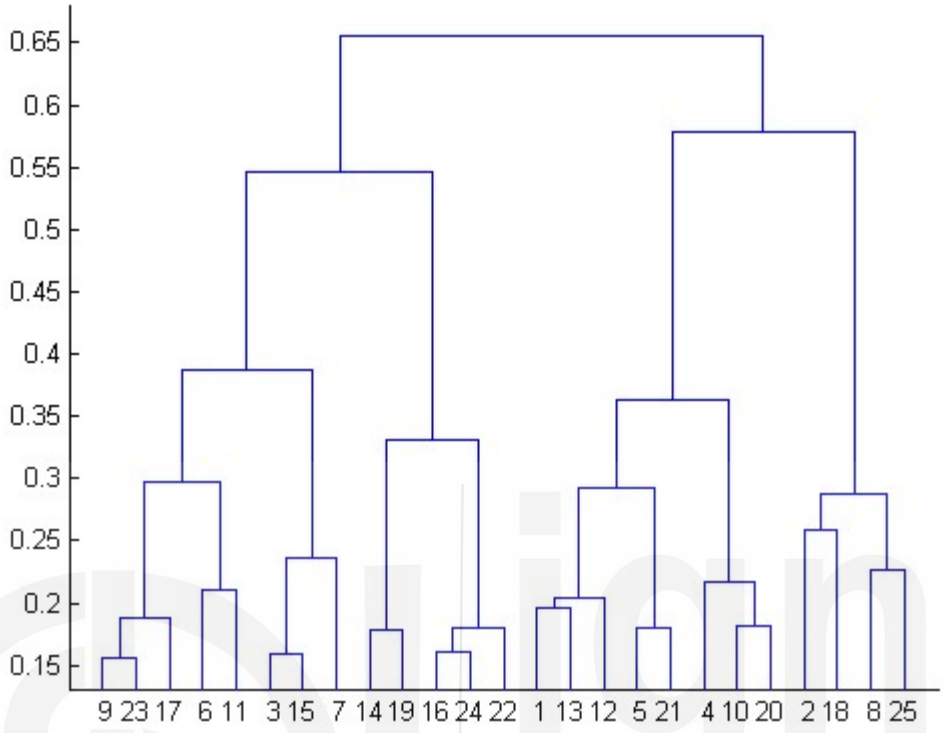


Fig 7. Agglomerative clustering

We begin at the bottom with 25 data points, each of which is assigned to a different cluster. Then two nearest clusters are selected to merge till we get only one cluster at the topmostposition. The distance between two clusters in the data space is represented by the height in the dendrogram at which two clusters are merged. Based on how the distance between each cluster is measured, we can have 3 different methods

- **Single linkage:** Where the shortest distance between the two points in each cluster is defined as the distance between the clusters.
- **Complete linkage:** In this case, we will consider the longest distance between each cluster's points as the distance between the clusters.

- **Average linkage:** In this situation, we'll take the average of each point in one cluster compared to every other point in the other.

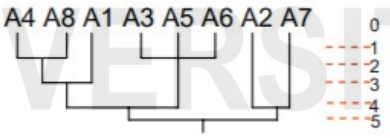
AGNES has few limitations one of this is it has a time complexity of at least **O(n2)**; hence it doesn't do well in scaling, and one other major drawback is that whatever has been done can never be undone, i.e. If we incorrectly group any cluster in an earlier stage of the algorithm, then we will not be able to change the outcome/modify it. But this algorithm has a bright side since there are many smaller clusters are formed; it can be helpful in the process of discovery. It produces an ordering of objects that is very helpful in visualization.

Problem-03: Hierarchical clustering (to be done at your own time, not in class) Use single-link, complete-link, average-link agglomerative clustering as well as medoid and centroid to cluster the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). The distance matrix is the same as the one in Exercise 1. Show the dendrograms.

Solution:

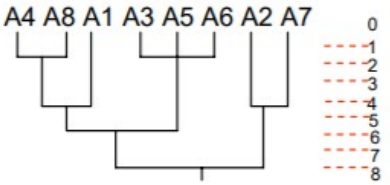
Single Link:

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	2	{A1, A3, A4, A5, A6, A8}, {A2, A7}
5	1	{A1, A3, A4, A5, A6, A8, A2, A7}



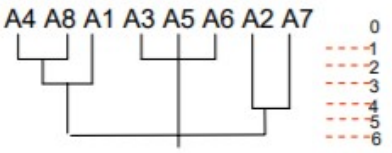
Complete Link

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	2	{A4, A8, A1, A3, A5, A6}, {A2, A7}
7	2	{A4, A8, A1, A3, A5, A6}, {A2, A7}
8	1	{A4, A8, A1, A3, A5, A6, A2, A7}



Average Link

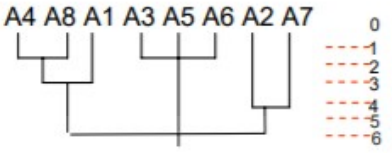
d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	1	{A4, A8, A1, A3, A5, A6, A2, A7}



Average distance from {A3, A5, A6} to {A1, A4, A8} is 5.53 and is 5.75 to {A2, A7}

Centroid

D	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	1	{A4, A8, A1, A3, A5, A6, A2, A7}

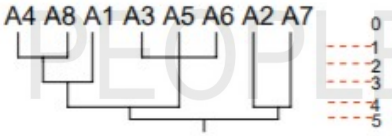


Centroid of {A4, A8} is B=(4.5, 8.5) and centroid of {A3, A5, A6} is C=(7, 4.33)
distance(A1, B) = 2.91 Centroid of {A1, A4, A8} is D=(3.66, 9) and of {A2, A7} is E=(1.5, 3.5)
distance(D,C)= 5.74 distance(D,E)= 5.90

Medoid

This is not deterministic. It can be different depending upon which medoid in a cluster we chose.

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	2	{A1, A3, A4, A5, A6, A8}, {A2, A7}
5	1	{A1, A3, A4, A5, A6, A8, A2, A7}



c) Divisive Clustering (DIANA)

Diana basically stands for Divisive Analysis; this is another type of hierarchical clustering where basically it works on the principle of top-down approach (inverse of **AGNES**) where the algorithm begins by forming a big cluster, and it recursively divides the most dissimilar cluster into two, and it goes on until we're all the similar data points belong in their respective clusters. These divisive algorithms result in highly accurate hierarchies than the agglomerative approach, but they are computationally expensive.

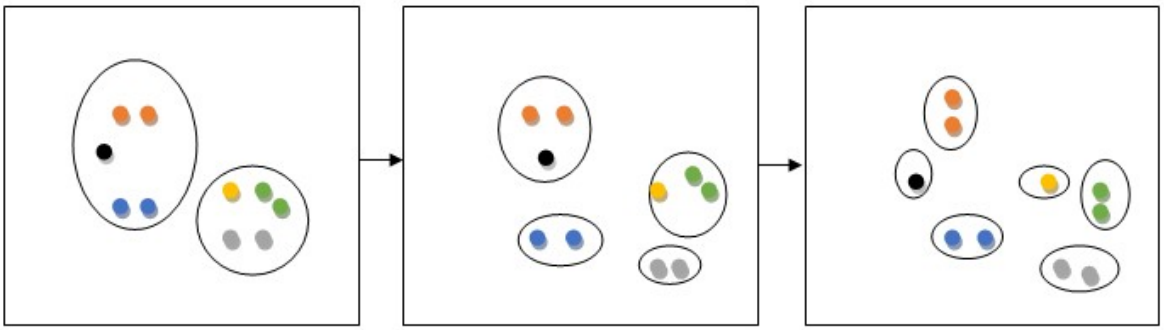


Fig 8. Divisive clustering step by step process

d) Density-Based Spatial Scan (DBSCAN)

The distance between objects is used to cluster things in most partitioning methods. Such algorithms can only locate spherical-shaped clusters and have trouble finding clusters of other forms. DBSCAN can form clusters in different shapes; this type of algorithm is most suitable when the dataset contains noise or outliers. Also, it depends on a density-based concept of cluster: A cluster is defined as the most densely connected set of points.

Other clustering algorithms based on the concept of density have been developed. Their basic concept is to keep creating a cluster till the density which consists of data points in the "neighbourhood" surpasses a certain threshold.

For instance, every data point present in a given cluster should meet the requirement of having minimum number of points in the neighborhood of a given radius. This type of method is extremely useful for detecting outliers or to filter out noise. This is also useful for finding clusters of arbitrary shape.

The best part in density-based methods is that they can divide a set of objects into multiple exclusive clusters, or a hierarchy of clusters. Density-based techniques often only evaluate exclusive clusters and ignore fuzzy clusters. Furthermore, density-based clustering algorithms can be extended from entire space to subspace.

DBSCAN uses noise to find clusters of any shape in spatial databases.

In density-based clustering we partition points into dense regions separated by not-so-dense regions. Characterization of points is done in following manner:

If a point has more than a given number of points (MinPts) within Eps, it is considered a core point.

▪ These points belong in a dense region and are at the interior of a cluster.

Within Eps, a border point has less points than MinPts, but it is close to a core point.

Any point that isn't a core point or a boundary point is referred to as a noise point.

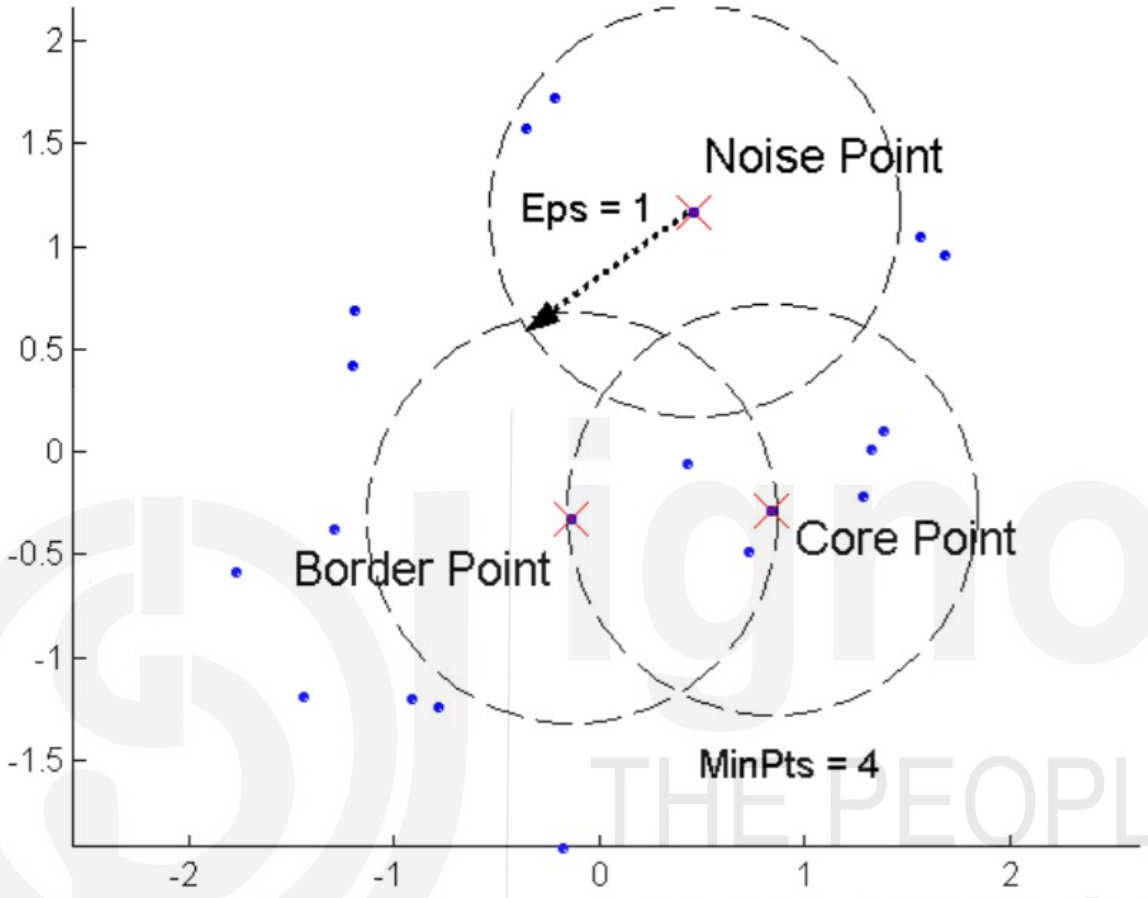
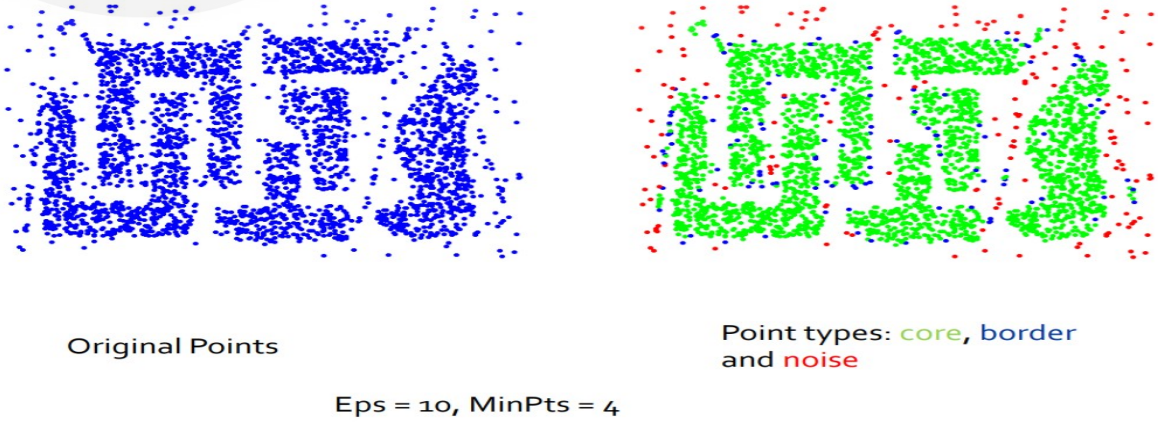


Fig 9. Core point, Border point and Noise point in DBScan



DBSCAN Algorithm

1. Label points as core, border and noise
2. Eliminate noise points for every core point p that has not been assigned to a cluster
3. Create a new cluster with the point p and all the points that are density-connected to p.
4. Assign border points to the cluster of the closest core point.

e) Fuzzy Clustering

Fuzzy clustering is the most used clustering algorithm present in real world and is a sort of clustering in where each data point is assigned to multiple clusters. The algorithm suggests that the data points can belongs to more than one cluster unlike hard clustering where data points can actually belong to only one cluster.

Illustrative Example
A Guava can either be Yellow and Green (hard clustering), but a Guava can also be Yellow and Green (this is fuzzy clustering). Here, the Guava can be Yellow to a certain degree as well as Green to a certain degree. Instead of the Guava belonging to Green [green = 1] and not Yellow [Yellow = 0], the Guava can belong to Green [green = 0.5] and Yellow [Yellow = 0.5]. These values are normalized between 0 and 1; however, they do not represent probabilities, so the two values do not need to add up to 1.

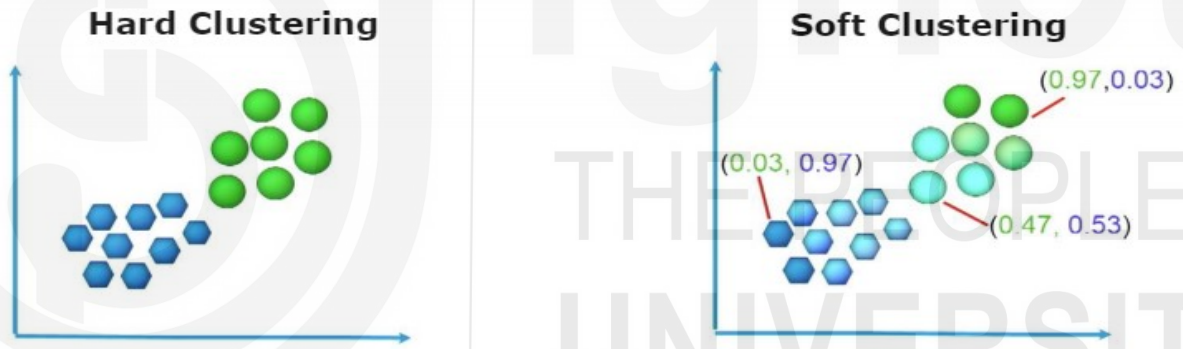


Fig 10. Understanding the algorithm

The Fuzzy Clustering algorithm attempts to group a finite set of n items.
$R = R1, \ldots, rnR = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}$ into a set of c fuzzy clusters with respect to some given criterion.

when a finite set of data is given, the algorithm returns a list of clusters centers
$F = f1, \ldots, ffF = \{\boldsymbol{f}_1, \ldots, f_f\}$And a partition matrix

$W = w_{i,j} \in [0,1], i = 1, \ldots, n, j = 1, \ldots, f$,where each element, w_{ij} tells the degree to which element,$\boldsymbol{r}_i$ belongs to cluster $\boldsymbol{f}_j$

The FCM aims to minimize an objective function:

$$\underset{C}{argmin} \sum_{i=1}^{n} \sum_{j=1}^{f} w_{ij}^{m} \parallel \boldsymbol{r}_i - \boldsymbol{f}_j \parallel^2,$$

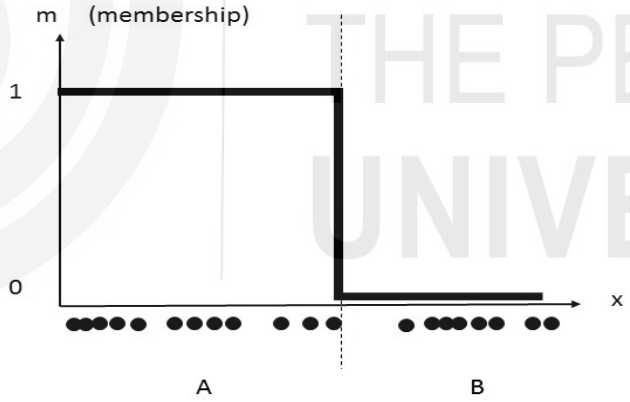
where:

$$w_{ij} = \frac{1}{\sum_{k=1}^{f} \left(\frac{\| \boldsymbol{r}_i - \boldsymbol{f}_j \|}{\| \boldsymbol{r}_i - \boldsymbol{f}_k \|} \right)^{\frac{2}{m-1}}}.$$

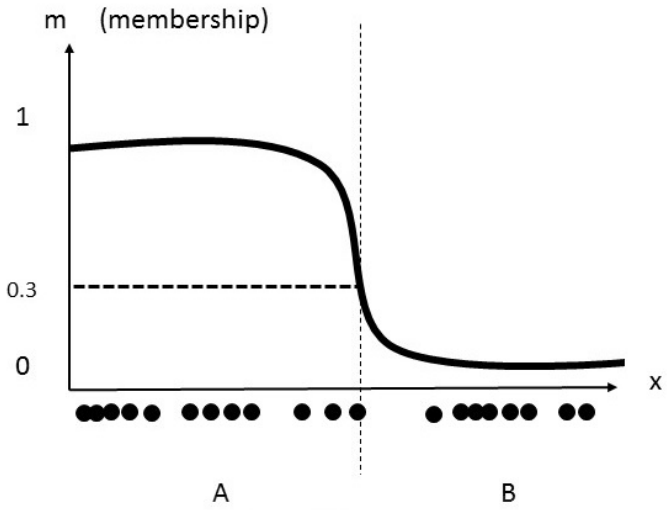
Illustrative Example

To better understand this principle, a classic example of mono-dimensional data is given below on an x axis.

This data set can be conventionally grouped into two clusters, and this is done by selecting a threshold on the x-axis.
The resulting two clusters are labelled 'A' and 'B', as seen in the image below. Each point belonging to the data set would therefore have a membership coefficient of either 0 or 1. This membership coefficient of each corresponding data point is represented by the inclusion of the y-axis.



In fuzzy clustering, each data point can have membership to multiple clusters. By relaxing the definition of membership coefficients from strictly 0 or 1, these values can range from inclusive value from 0 to 1. The following image shows the data set from the previous clustering, but now fuzzy c-means clustering has been applied. First, a new threshold value defining two clusters is generated. Next, new membership coefficients for each data point are generated based on clusters centroids, as well as distance from each cluster centroid.



As we can see, the middle data point belongs to cluster A and cluster B. The value (point) of 0.3 is this data point's membership coefficient for cluster A

Check Your Progress - 4

Qn1. Briefly explain 5 types of clustering.

Qn2. What is Agglomerative Clustering? What are 3 methods that can be used for Agglomerative Clustering?

Qn3. Briefly explain fuzzy clustering and provide a Real-World Example for the same.

Qn4. Differentiate Between AGNES and DIANA

Qn5. What is DBSCAN?

Qn6. Explain how K-Means Algorithm works.

15.7 SUMMARY

The technique of separating a set of data objects into subgroups based on some observation is known as clustering or cluster analysis. Each subset is referred to as a 'cluster,' with objects that are related to one another but different from those in other clusters. In simple terms, the process of separating a population or set of data points into several groups so that data points from the same group can be compared to data points from different groups. Clustering is often confused with classification in data analysis, where separation of data happens on the basis of class labels, while in clustering partitioning of large data sets occurs in groups occurs on the basis of similarity.

Clustering algorithms fall into different categories based on the underlying methodology of the algorithm, the structure of the final solution, or the density based.

In Density Based Clustering Technique, the clusters are created based upon the density of the data points which are represented in the data space.

K-Means: Among clustering algorithms, is an algorithm that tries to minimize the distance of the points in a cluster with their centroid the k-means clustering technique. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

AGNES is a type of agglomerative clustering that combines the data objects into a cluster based on similarity.

DBSCAN uses noise to find clusters of any shape in spatial databases.

Fuzzy Clustering is the most widely used clustering technique in practice, and it is a sort of clustering in which each data point belongs to many clusters.

15.8 SOLUTIONS TO CHECK YOUR PROGRESS

Check Your Progress - 1

For detailed answers refer to Section 15.2.

Answer 1.

Clustering is the process of grouping data into groups (also known as clusters) based on patterns found in the data. To put it another way, the work of dividing a population or Data points into groups so that data points in the same group are more comparable to data points in other groups.

Answer 2.

Clustering is a widely used technique in the industry. It is being used in almost every domain, ranging from banking to recommendation engines, document clustering to image segmentation.

- Customer Segmentation
- Document Clustering
- Image Segmentation

Check Your Progress - 2

For detailed answers refer to Section 15.3

Answer 1.

Partitioning approaches start with a set of beginning seeds (or clusters), which are then improved iteratively by shifting objects from one group to another. According to the general criterion of good partitioning, objects in the same cluster are "close" or connected to one another, whereas objects in other clusters are "far away" or notably distinct.

Answer 2.

Popular partitions based on clustering are:

- K-Mean,
- PAM (K-Medoids),
- CLARA algorithm (Clustering Large Applications)

Check Your Progress - 3

For detailed answers refer to Section 15.4.

Answer 1:

Differences in assumptions, runtime, input, and output are all factors. Partitional clustering is, on average, faster than hierarchical clustering. Stronger assumptions are required for partitional clustering. Hierarchical clustering, on the other hand, simply requires a similarity metric. There are no input parameters required for hierarchical clustering, but partitional clustering techniques require a number of clusters to begin. Clusters are divided more meaningfully and subjectively using hierarchical clustering. Partitional clustering, on the other hand, produces k clusters.

Answer 2:

The major distinction between Hierarchical and Partitional Clustering is that each cluster begins as a singleton or individual cluster. The nearest clusters are joined with each iteration. This technique is repeated until only one cluster for Hierarchical clustering remains.

Partitional clustering, on the other hand, needs the analyst to designate K clusters before executing the algorithm, and objects that are closest to the clusters are clustered together. The spacing between the cluster's changes with each repetition. This process continues until the centroid of each cluster no longer moves, or until the halting requirement is reached.

Answer 3:

Hierarchical clustering Technique in terms of space and time complexity:

- **Space complexity:** When the number of data points is large, the space required for the Hierarchical Clustering Technique is large since the similarity matrix must be stored in RAM. The space complexity is measured by the order of the square of n.

  Space complexity = $O(n^2)$ where n is the number of data points.

- **Time complexity:** The time complexity is also very high because we have to execute n iterations and update and restore the similarity matrix in each iteration. The order of the cube of n is the time complexity.

  Time complexity = $O(n^3)$ where n is the number of data points.

Answer 4:

Limitations ofHierarchical Clustering Technique:

1. Hierarchical clustering does not have a mathematical goal.

2. Every method for calculating cluster similarity has its own set of drawbacks.
3. Hierarchical clustering has a high spatial and temporal complexity. As a result, when we have a lot of data, we can't apply this clustering approach.

Check Your Progress - 4

For detailed answers refer to Section 15.5.

Answer 1.

The term "Clustering" is a technique in which the clusters are created based upon the density of the data points which are represented in the data space. The regions that become dense due to the huge number of data points residing in that region are considered as clusters.

Answer 2.

Refer to Page 8

Check Your Progress - 5

For detailed answers refer to Section 15.6.

Answer 1.

a) K-Means

Among clustering algorithms, is an algorithm that tries to minimize the distance of the points in a cluster with their centroid – the k-means clustering technique.
K-means is a centroid-based or distance-based technique in which the distances between points are calculated to allocate a point to a cluster. Each cluster in K-Means is paired with a centroid.

b) AGNES (Agglomerative Nesting)

AGNES is a type of agglomerative clustering that combines the data objects into a cluster based on similarity. The result of this algorithm is a tree-based structure called Dendrogram. Here it uses the distance metrics to decide which data points should be combined with which cluster. Basically, it constructs a distance matrix and checks for the pair of clusters with the smallest distance and combines then.

c) Divisive Clustering (DIANA)

Diana basically stands for Divisive Analysis; this is another type of hierarchical clustering where basically it works on the principle of top-down approach (inverse of AGNES) where the algorithm begins by forming a big cluster, and it recursively divides the most dissimilar cluster into two, and it goes on

d) Density-Based Spatial Scan (DBSCAN)

The distance between objects is used to cluster things in most partitioning methods. Only spherical-shaped clusters can be found using these approaches, and clusters of any shape are difficult to find. DBSCAN may create clusters of various shapes; this technique is best suited to datasets with noise or outliers.

Fuzzy Clustering

The most widely used clustering technique in the real world is fuzzy clustering, which is a sort of clustering in which each data point belongs to many clusters. The algorithm suggests that the data points can belongs to more than one cluster unlike hard clustering where data points can actually belong to only one cluster.

Answer 2.

a) **AGNES** (AGglomerativeNESting)

AGNES is a type of agglomerative clustering that combines the data objects into a cluster based on similarity. The result of this algorithm is a tree-based structure called Dendrogram. Here it uses the distance metrics to decide which data points should be combined with which cluster. Basically, it constructs a distance matrix and checks for the pair of clusters with the smallest distance and combines then.

Based on how the distance between each cluster is measured, we can have 3 different methods

- **Single linkage:** Where the shortest distance between the two points in each cluster is defined as the distance between the clusters.
- **Complete linkage:** In this case, we will consider the longest distance between each cluster's points as the distance between the clusters.
- **Average linkage:** In this situation, we'll take the average of each point in one cluster compared to every other point in the other.

Answer 3.

The most widely used clustering technique in the real world is fuzzy clustering, which is a type of clustering in which each object belongs to many clusters**.** The algorithm suggests that the data points can actually belongs to more than one cluster unlike hard clustering where data points can actually belong to only one cluster.

Real-World Example:

A Guava can either be Yellow and Green (hard clustering), but a Guava can also be Yellow and Green (this is fuzzy clustering). Here, the Guava can be Yellow to a certain degree as well as Green to a certain degree. Instead of the Guava belonging to Green [green = 1] and not Yellow [Yellow = 0], the Guava can belong to Green [green = 0.5] and Yellow [Yellow = 0.5]. These values are normalized between 0 and 1; however, they do not represent probabilities, so the two values do not need to add up to 1.

Answer 4:

DIANA is like the reverse of **AGNES**. It begins with the root, in which all observations are included in a single cluster. At each step of the algorithm, the current cluster is split into two clusters that are considered most heterogeneous. The process is iterated until all observations are in their own cluster.

Answer 5:

DBSCAN can form clusters in different shapes; this type of algorithm is most suitable when the dataset contains noise or outliers.

Also, it depends on a density-based concept of cluster: A cluster is defined as the most densely connected set of points.

Answer 6:

Refer Page 10

MULTIPLE CHOICE QUESTIONS

Q1. The objective of clustering is to-

A. Sort the data points into categories.
B. To classify the objects into different classes
C. To predict the values of input data points and generate output.
D. All of the above

Solution: (A)

Q2. Clustering is a-

A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning
D. None

Solution:(B)

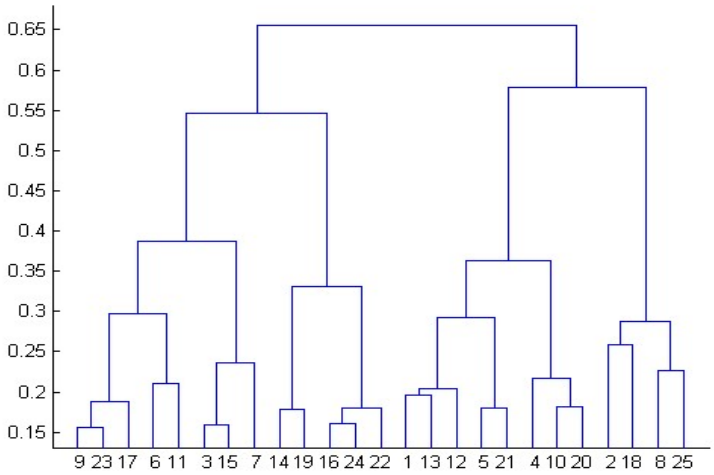
Q3. Which of the following clustering algorithm is most sensitive to outliers?

A. K-means
B. K-modes
C. K-medians
D. K-medoids

Solution: (A)

Explanation: The K-Means clustering approach, which employs the mean of cluster data points to locate the cluster center, is the most sensitive to outliers of all the options.

Q4. You saw the dendrogram below after doing K-Means Clustering analysis on a dataset. From the dendrogram, which of the following conclusions can be drawn?



A. There were 32 data points in clustering analysis

B. The data points analyzed has best number of clusters is 4

C. The proximity function used here is Average-link clustering

D. The above interpretation of dendrogram is not possible for K-Means clustering analysis

Solution: (D)

Explanation: Dendrogram is not possible for K-Means clustering analysis. However, one can create a cluster gram based on K-Means clustering analysis.

.

Q5. What are the two types of Hierarchical Clustering?

A. Top-Down Clustering (Divisive)

B. Bottom-Top Clustering (Agglomerative)

C. Dendrogram

D. K-means

Solution: Both A & B

Q6. Is it reasonable to assume the same clustering results from two K-Mean clustering runs?

A. Yes

B. No

Solution: (B)

Explanation: Instead, the K-Means clustering technique talks about local minima, which may or may not equate to global minima in some situations. As a result, it's a good idea to run the K-Means method several times before making any conclusions about the clusters.

It's worth noting, though, that by using the same seed value for each run, you may get the same clustering results via K-means. However, this is accomplished by simply instructing the algorithm to select the same set of random numbers for each iteration.

Q7. Which of the following clustering techniques has a difficulty with local optima convergence?

A. Agglomerative clustering algorithm

B. K- Means clustering algorithm

C. Expectation-Maximization clustering algorithm

D. Diverse clustering algorithm

Options:

A. A only
B. B and C
C. B and D
D. A and D

Solution: (B)

Out of the options given, only K-Means clustering algorithm and EM clustering algorithm has the drawback of converging at local minima.

Q8. Which of the following is a bad characteristic of a dataset for clustering analysis-?

A. Objects with outliers
B. Objects with different densities
C. Objects with non-convex shapes
D. All of the above

Solution: (D)

Q9. Which is the following statement being incorrect?

A. k-means clustering is a method of vector quantization.
B. k-means clustering groups number of observations into k clusters.
C. k-means is same ask-nearest neighbor.
D. None

Solution: (C)

Q10. What do you understand by dendrogram?

A. A hierarchical structure
B. A diagrammatical view
C. A graph
D. None

Solution: (A)