Mon Tue Wed Thu Fri Sat

Topic 5:- Clustering

What is Clustering?

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities.

Application of Clustering:-

→ Widely used as image processing, data analysis and pattern recoginition.

→ Helps markerters to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.

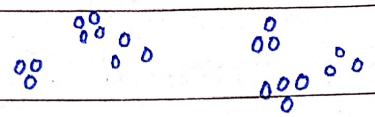
→ Can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.

→ It also helps in information discovery by classifying documents on the web.

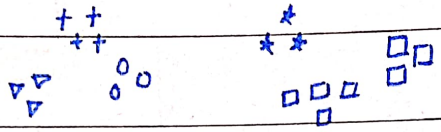
Mon Tue Wed Thu Fri Sat

تاریخ: _____

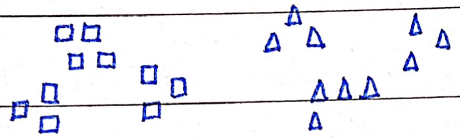
Notion of Cluster can be ambiguous.



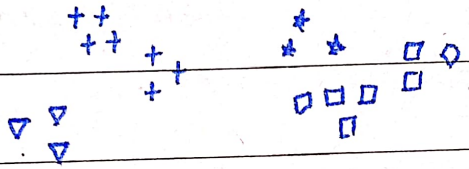
how many clusters?



Six clusters



2 clusters



Four clusters

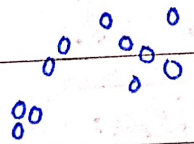
Types of Clustering:-

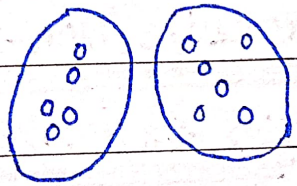
→ Partitional Clustering

→ Hierarchical Clustering

Partitional Clustering:-

Partitional Clustering decomposes a data set into set of disjoint clusters.



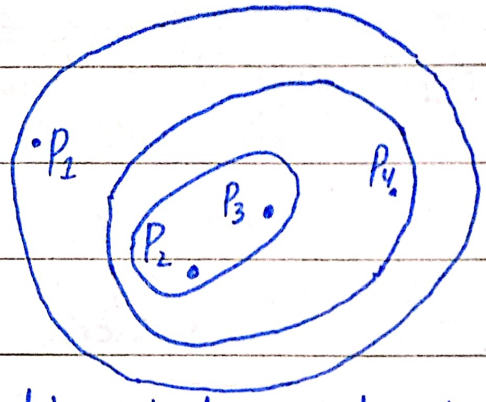


Mon Tue Wed Thu Fri Sat

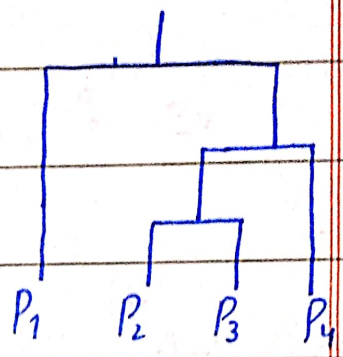
تاریخ:

Hierarchical Clustering:-

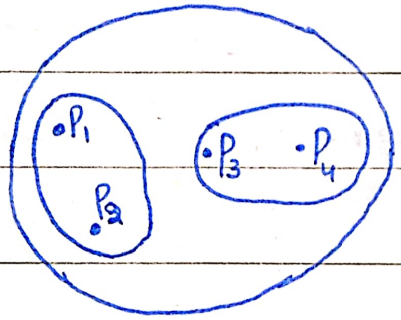
A set of nested clusters organized as a hierarchical tree.



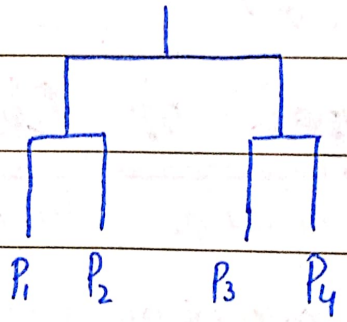
Traditional hierarchical tree



Traditional Dendrogram



Non-traditional hierarchical tree



Non-traditional Dendogram.

Other types of Clustering

Exclusive/non-overlapping VS non-exclusive/overlapping:-

Mon Tue Wed Thu Fri Sat

تاریخ:

An exclusive classification is a partition of the set of objects where each object belongs to exactly one cluster. In non-exclusive (or overlapping) clustering, objects maybe assigned to multiple clusters.

Fuzy/Soft VS Non-fuzzy (hard)

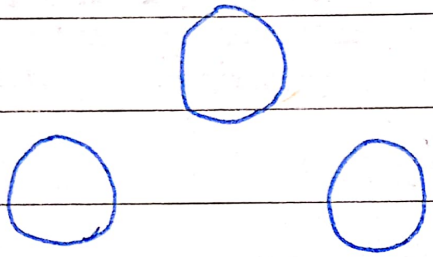
In non-fuzzy clustering, data is divided into distinct clusters, where each data point can only belong to exactly one cluster. In fuzzy clustering, data points can potentially belong to multiple clusters.

Partial VS Complete:-

A complete clustering assigns every object to a cluster, whereas a partial clustering does not. The motivation for a partial clustering is that some objects in a data set may not belong to well-defined groups.

Well-Separated Clusters:-

A cluster is a set of objects in which each object is closer (or more similar) to every other object in the cluster than to any object not in the cluster.



3- Well Separated Clusters.

Center Based Clusters:-

A cluster is a set of objects such that an object in a cluster (more similar) to the "center" of a cluster, than to the center of any other cluster.

The center of a cluster is often a centroid, the minimizer of

distances from all points in the cluster, or a medoid, the most 'representative' point of a cluster.

Center-based algorithms are not good choices for finding cluster of arbitrary shapes.

Contiguous Cluster:-

A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

Density-Based clusters:-

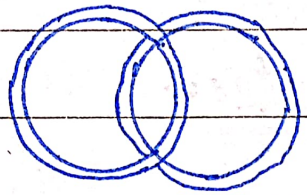
A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. It is used when clusters are irregular or interwined, and when noise and outliers are present.

Mon Tue Wed Thu Fri Sat

تاریخ:

Shared-Property or Conceptual Clusters:-

Find clusters that share some common property or represent a particular concept.



A clustering algo whould require a specific concept of a cluster to recognize these clusters effectively. The way discovering such clusters is called conceptual clustering.

Objective Function:-

The goal is to find groups or clusters of like data. The clusters will be interest to the person applying the algorithm. An objective function-based clustering algorithm tries to minimize (or maximize) a function

Mon Tue Wed Thu Fri Sat

تاریخ: _____

such that the clusters that are obtained when the minimum/maximum is reached are homogeneous.

Clustering Algorithms:-

→ K-means and its variants

→ Hierarchical clustering

→ DBSCAN

K-means Clustering:-

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In k-means, each cluster is associated with a centroid.

The main objective of k-means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

Mon Tue Wed Thu Fri Sat

تاریخ: _____

Working of K-means algorithm:-

→ Select the number of K to decide the no. of clusters.

→ Select random K points.

→ Assign each data point to their closest centroid, which will form the predefine cluster.

→ Calculate the variance and place a new centroid of each cluster.

→ Repeat the 3rd step which means reassign each data point to the new closest centroid of each cluster.

→ If any reassignment occur then go to step 4 - else go to finish.

→ The model is ready.

Also known as **Lloyd's Algorithm.**

Example:-

{2, 3, 5, 8, 9, 11, 12, 16, 18, 19, 30}

Solution with 2 Means

Let's assume 2 centroid/mediod/means.

$m_1 = 5$ $m_2 = 18$

{2, 3, 5, 8, 9, 11, 12, 16, 18, 19, 30}

$K_1 = \{2, 3, 5, 8, 9, 11\}$ $K_2 = \{12, 16, 18, 19, 30\}$.

$m_1 = 6$ $m_2 = 19$

{2, 3, 5, 8, 9, 11, 12, 16, 18, 19, 30}

$K_1 = \{2, 3, 5, 8, 9, 11, 12\}$ $K_2 = \{16, 18, 19, 30\}$.

$m_1 = 7$ $m_2 = 21$

{2, 3, 5, 8, 9, 11, 12, 16, 18, 19, 30}

$K_1 = \{2, 3, 5, 8, 9, 11, 12\}$ $K_2 = \{16, 18, 19, 30\}$.

Limitations of K-means:-

K-means has problems when clusters are of different

- → sizes
- → Densities
- → Non-globular shapes

K-means has problems when data contains outliers.

Mon Tue Wed Thu Fri Sat

تاریخ:

Variations:-

→ **K-medoids:-**

Similar problem definition as in K-means, but the centroid of the cluster is define to be one of the points in the cluster (the medoid).

→ **k-centers:-**

Similar problem definitions as in K-means, but the goal now is to minimize the maximum diameter of the clusters.

Hierarchical Clustering:-

A hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster.

In hierarchical clustering, the aim is to produce a hierarchical series of nested clusters.

A diagram called Dendogram - (A dendogram is a tree-like diagram that statistics the sequence of merges or splits) graphically represent this hierarchy.

Types of Hierarchical Clustering:-

→ Agglomerative

→ Divisive

Agglomerative:-

Initially consider every data point as an individual cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom up method).

Divise:-

Divise hierarchical clustering is precisely the opposite of the agglomerative hierarchical clustering. In this, we take into account of all the data points as a single cluster and in every

iteration, we separate the data points from the clusters which aren't comparable.

Strength of Hierarchical Clustering:-

→ It is easy to understand and implement.

→ We don't have to pre-specify any particular number of clusters.

→ can obtain any desired number of clusters by cutting the dendrogram at proper level.

→ They may correspond to meaningful classification.

→ Easy to decide the number of clusters by merely looking at the dendrogram.

Limitations of Hierarchical Clustering:-

→ Hierarchical clustering does not work well on vast amounts of data.

→ All the approaches to calculate the similarity between clusters have their own disadvantages.

Mon Tue Wed Thu Fri Sat

تاریخ: _____

→ In hierarchical clustering, once a decision is made to combine two clusters, it can not be undone.

→ Different measures have problems with one or more of the following:

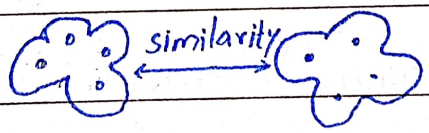
→ sensitivity to noise and outliers.

→ Faces difficulty when handling with different sizes of clusters.

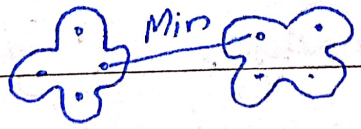
→ It is breaking large clusters.

→ In this technique, the order of the data has an impact on the final results.

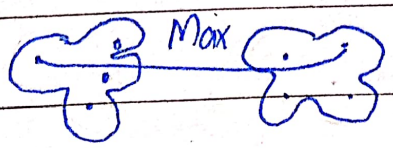
How to define inter-cluster:-



Min

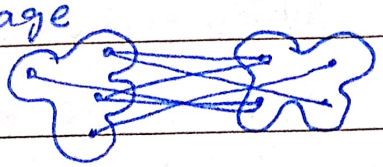


Max



Mon Tue Wed Thu Fri Sat

Group Average



Linkage Criteria:-

Determines the distance between set of observations as a function of the Pairwise distance between observations.

In **Single Linkage,** the distance between two clusters is the minimum distance between members of two clusters.

In **Complete Linkage,** the distance between two clusters is the maximum distance between members of the two clusters.

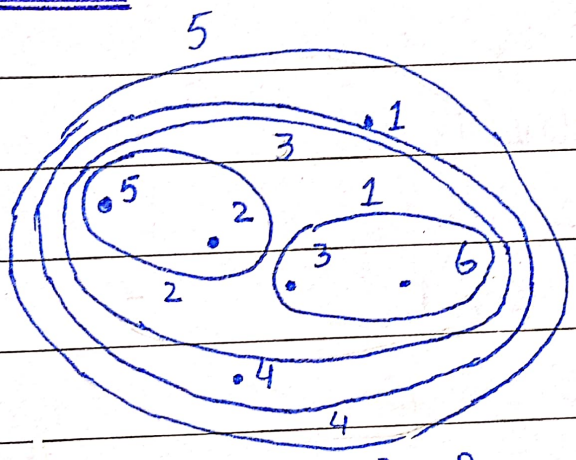
In **Average Linkage,** the distance between two clusters is the average of all distances between member of two clusters.

Mon Tue Wed Thu Fri Sat

In **Centroid Linkage**, the distance between two clusters is the distance between their centroids.

Single Linkage:-

MIN:-



	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	24	0				
P_3	22	15	0			
P_4	37	20	15	0		
P_5	34	14	28	29	0	
P_6	23	25	11	22	39	0

Mon Tue Wed Thu Fri Sat

تاريخ:

	P_1	P_2	P_3-P_6	P_4	P_5
P_1	0				
P_2	24	0			
P_3-P_6	22	15	0		
P_4	37	20	15		
P_5	34	14	28	29	0

min diff (P_1 (P_3-P_6))

22 , 23 ⇒ 22

d (P_2 (P_3 - P_6))

15 , 25 ⇒ 15

d (P_4 (P_3 - P_6))

15 , 22 ⇒ 15

d (P_5 (P_3 - P_6))

28 - 39 ⇒ 28

	P_1	P_2-P_5	P_3-P_6	P_4
P_1	0			
P_2-P_5	24	0		
P_3-P_6	22	15		
P_4	37	20	15	

Mon Tue Wed Thu Fri Sat

تاريخ: _____

d (P1 (P2 - P5))

24 → 34 ⇒ 24

d [(P3 - P6)(P2 - P5)]

15, 28 ⇒ 15

d (P4 (P2 - P5))

20, 29 ⇒ 20

	P1	(P2 - P5)(P3 - P6)	P4
P1	0		
(P2 - P5)(P3 - P6)	22	0	
P4	37	15	

d P1 [(P2 - P5), (P3 - P6)]

24, 22 ⇒ 22

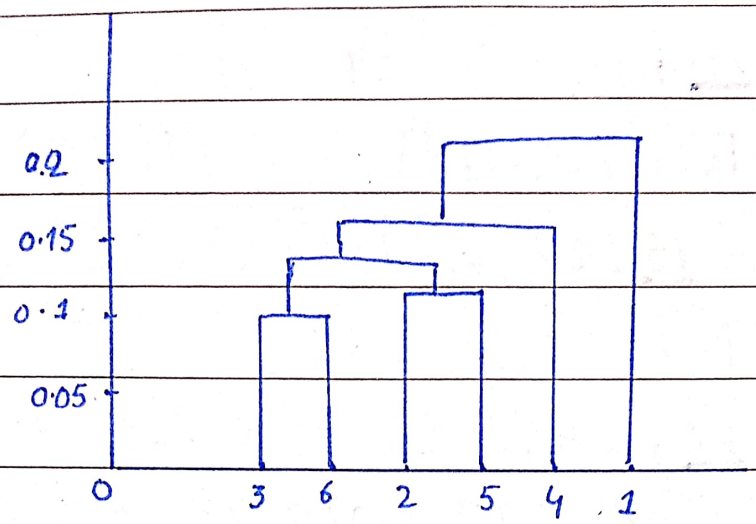
d P4 (P2 - P5), (P3 - P6)

15, 15 ⇒ 15

	P1	(P2 - P5)(P3 - P6)(P4)
P1	0	
(P2 - P5)(P3 - P6)(P4)	22	0

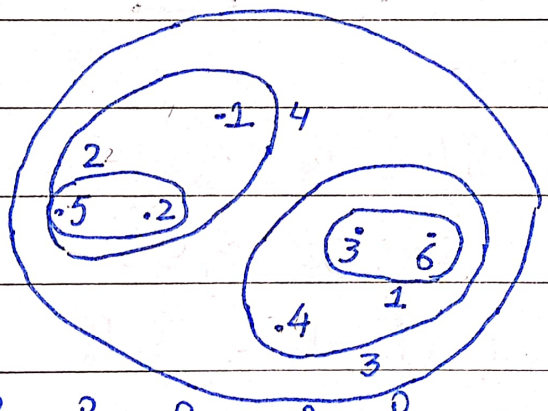
diff P1 [(P2 - P5)(P3 - P6) P4]

22, 37



Complete Linkage:-

MAX:-



	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	24	0				
P_3	22	15	0			
P_4	37	20	15	0		
P_5	34	14	28	29	0	
P_6	23	25	11	22	39	0

Mon Tue Wed Thu Fri Sat

تاریخ: _____

	P_1	P_2	P_3-P_6	P_4	P_5
P_1	0				
P_2	24	0			
P_3-P_6	23	25	0		
P_4	37	20	22	0	
P_5	34	**14**	39	29	0

max diff P_1 (P_3-P_6)

22, 23 => 23

P_2 (P_3-P_6)

15, 25 => 25

P_4 (P_3-P_6)

15, 22 => 22

Second cluster (P_2-P_5)

	P_1	P_2-P_5	P_3-P_6	P_4
P_1	0			
P_2-P_5	34	0		
P_3-P_6	23	39	0	
P_4	37	29	**22**	0

Third cluster $P_4-(P_3-P_6)$

	P_1	P_2-P_5	$P_4-(P_3-P_6)$
P_1	0		
P_2-P_5	34	0	
$(P_3-P_6)-P_4$	37	39	0

$P_1\ (P_4-(P_3-P_6))$

$37, 23 \Rightarrow 37$

$(P_2-P_5)(P_4-(P_3-P_6)$

$29, 39 \Rightarrow 39$

Next cluster $(P_1-(P_2-P_5))$

	$P_1-(P_2-P_5)$	$P_4-(P_3-P_6)$
$P_1-(P_2-P_5)$	0	
$P_4-(P_3-P_6)$	39	0

$(P_4-(P_3-P_6))(P_1-(P_2-P_5))$

$37, 39 \Rightarrow 39$

So, next all these clusters will make final cluster.

Mon Tue Wed Thu Fri Sat

تاریخ: _____