UNIT 5

Cluster Analysis: Basic Concepts and Algorithms: Overview, What Is Cluster Analysis? Different Types of Clustering, Different Types of Clusters; K-means: The Basic K-means Algorithm, K-means Additional Issues, Bisecting K-means, Strengths and Weaknesses; Agglomerative Hierarchical Clustering: Basic Agglomerative Hierarchical Clustering

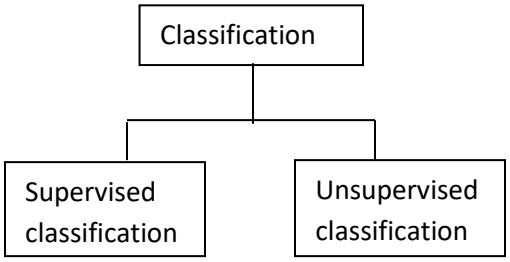
Algorithm DBSCAN: Traditional Density Center-Based Approach, DBSCAN Algorithm, Strengths and Weaknesses. (Tan &Vipin)

PART 1

Introduction

Que 1: What is Cluster Analysis? What are the applications of cluster analysis?

Classification is divided into two categories:



Supervised classification is simply known as classification in which the unknown class labels are identified.

Unsupervised classification is known as cluster analysis.

Clustering

Grouping of similar data items together is known as clustering. This is mainly used for summarization of data.

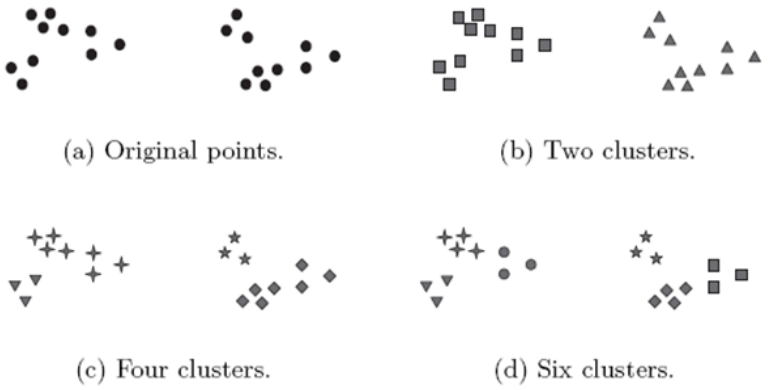


Fig: Different ways of clustering same set of points

Cluster analysis

The process of grouping data basing on the information found is known as cluster analysis. The main goal of cluster analysis is that the objects within a group be similar to one another and different from the objects in other groups.

Applications of clustering

1. **Information retrieval**
   Clustering can be used to group search results into a smaller number of clusters. For example, app like news in shots can cluster articles like crime, politics, prime minister.
2. **Climate**
   Cluster analysis has been applied to find patterns about related climatic conditions. And cluster the locations based on climate (like, volcanic area, earthquake areas, and Avalanche areas)
3. **Medicine**
   Clustering is used in medical to identify the diseases basing on the symptoms. And, patients can be clustered based on their diseases.
4. **Biology**
   Clustering is used in biology to analyze the large amount of genetic information. And, virus and bacteria can be grouped based on their behavior.
5. **Business**
   Clustering can be used to group the list of customers basing on the buying trends and sales of various products.

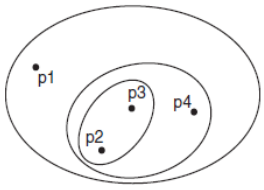
Que 2 what are different types of clustering techniques?

An entire collection of clusters is commonly referred to as **clustering**.

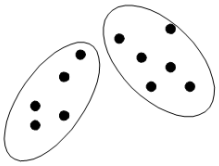
The various types of clusterings are as follows:

1. **Hierarchical versus partitional clustering**

   Hierarchical clustering: In this type of clustering, the set of clusters are nested clusters that are organized in the form of a tree known as dendrogram. Each node in the tree is the union of its children and root of the tree is the cluster containing all the objects.



   Partitional clustering: In this type of clustering, the set of clusters are unnested clusters. It is simply a division of the set of data objects into non-overlapping subsets such that each data object is assigned only to one subset.



2. **Exclusive versus overlapping versus fuzzy clustering**

   Exclusive clustering: In this type of clustering, each data object is exactly assigned to only one cluster.

   Overlapping clustering: There are some cases where one data object is assigned to more than one cluster. Such situations can be handled by overlapping clustering also known as **non-exclusive clustering** i.e. a data object can be assigned to more than one cluster.

   Fuzzy clustering: In this type of clustering, every data object can be assigned to every cluster, but basing on the membership function or weight which lies between 0 and 1. The value 0 implies that it doesn't belong to any cluster and 1 implies that it belongs to a cluster.

3. Complete versus partial clustering

Complete clustering: In this type of clustering, every data object is assigned to a cluster even if the data has some outliers or noise.

Partial clustering: The data object is assigned to a cluster only if it doesn't contain any noise or outliers.

Que3: What are different types of clusters?

1. Well-Separated clusters

A cluster is a set of objects in which each object is closer to every other object in the cluster than to any object not in the cluster i.e. the distance between any two points in different groups is larger than the distance between points in a group.

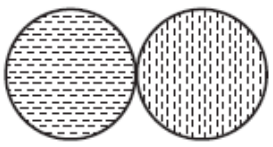
(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

2. Prototype based cluster

A cluster is a set of objects in which each object is closer to the prototype that defines the cluster than to the prototype of any other cluster. Prototype based cluster is also known as **center based cluster.**

For continuous attribute, the prototype of a cluster is centroid, i.e., the average of all points in the cluster.

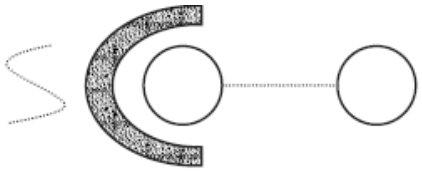
For categorical attribute, the prototype is medoid, i.e., the most representative point of a cluster.



(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

3. Graph based cluster

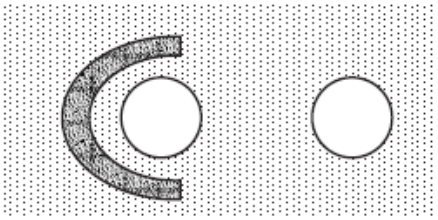
If the data is represented as a graph, where the nodes are objects and the links represent connections among objects, then a cluster can be defined as a connected component; i.e., group of objects that are connected to one another, but that have no connections to objects outside the group. **Contiguity based cluster** is an example of graph based cluster.



(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.

4. Density based clustering

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 the clusters are irregular or intertwined, and when noise and outliers are present.



(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.

K-Means

Que 4: Explain K-Means partitioning Technique with example.

(Or)

What is partitioning method? Describe any one partition based clustering algorithm.

(Or)

With a suitable example, explain K-Means Clustering algorithm.

(Or)

Consider five points {A1, A2, A3, A4, and A5} with the following coordinates as a two dimensional sample for clustering:

A1= (0.5, 2.5); A2= (0, 0); A3= (1.5, 1); A4= (5, 1); A5= (6, 2);

Illustrate the K-means partitioning algorithms using the above data set.

Ans:

k- Means defines a prototype in terms of a centroid, which is usually the mean of a group of points and is applied to objects in a continuous n-dimensional space.

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

Example1:

Consider five points {X_1,X_2,X_3,X_4,X_5} with the following coordinates as a two dimensional sample for clustering :

A_1= (0.5, 2.5); A_2= (0, 0); A_3= (1.5, 1); A_4= (5, 1); A_5= (6, 2)

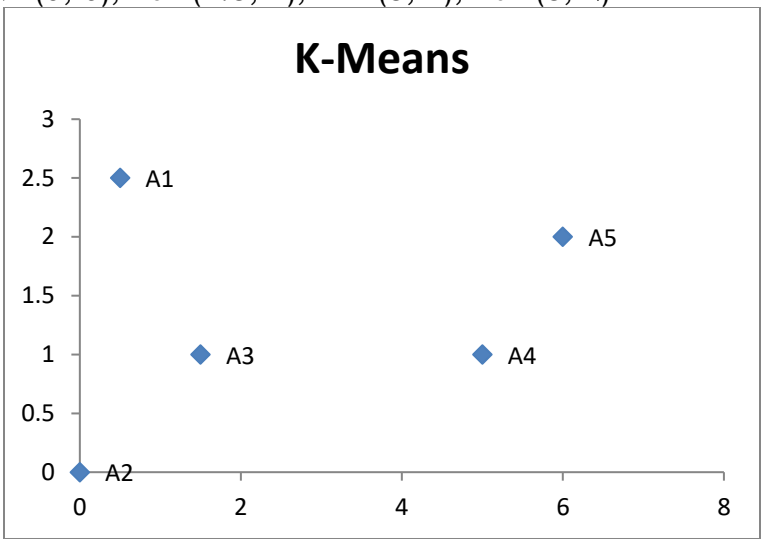


Figure 1

- Let us consider two random centroids 'A1' and 'A5' from above 5 points:
- Now measure distance from all elements to the two centroids. This can be done by using the formula $\sqrt{(\mathbf{xb}-\mathbf{xa})\mathbf{2}+(\mathbf{yb}-\mathbf{ya})\mathbf{2}}$

Calculate distance from all points to A1

Dist(A2,A1)= $\sqrt{(0-0.5)2+(0-2.5)2}$=2.54

Dist (A3,A1)= $\sqrt{(1.5-0.5)2+(1-2.5)2}$=1.802

Dist (A4,A1)= $\sqrt{(5-0.5)2+(1-2.5)2}$=4.743

Dist(A5,A1)= $\sqrt{(6-0.5)2+(2-2.5)2}$=5.52

Calculate distance from all points to A5

Dist(A1,A5)= $\sqrt{(0.5-6)2+(2.5-2)2}$=5.52

Dist (A2,A5)= $\sqrt{(0-6)2+(0-2)2}$=6.324

Dist (A3,A5)= $\sqrt{(1.5-6)2+(1-2)2}$=4.60

Dist (A4,A5)= $\sqrt{(5-6)2+(1-2)2}$=1.414

	A1	A5	Cluster to map
A1	0	5.52	A1
A2	2.54	6.324	A1
A3	1.802	4.60	A1
A4	4.743	1.414	A5
A5	5.52	0	A5

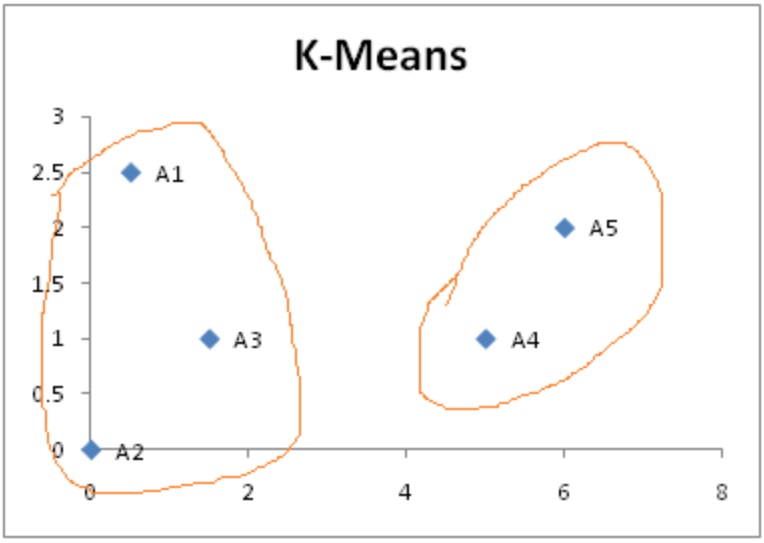


Figure 2

- Now 'A1','A2','A3' belongs to one cluster and 'A4' and 'A5' belongs to another cluster.
- Now, Calculate the average of A1, A2, and A3

  X-Axis= (0.5+0+1.5)/3=0.666

  Y-Axis = (2.5+0+1)/3=1.16

  Similarly calculate the average of 'A4' and 'A5' =

  X-Axis= (5+6)/2=5.5

  Y-Axis= (2+1)/2=1.5

  New centroids are,

  P= (0.66, 1.16)

  Q= (5.5, 1.5)
- Repeat same process, Calculate distance from all points to 'P' and 'Q':

  Calculate distance from all points to 'P'

  Dist(A1,P)= $\sqrt{(0.5-0.66)2+(2.5-1.16)2)}$=1.349

  Dist (A2,P)= $\sqrt{(0-0.66)2+(0-1.16)2)}$=1.33

  Dist(A3,P)= $\sqrt{(1.5-0.66)2+(1-1.16)2)}$=0.85

  Dist (A4,P)= $\sqrt{(5-0.66)2+(1-1.16)2)}$=4.34

  Dist (A5,P)= $\sqrt{(6-0.66)2+(2-1.16)2)}$=5.40

  Calculate distance from all points to Q.

  Dist(A1,Q)= $\sqrt{(0.5-5.5)2+(2.5-1.5)2)}$ =5.09

  Dist(A2,Q)= $\sqrt{(0-5.5)2+(0-1.5)2)}$=5.70

  Dist(A3,Q)= $\sqrt{(1.5-5.5)2+(1-1.5)2)}$=4.031

  Dist(A4,Q)= $\sqrt{(5-5.5)2+(1-1.5)2)}$=0.707

  Dist(A5,Q)= $\sqrt{(6-5.5)2+(2-1.5)2)}$=0.707

	P	Q	Cluster to map
A1	1.349	5.09	P
A2	1.33	5.70	P
A3	0.85	4.031	P
A4	4.34	0.707	Q
A5	5.40	0.707	Q

- Since ‘A1’,’A2’ and ‘A3’ belongs to one cluster and ‘A4’ and ‘A5’ belongs to another cluster. The process of clustering can be halted.

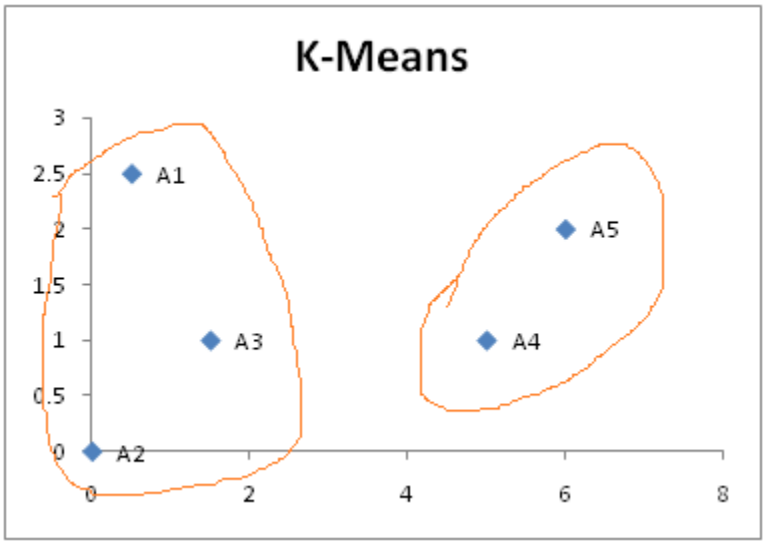


Figure 3

- Note: The process of making cluster is halted because we got the same type of cluster in 1st iteration and 2nd iteration. And there is no movement between elements in clusters.

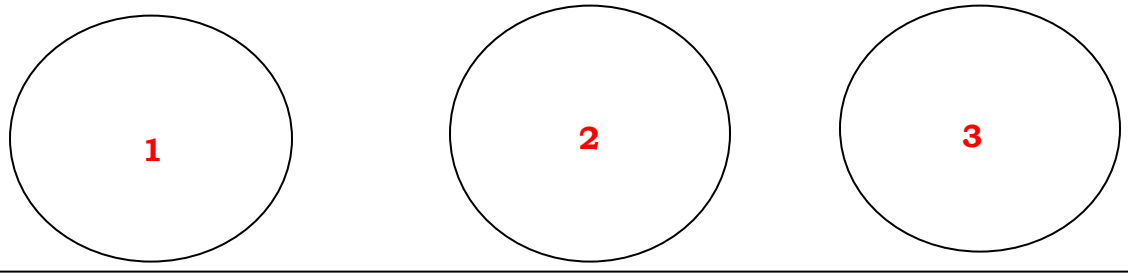
Example 2

(I recommend students to write this example in exam only if you have very-very less time)

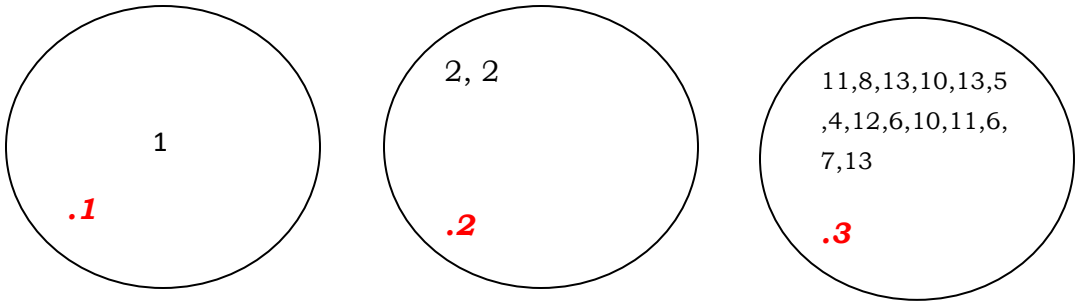
Consider samples

11,8,13,10,13,5,4,12,6,10,11,3,6,7,13,2,2,2,1,1

Consider 3 random centroids 1, 2, 3



Now, Move all elements to nearest centroid



Mean: 1 **Mean: 2** **Mean 3: 8.8**

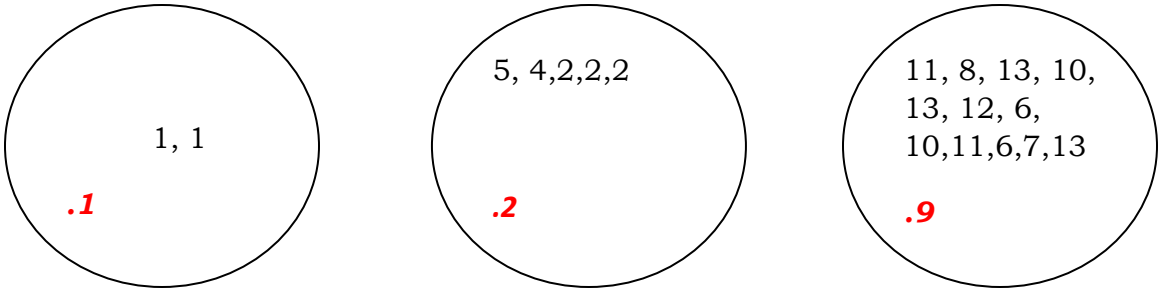
Now calculate the average of all clusters to find the new centroids:

Calculate the mean of 1st, 1=1=new centroid

Calculate the mean 2nd centroid 2, 2=2=new centroid

Calculate the mean 3rd centroid: 11,8,13,10,13,5,4,12,6,10,11,6,7,13,3=9=new centroid

Again, Move the elements to new centroids



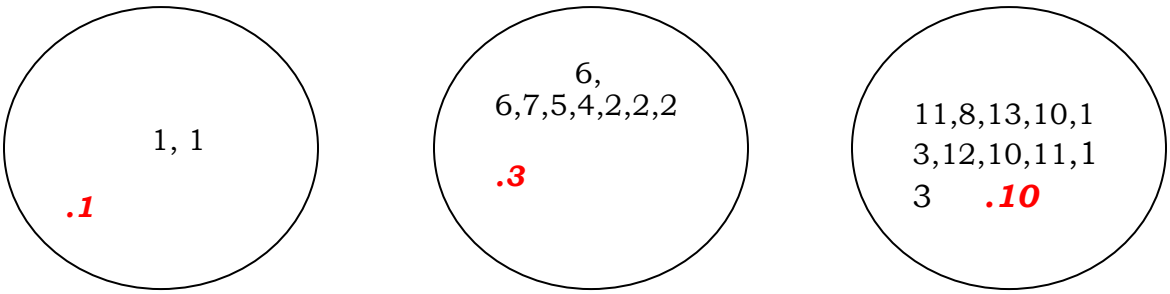
Mean: 1 **Mean: 3** **Mean: 10**

Centroid 1: 1, 1=1

Centroid 2: 5, 4, 2, 2, 2=3

Centroid 3: 11, 8, 13, 10, 13, 12, 6, 10, 11, 6, 7, 13=10

Again, Move the elements to new centroid



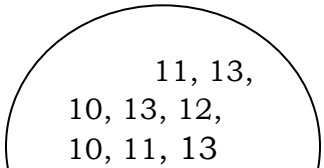
Mean: 1 **Mean: 4.25** **Mean: 11.22**

Centroid 1: 1, 1=1

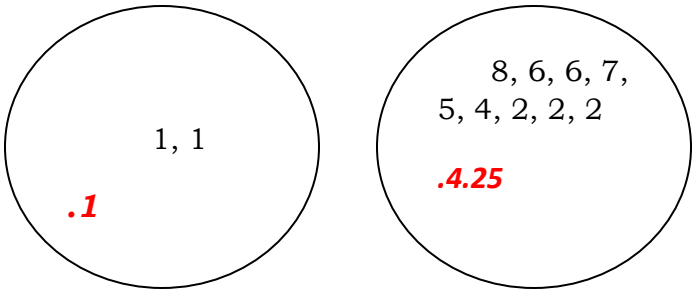
Centroid 2: 6, 6,7,5,4,2,2,2 =4.25

Centroid 3: 11,8,13,10,13,12,10,11,13 =11.22

Again, Move the elements to nearest centroid



EP



Centroid 1: 1, 1=1

Centroid 2: 8, 6, 6, 7, 5, 4, 2, 2, 2=4.66

Centroid 3: 11, 13, 10, 13, 12, 10, 11, 13=11.62

Since, all elements are in nearest centroid. So, Clustering can be stopped.

Que 5: What are the additional Issues of K-Means?

K-means: Additional issues

1. **Handling empty clusters**

   There are some cases where a cluster has only a single point i.e.; its centroids. In such case we either need to replace the centroids or eliminate it.

2. **Outliers**

   There are two ways to handle outliers. They are:

   a. **Preprocessing**

   b. **Post processing**

   In preprocessing technique, we need to first identify the outliers, remove the outliers and perform clustering. There are some cases where outliers are most important, such as, in fraud detections. In such cases, we use post processing techniques, i.e.; first cluster the data objects and later on remove the outliers.

3. **Reducing the SSE with post processing**

   SSE stands for sum of squared error.

$$\text{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

Low SSE of clusters means good clusters. There are some techniques to reduce SSE, they are as follows:

a. **Decreasing SSE by increasing number of clusters**
    i. **Split a cluster:** The cluster with the largest SSE is usually chosen and further splitted.
    ii. **Introduce a new cluster centroid:** The SSE can be decreased by replacing an old centroid with the new centroid. Or, remove an element from cluster and make it as new centroid. Now map the elements in the cluster to new centroid and make a new cluster.
b. **Decreasing SSE by decreasing number of clusters**
    i. **Disperse a cluster:** a cluster with high SSE can be dispersed and those points can be reassigned to other clusters.
    ii. **Merge two clusters:** The clusters with closest centroids are merged in order to reduce SSE.

4. Updating Centroids incrementally

In the basic K-means algorithm, centroids are updated after all points are assigned to a centroid. An alternative is to update the centroids after each assignment (incremental approach)

a. Each assignment updates centroids
b. More expensive
c. Introduces an order dependency
d. Never get an empty cluster
e. Can use “weights” to change the impact

Que 6: What is bisecting K-Means? Explain with Algorithm.

Please note: Here, there are two algorithms. 1st one is from text book and second one is a simplified version. Students can read which ever algorithm they fell easy.

This is an extension to k- means. To obtain k-clusters, split the set of all points into two clusters, select one of these clusters with large SSE(Sum of square error) to split until k-clusters have been produced. The procedure for splitting into clusters is same for k- Means and bisecting k-Means.

The algorithm for bisecting k-Means is as follows:

Algorithm 3 Bisecting K-means Algorithm.

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

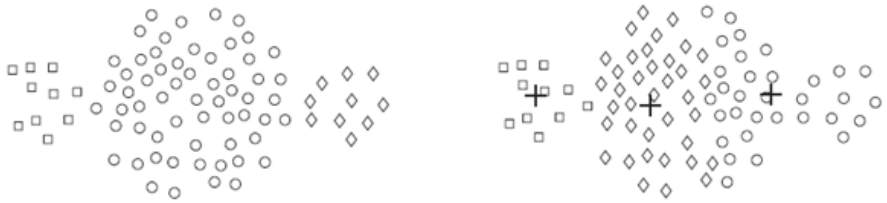
1. Repeat
2. Choose the parent cluster to be split C.
3. Repeat
4. Select two centroids at random from C.
5. Map all elements in the cluster C to nearest centroid
6. Recompute centroids by calculating mean of all elements of cluster and have new centroid assignment. (Apply K-Means on two centroids)
7. Calculate SSE for the 2 sub clusters. Choose the subclusters with maximum SSE
8. Consider the cluster with large SSE as new parent.
9. Again split the new parent
10. Until K Clusters are obtained

Que7 : Write a brief note on K-means and Different Types of Clusters.

k- Means has some difficulties when clustering have non-spherical shapes, different sizes and densities.

1. K- Means with clusters of different sizes

In the below fig, one of the cluster is much larger than the other two clusters, hence larger cluster is broken and combined with one of the smaller cluster.

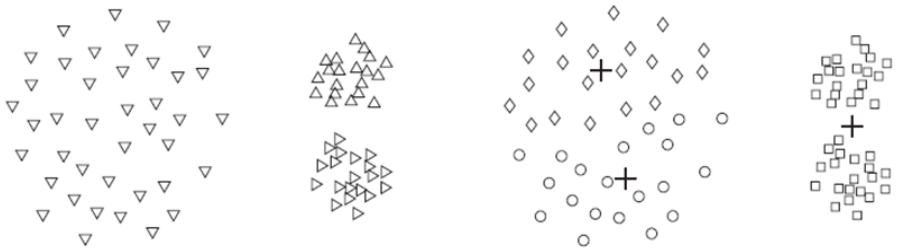


(a) Original points. (b) Three K-means clusters.

Fig: K-Means with clusters of different sizes

2. K- Means with clusters of different density

k- Means fails to find three natural clusters since the smaller cluster are much denser than the larger cluster. Therefore, larger cluster is divided into two smaller clusters.

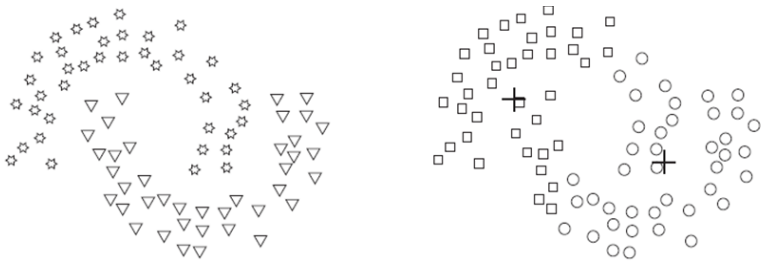


(a) Original points. (b) Three K-means clusters.

Fig: K- Means with clusters of different density

3. K- Means with non-globular clusters

In the above fig, the two natural clusters are combined, since the shapes of clusters are non- globular.



(a) Original points. (b) Two K-means clusters.

Fig: K- Means with non-globular clusters

Que 8: Write the Strengths, Weaknesses, Time and space complexity of K-Means?

Strengths

1. It is simple and useful for all varieties of data types.
2. Even though multiple iterations are performed, it is quite efficient.
3. Bisecting k- means is also efficient.

Weakness

1. It cannot handle non-globular clusters, clusters of different sizes and densities even though it can find pure sub-clusters.
2. Outliers cannot be clustered.
3. K-means is restricted for the notion of centroid.

Time complexity

O(I*K*m*n)

Where,

I- Number of Iteration for making perfect clusters

K- Number of clusters

m- Number of attributes

n- Number of elements

Space complexity

O((m+K)n)

Que 9 : Write the procedure to handle document data for clustering.

K-Means is not restricted to points and numbers. K-Means can be used to handle documents. The documents can be represented as document term matrix as shown below.

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2

Here, DOC = Document and

T1, T2,...., T8 are terms

For example, DOC5 has term T6 4 times.

The main objective of document clustering is to group similar documents. For example NEWS portal organizes article in crime related, politics, social cause, sports etc. The quality of a document cluster is represented by cohesion which is represented as:

$$\text{Total Cohesion} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} cosine(\mathbf{x}, \mathbf{c}_i)$$

Here,

- K-Number of clusters
- Ci= Cluster and
- ci= Centroid of cluster
- x= document

PART 2

Agglomerative Hierarchical Clustering

Que 10: What do you mean by Hierarchical clustering? What are different types of hierarchal clustering? How hierarchal clusters are represented?

Hierarchical clustering

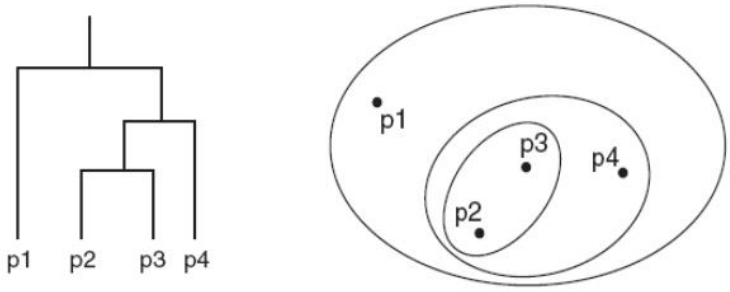
In this type of clustering, the set of clusters are nested clusters that are organized in the form of a tree known as dendrogram. Each node in the tree is the union of its children and root of the tree is the cluster containing all the objects.

There are two types of hierarchical clustering. They are:

1. Agglomerative hierarchical clustering
2. Divisive hierarchical clustering

Agglomerative hierarchical clustering

Start with the points as individual clusters and at each step, merge the closest pair of clusters.

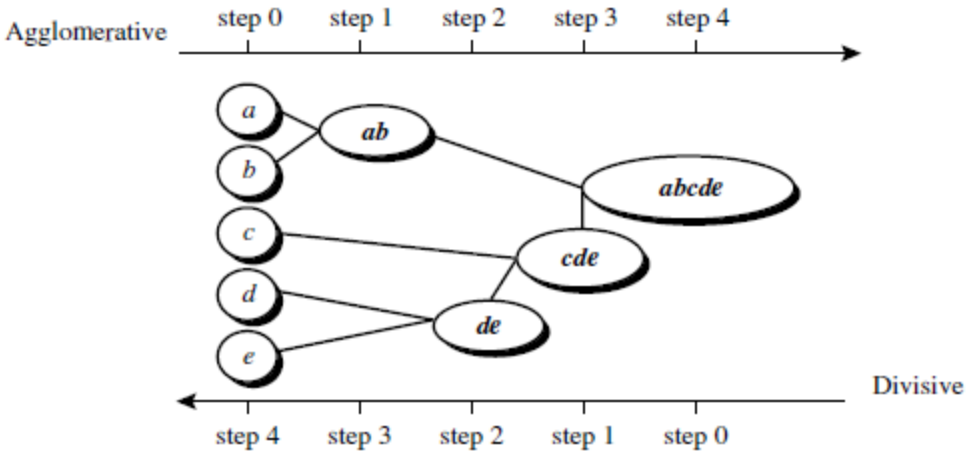


(a) Dendrogram. (b) Nested cluster diagram.

A hierarchical clustering of four points shown as a dendrogram and as nested clusters.

Divisive hierarchical clustering

Start with one i.e.; group all data objects into a single cluster, at each step, split a cluster until only single cluster of individual points remains.



Basic agglomerative hierarchical clustering algorithm

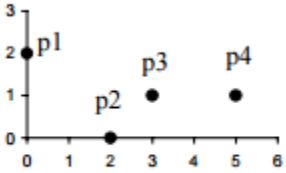
Starting with initial points as clusters, successively merge the two closest clusters until only one cluster remains.

Algorithm 8.3 Basic agglomerative hierarchical clustering algorithm.

1: Compute the proximity matrix, if necessary.
2: **repeat**
3: Merge the closest two clusters.
4: Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5: **until** Only one cluster remains.

Que 11: What is meant by cluster proximity? Explain

Cluster proximity is defined as similarity or dissimilarity between elements or clusters. They are generally defined by proximity matrix.



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0.000	2.828	3.162	5.099
p2	2.828	0.000	1.414	3.162
p3	3.162	1.414	0.000	2.000
p4	5.099	3.162	2.000	0.000

Four points and their corresponding data and proximity (distance) matrices.

Proximity matrix is a matrix containing distance between elements.

Que12: What are the methods for defining the proximity between the clusters?

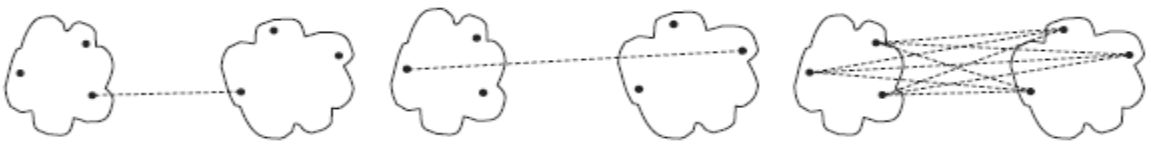
Proximity between clusters can be defined in three ways. They are:

1. Max
2. Min
3. Group average

Min defines cluster proximity between the closest two points that are in different clusters. This type of technique is also known as **single link technique.**

Max defines cluster proximity between the farthest two points that are in different clusters. This type of technique is also known as **complete link technique.**

Group average defines cluster proximity to be the average proximities of all pairs of points from different clusters.



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

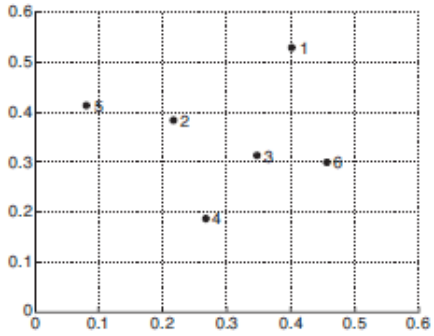
Graph-based definitions of cluster proximity

Que 13: How to define proximity between two clusters using MIN technique?

(Or)

Explain Single- Link hierarchical clustering with example?

Let us consider a data set with 6 data points.



Set of 6 two-dimensional points.

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

xy coordinates of 6 points.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Euclidean distance matrix for 6 points.

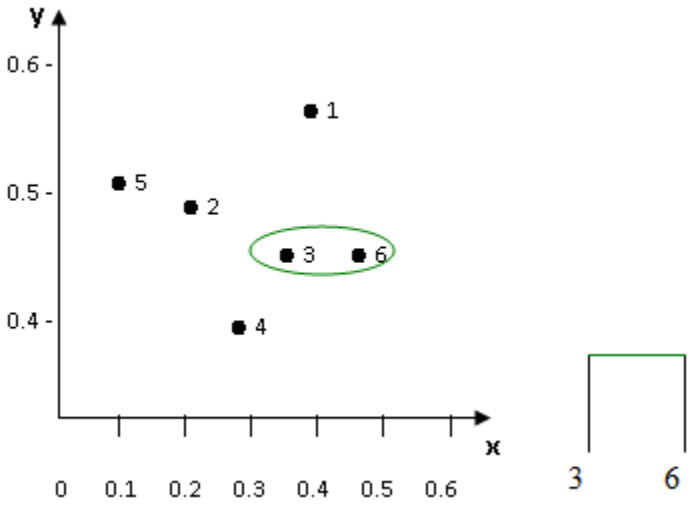
Min or single link technique

The proximity of two clusters is defined as the minimum of the distance between any two points I the two different clusters. The single link technique is good at handling non-elliptical shapes, but is sensitive to noise and outliers.

	P1	P2	P3	P4	P5	P6
P1	0	0.24	0.22	0.37	0.34	0.23
P2	0.24	0	0.15	0.20	0.14	0.25
P3	0.22	0.15	0	0.15	0.28	0.11
P4	0.37	0.20	0.15	0	0.29	0.22
P5	0.34	0.14	0.28	0.29	0	0.39
P6	0.23	0.25	0.11	0.22	0.39	0

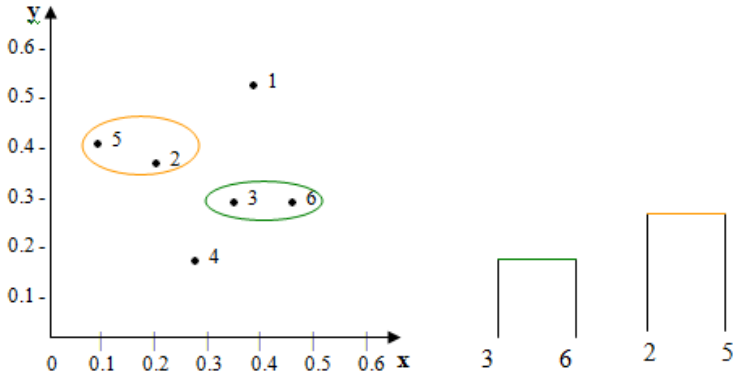
In the above table, p3, p6 is having the lowest distance, so merge the data points into a single cluster.

	P1	P2	P3P6	P4	P5
P1	0	0.24	0.22	0.37	0.34
P2	0.24	0	0.15	0.20	0.14
P3P6	0.22	0.15	0	0.15	0.28
P4	0.37	0.20	0.15	0	0.29
P5	0.34	0.14	0.28	0.29	0



The next lowest distance is for P2P5, so merge those two data points into a single cluster. The matrix obtains as follows:

	P1	P2P5	P3P6	P4
P1	0	0.24	0.22	0.37
P2P5	0.24	0	0.15	0.20
P3P6	0.22	0.15	0	0.15
P4	0.37	0.20	0.15	0



The distance between (p3, p6) and (p2, p5) would be calculated as follows:

dist((p3, p6), (p2, p5)) =MIN (*dist*(p3, p2) , *dist*(p6, p2), *dist*(p3, p5), *dist*(p6, p5))

=MIN (0.15, 0.25, 0.28, 0.39)

=0.15

dist((p3, p6), (p1)) =MIN (*dist*(p3, p1) , *dist*(p6, p1)

=MIN (0.22, 0.23)

=0.22

dist((p3, p6), (p4)) =MIN (*dist*(p3, p4) , *dist*(p6, p4)

=MIN (0.15, 0.22)

=0.15

dist((p2, p5), (p1)) =MIN (*dist*(p2, p1) , *dist*(p5, p1)

=MIN (0.24, 0.34)

=0.24

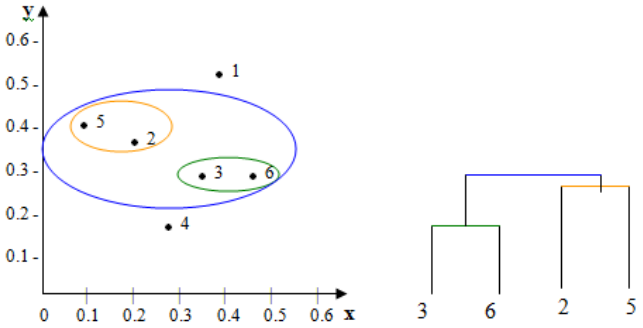
dist((p2, p5), (p4)) =MIN (*dist*(p2, p4) , *dist*(p5, p4)

=MIN (0.20, 0.29)

=0.20

So, looking at the last distance matrix above, we see that (p2, p5) and (p3, p6) have the smallest distance from all - 0.15. We also notice that p4 and (p3, p6) have the same distance - 0.15. In that case, we can pick either one. We choose (p2, p5) and (p3, p6). So, we merge those two in a single cluster, and re-compute the distance matrix.

	P1	P2P5P3P6	P4
P1	0	0.22	0.37
P2P5P3P6	0.22	0	0.15
P4	0.37	0.20	0



The distance between (P2, P5, P3, P6) and P1 would be calculated as follows:

dist((P2, P5, P3, P6), (p1)) =MIN (*dist*(p2, p1) , *dist*(p5, p1), *dist*(p3, p1) , *dist*(p6, p1))

=MIN (0.24, 0.34, 0.22, 0.23)

=0.22

The distance between (P2, P5, P3, P6) and P4 would be calculated as follows:

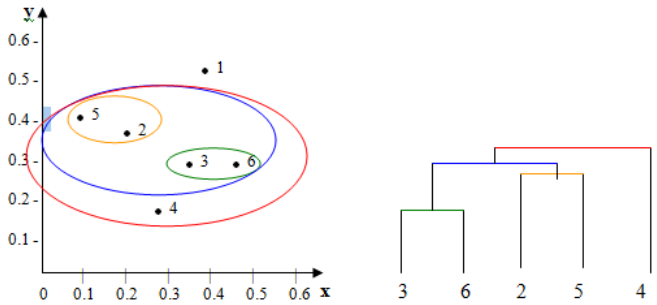
dist((P2, P5, P3, P6), (p4)) =MIN (*dist*(p2, p4) , *dist*(p5, p4), *dist*(p3, p4) , *dist*(p6, p4))

=MIN (0.20, 0.29, 0.15, 0.22)

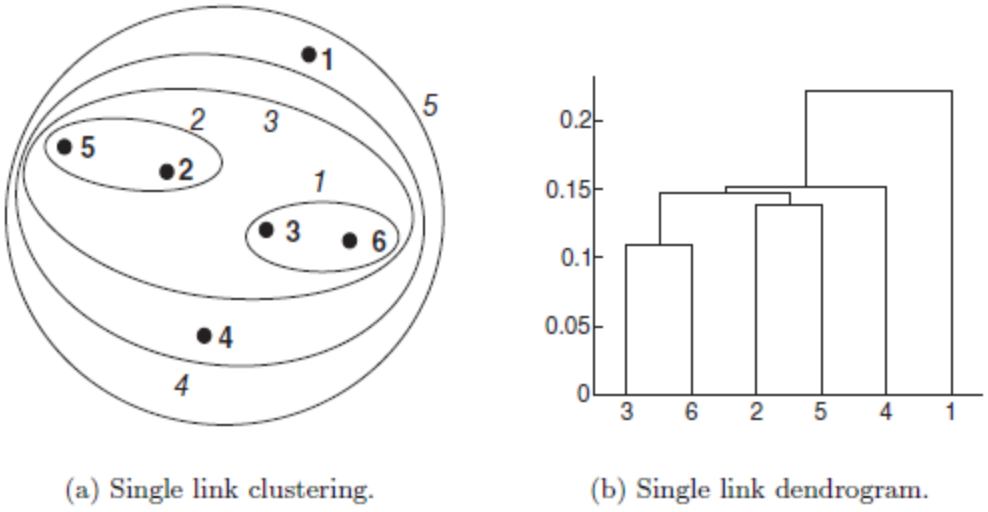
=0.15

So, looking at the last distance matrix above, we see that (p2, p5, p3, p6) and p4 have the smallest distance from all - 0.15. So, we merge those two in a single cluster, and re-compute the distance matrix.

	P1	P2P5P3P6P4
P1	0	0.22
P2P5P3P6P4	0.22	0



Finally merge clusters (P2, P5, P3, P6, P4) and P1. The clusters and dendrogram are formed as follows:



(a) Single link clustering. (b) Single link dendrogram.

Fig: single link clustering and the dendrogram of the 6 data points

Que 14: How to define proximity between two clusters using MAX technique?

(Or)

Explain complete Link hierarchical clustering with example?

(Or)

Explain Clique technique for clustering.

The proximity of two clusters is defined as the maximum of the distance between any two points in the two different clusters. Complete link is less susceptible to noise and outliers, but it can break large clusters and it favors globular shapes.

The procedure is same as Min but the difference is max distance is considered while combing to closest clusters.

For Example,

After the clusters (P2, P5) and (P3, P6) are formed, the distances are calculated as follows:

dist((p3, p6), (p2, p5)) =MAX (*dist*(p3, p2) , *dist*(p6, p2), *dist*(p3, p5),

dist(p6, p5))

=MAX (0.15, 0.25, 0.28, 0.39)

=0.39

dist((p3, p6), (p1)) =MAX (*dist*(p3, p1) , *dist*(p6, p1)

=MAX (0.22, 0.23)

=0.23

dist((p3, p6), (p4)) =MAX (*dist*(p3, p4) , *dist*(p6, p4)

=MAX (0.15, 0.22)

=0.22

dist((p2, p5), (p1)) =MAX (*dist*(p2, p1) , *dist*(p5, p1)

=MAX (0.24, 0.34)

=0.34

dist((p2, p5), (p4)) =MAX (*dist*(p2, p4) , *dist*(p5, p4)

=MAX (0.20, 0.29)

=0.29

Here the proximity is defined as maximum of distance but minimum of similarity. The lowest among all the distances is 0.22. So, merge clusters (P3, P6) and P4.

The distance between (P3, P6, P4) and remaining points are calculated as follows:

dist((p3, p6, p4), (p2, p5)) =MAX (*dist*(p3, p2) , *dist*(p6, p2), *dist*(p4, p2), *dist*(p3, p5) , *dist*(p6, p5), *dist*(p4, p5))

=MAX (0.22, 0.23, 0.37, 0.28, 0.39, 0.29)

=0.39

dist((p3, p6, p4), (p1)) =MAX (*dist*(p3, p1) , *dist*(p6, p1), *dist*(p4, p1))

=MAX (0.22, 0.23, 0.37)

=0.37

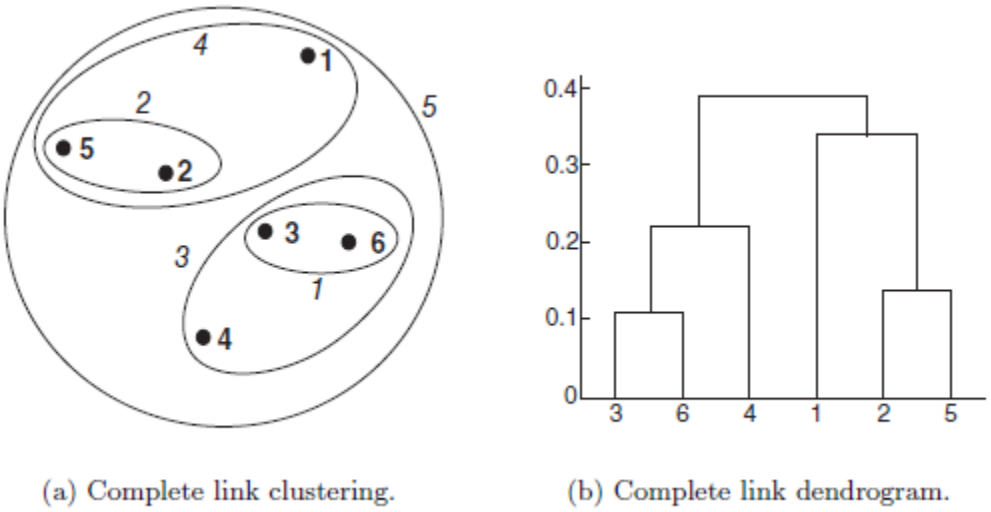
dist((p2, p5), (p1)) =MAX (*dist*(p2, p1) , *dist*(p5, p1)

=MAX (0.24, 0.34)

=0.34

The lowest among all the distances is 0.34. So, merge clusters (P2, P5) and P1.

Now merge clusters (P2, P5, P1) and (P3, P6, P4). The final clusters and dendrogram are formed as follows:



(a) Complete link clustering. (b) Complete link dendrogram.

Fig: Complete link clustering and the dendrogram of the 6 data points

Que 15: How to define proximity between two clusters using Group Average approach.

Group average technique

The proximity of two clusters is defined as the average pairwise proximity among all pairs of points in the different clusters. This is an intermediate approach between the single and complete link approaches. The cluster proximity (c_i, c_j) of clusters c_i,c_j, which are of size m_i and m_j, respectively, is expressed by the following equation:

$$proximity(C_i, C_j) = \frac{\sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_j}} proximity(\mathbf{x}, \mathbf{y})}{m_i * m_j}.$$

Initially the clustering is same as for single link and complete link. But the distance is calculated using the above equation.

$dist(\{p3, p6\}, \{p2, p5\}) = (0.15 + 0.28 + 0.25 + 0.39)/(2 * 2) = 0.26$

$dist(\{p3, p6\}, \{p1\}) = (0.22 + 0.23)/(2 * 1) = 0.225$

$dist(\{p3, p6\}, \{p4\}) = (0.15 + 0.22)/(2 * 1) = 0.185$

$dist(\{p2, p5\}, \{p1\}) = (0.24 + 0.34)/(2 * 1) = 0.29$

$dist(\{p2, p5\}, \{p4\}) = (0.20 + 0.29)/(2 * 1) = 0.245$

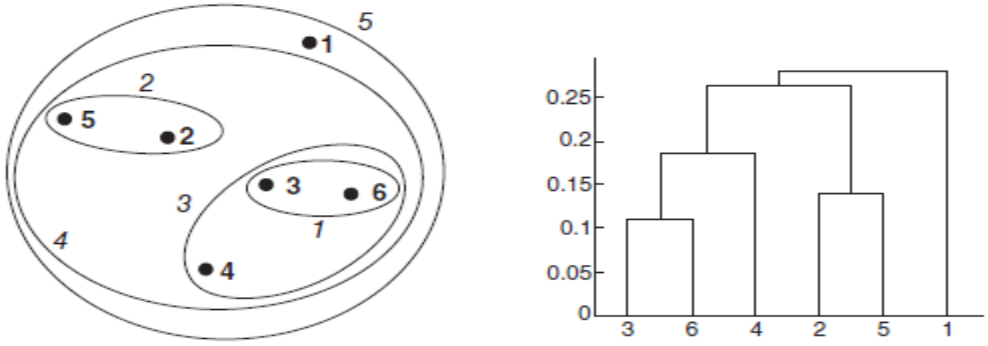
The lowest among all these distances is(P3, P6) and P4. therefore, the two clusters are merged.

The distance between (P3, P6, P4) and other data points are calculated as follows:

$dist(\{p3, p6, p4\}, \{p2, p5\}) = (0.22 + 0.28 + 0.25 + 0.39 + 0.20 + 0.29)/(3 * 2) = 0.271$

$dist(\{p3, p6, p4\}, \{p1\}) = (0.22 + 0.23 + 0.37)/(3 * 1) = 0.273$

The lowest among these two distances is 0.271, the two clusters (p3, p6, p4), (p2, p5) are merged and finally merged with P1. The final cluster and dendrogram are represented as follows:



(a) Group average clustering. (b) Group average dendrogram.

Fig: Group average clustering and the dendrogram of the 6 data points

Que 16: How to define proximity between two clusters using ward's approach.

The proximity between two clusters is defined as the increase in SSE that results after when two clusters are merged. Thus this method is similar to K-Means. This method makes ward's method distinct from other clustering techniques.

For example, Cluster A's initial SSE (Sum of square error) be 0. Let there are 3 clusters 'B', 'C', 'D'

SSE with nesting 'A' with 'B'=25.6

SSE with nesting 'A' with 'C'=10.6

SSE with nesting 'A' with 'D'=5.2

Since SSE with A and D is less, they are merged.

This method is similar to K-means because of SSE usage.

Disadvantage of Ward's Technique:

- Possibility of inversions: Two clusters that were merged may be more similar to clusters merged in previous step.

Que 17: what are the Key issues in hierarchical clustering?

1. **Lack of global objective function**
   Unlike K-Means which have some object function to measure the distance between elements and centroids. In hierarchical clustering, nearest neighbors (local merging) are merged.

2. **Ability to handle different cluster sizes**
   There are two approaches for handling clusters of different sizes. They are:
   a. **Weighted approach**, which treats all clusters equally.
   b. **Unweighted approach**, which takes the number of points in a cluster into account.

3. **Merging decisions are final**
   In this type of clustering, once two clusters are nested, Reverting back is near impossible. To address this issue, it's better to start with making clusters using K-means and then go with hierarchical clustering.

Que 18: Write the strengths, weakness, time and space complexity of hierarchical clustering?

Strengths

1. Produces better quality cluster.

2. Helps in specialized applications like creation of taxonomy (Dealing with hierarchical data).

Weaknesses

1. It is expensive.

2. Since the merges are final, it might cause some trouble to noisy and high dimensional data.

Time Complexity

O ($m^2 \log m$)

Where, 'm' is number of elements.

Space complexity

O (m^2)

DENSITY BASED CLUSTERING

Que 19: Explain center based approach for density based clustering?

In the center based approach, density is estimated for a particular point in the dataset as the number of points within the EPS (radius) of that point.

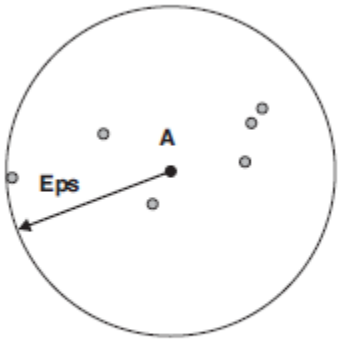
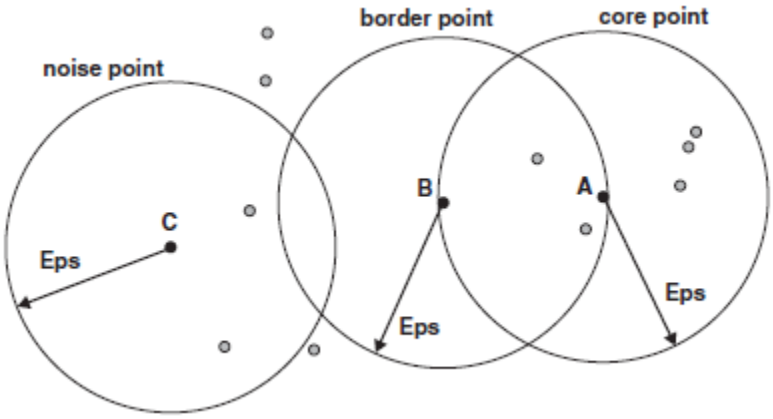


Figure 8.20. Center-based density.

A cluster is called denser, if it has minimum points (defined by Min_Pts) within the specified EPS(radius).

In density based clustering, the points are classified as core points, border points, and noise points.

1. **Core point:** These points are in the interior of the density based cluster. In the figure below, A is core point.
2. **Border point:** A border point is not a core point, but falls within the neighborhood of core point. In the figure below, B is border point.
3. **Noise point:** A noise point is any point that is neither a core point nor a border point which lies outside a density based cluster. In the figure below, C is noise point.
4. **Core object:** If the e-neighborhood of an object contains at least a minimum number, *MinPts*, of objects, then the object is called a core object.



Que 20 : Explain DBSCAN algorithm in detail.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.

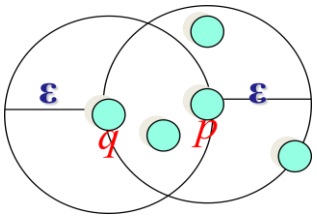
Algorithm:

Algorithm 8.4 DBSCAN algorithm.

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.

The points in the cluster space can be classified as follows:

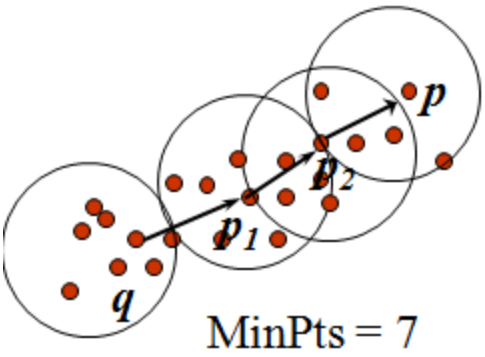
i. **Directly density reachable:** Given a set of objects, *D*, we say that an object ***p*** is directly density-reachable from object ***q*** if ***p*** is within the e-neighborhood of ***q***, and ***q*** is a core object. (An object q is directly density-reachable from object p if p is a core object and q is in p's ε-neighborhood.)



MinPts = 4

- q is directly density-reachable from p
- p is not directly density- reachable from q
- Density-reachability is asymmetric.

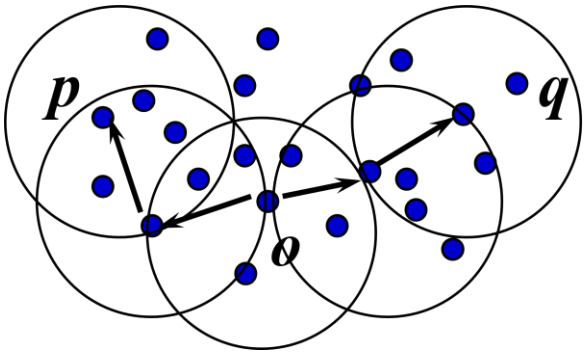
ii. **Density reachable:** An object ***p*** is density-reachable from object ***q*** with respect to e and *MinPts* in a set of objects, *D*, if there is a chain of objects ***p*1**, : : : , ***pn***, where ***p*1** = ***q*** and ***pn*** = ***p*** such that ***pi*+1** is directly density-reachable from ***pi*** with respect to e and *MinPts*, for 1 _ *i* _ *n*, ***pi*** □□***D***.



(A point p is directly density-reachable from p2; p2 is directly density-reachable from p1; p1 is directly density-reachable from q; p←p2←p1←q form a chain.)

- p is (indirectly) density-reachable from q
- q is not density- reachable from p

iii. **Density connected:** An object ***p*** is density-connected to object ***q*** with respect to e and *MinPts* in a set of objects, *D*, if there is an object **o** □*D* such that both ***p*** and ***q*** are density-reachable from ***o*** with respect to □ and *MinPts*.



- A pair of point's p and q is density-connected if they are commonly density-reachable from a point o. Density-connectivity is symmetric.

A density-based cluster is a set of density-connected objects that is maximal with respect to density-reachability. Every object not contained in any cluster is considered to be *noise*.

Que 21 : Write the strengths, weakness, time and space complexity of DBSCAN?

Ans:

Strengths:

1) Resistant to noise
2) Can handle clusters of arbitrary size.

Weakness:

1) Trouble when clusters have varying density.
2) Hard to handle high dimensional data.

Time complexity:

O(m* time to find points in EPS-Neighborhood)

Where, m is number of points.

Space complexity:

$O(m^2)$

Que 22: What is the difference between classification and clustering?

Classification	Clustering
1) Supervised learning	1) Unsupervised learning
2) Class labels are needed to learn the data	2) No Need of Class labels
3) Mostly do predictive tasks	3) Mostly do descriptive tasks
4) Algorithms: Decision tree induction, Naïve bayes approach	4) K-Means, DBSCAN
5) Mostly used to predict class label of new sample	5) Mostly used group data based on a property
6) Example application: To predict whether a customer will buy a computer or not	6) Example application: To cluster customers are premium customers, average spenders, low spenders.