Unit - 5

Clustering

"Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster is known as "clustering".

Clustering is an unsupervised Machine learning-based algorithm that comprises a group of datapoints into clusters so that the objects belong to the same group.

Clustering helps to split data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters.

It is the datamining technique used to place its data elements into their related groups.

Clustering is the process of partitioning the data (or) objects into the same class. the process of partitioning data objects into subclasses is called as 'cluster'

Clustering is also called as data segmentation because of it partitions large datasets into groups according to their similarities.

- Clustering can be helpful in many field such as
  - Marketing.

i.e., clustering helps to find group of elements with similar behaviour from a given dataset customer record.

Biology :- classification of plants and animals according to their features.

Library :- Useful in book ordering.

Difference b/w classification and clustering.

classification	clustering
→ It is used for Supervised learning	→ It is used for unsupervised learning.
→ process of classifying the i/p instances based on their corresponding class labels.	→ Grouping the instances based on their similarities without the help of class labels.
→ It has the labels so there is need of training and testing dataset for verifying the model created.	→ There is no need of training and testing dataset.
→ More complex as compared to clustering.	→ less complex as compared to classification.
Eg :- Logistic regression, Naive Bayes classifier, Support Vector Machines (SVM) etc.	Eg :- k-mean clustering Algorithm, Fuzzy C-means clustering algorithm, Gaussian (EM) Clustering Algorithm etc.

Unit - 5

Clustering :-

A group of objects such that the objects in the group are similar to one another. And the objects in one group are different from another group (or) cluster.

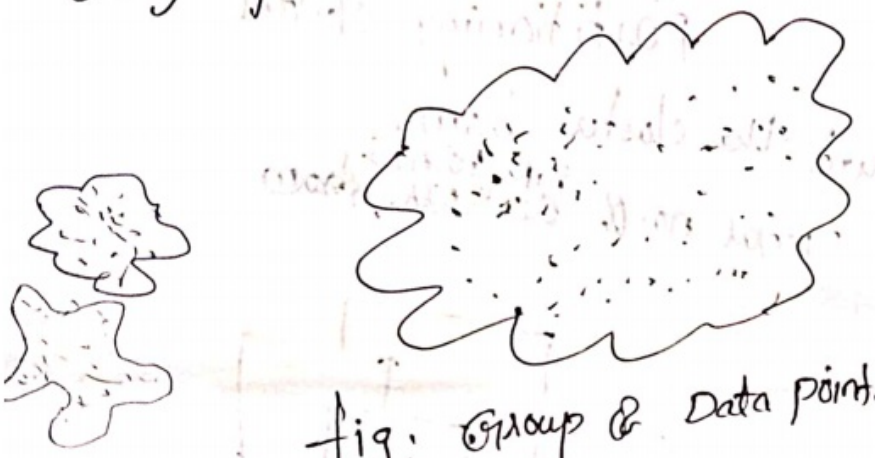


fig: Group & Data points.

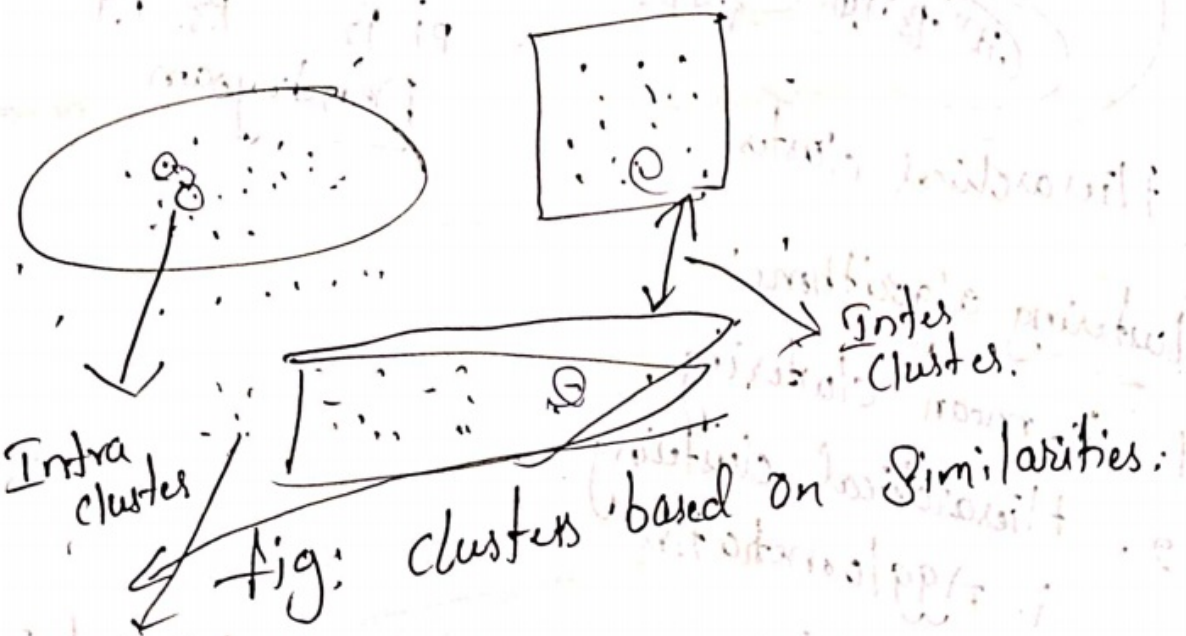
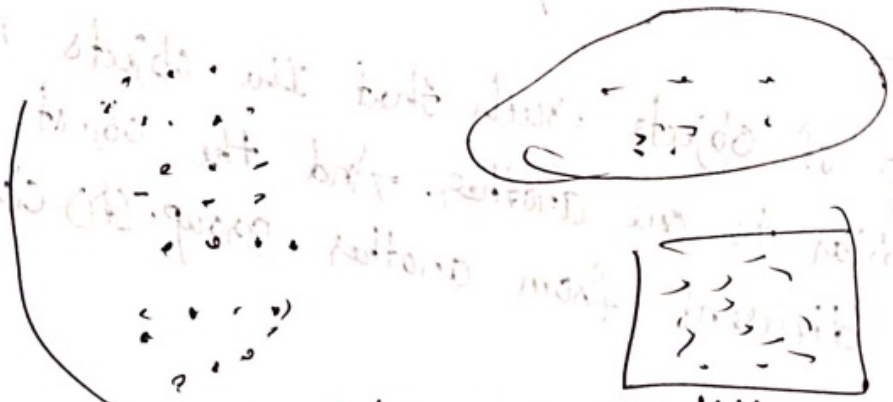


fig: clusters based on Similarities.

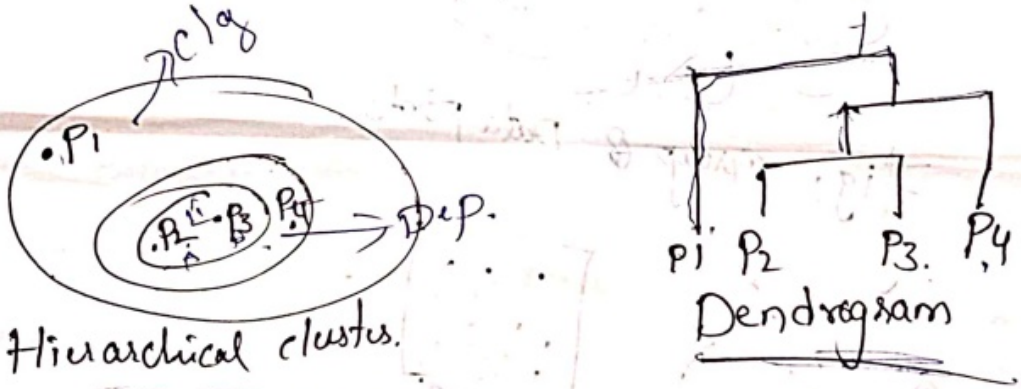
Types of clustering :-

1. partitional.



original points. partitioning cluster

2. Hierarchical :- count the clusters from Total no. of data points.



Hierarchical cluster. Dendrogram

Clustering algorithms :-

1. K-mean clustering.
2. Hierarchical clustering.
   i. Agglomerative
   ii. Divisive.
3. DBSCAN (Density Based Spatial clustering & Application with Noise).

Cluster Analysis:-

Cluster Analysis is like classification but the class label each object is not known.

Clustering is the process of grouping the data into classes (or) clusters. So that objects within a cluster have high Similarity.

In comparision to one another that can form one group its dissimilar to objects in other cluster.

Applications:-

Clustering Analysis is broadly used in many applications such as market research, pattern recognition, data analysis and image processing.

In the field of biology it can be used to derive Plant and animal taxonomies categorize genes with Similar functionalities and gain insight into structure inherent to population.

clustering also helps in classifying documents on the web for information discovery.

Clustering is also used in outlier detection application such as detection of credit card fraud.

Requirements of clustering in Datamining:

The following are some points why clustering is important in Datamining.

1. **Scalability:-**

We require highly Scalable clustering algorithms to work with large databases.

2. **Ability to deal with different kinds of attributes:-**

Algorithm should be able to work with the type of data such as categorical numerical and binary data.

3. **Discovery of clusters with attribute shape:**

The algorithm should be able to detect clusters in arbitary shapes and it should not be bounded to distance measures.

4. **Interpretability:-**

The results should be comprehensive, usable, and interpretable.

5. **High Dimensionality:-**

The algorithm should be able to handle high dimensional space instead of only handling low dimensional data.

Evaluation & clustering Algorithm :-

Three important factors by which clustering can be evaluated are

a) clustering tendency.
b) Number of clusters, k
c) clustering quality.

a) clustering tendency :-

Before evaluating the clustering performance, making sure that data set we are working has clustering tendency and does not contain uniformly distributed points is very important.

If the data doesnot contain clustering tendency, then clusters identified by any state of the art clustering algorithms may be irrelevant.

Non-uniform distribution of points in data set becomes important in clustering.

b) Number of clusters, k

Some of the clustering algorithms like k-means, require no. of clusters k, as a cluster parameter. Getting the optimal no. of clusters is very significant in the analysis.

if k is too high, each point will broadly start representing a cluster and
if k is too low, then data points are incorrectly clustered.

* Finding the optimal no. of clusters leads to granularity in clustering.

There are two major approaches to find optimal no. of clusters:

1. Domain knowledge
2. Data driven approach.

1. Domain knowledge - Domain knowledge might give some prior knowledge on finding no. of clusters. For example, in case of clustering iris data set, if we have the prior knowledge of species (virginica, versicolor), then k=3. Domain knowledg driven k value gives more relevant insights.

2. Data driven approach:- If the domain knowledge is not available, mathematical methods help in finding out right no. of clusters.

c) clustering quality :-

Once clustering is done, how well the clustering has performed can be quantified by a no. of metrics. Ideal clustering is characterized by minimal intra cluster distanc and maximal inter cluster distance.

There are majority two types of measures to assess the clustering performance.

(i) Extrinsic measures:- which require ground truth labels. Examples are Adjusted Rand index, ~~Fowlkes~~ Fowlkes-Mallows scores, Mutual Information based scores, Homogeneity, completeness and V-measure.

ii) Intrinsic Measures:- That does not requires ground truth labels, Some of the clustering performance measures are coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

Partitioning Method (K-mean) in Data mining:-

Partitioning Method:-

This clustering method classifies the information into multiple Groups based on the characteristics and Similarity of the data.

It is the data analysis to Specify the no: of clusters that has to be generated for the clustering methods.

In the partitioning method when database (D) that contains multiple (N) objects then the partitioning method constructs user-Specified (K) partitions of the data in which each partition represents a cluster and a particular region.

There are many algorithms that come under Partitioning method some of the popular ones are

k-mean

PAM (k-medoids)

CLARA algorithm (Clustering Large Applications).

Algorithm:- k-mean

Input:

k: The no. of clusters in which the dataset must be divided.

D: A dataset containing N no. of objects.

Output: A dataset of k clusters.

Method:-

1. Randomly assign k objects from the dataset (D) as cluster centres (C)
2. Re-Assign each object to which object is most similar based upon mean values.
3. Update cluster means i.e., Recalculate the mean of each cluster with the updated values.
4. Repeat step 2 until no change occurs.

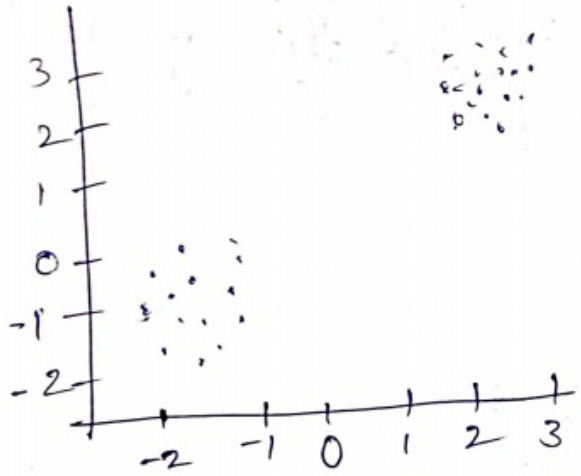
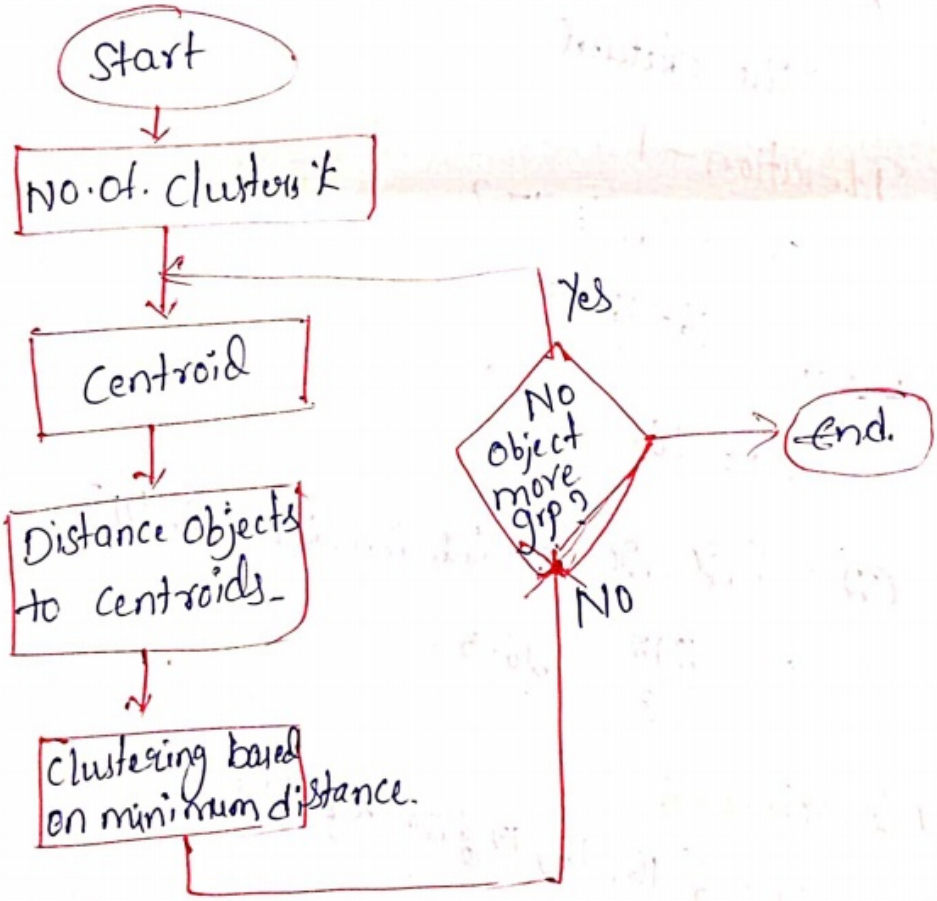


Figure:- k-mean clustering flowchart.



Step 1:- Take the mean value.

Step 2:- Find the nearest no. of mean and put in cluster.

Step 3:- Repeat ① & ② until we get the same mean.

Eg:- Suppose we want to group the visitors to a website using just their age as follows.

16, 16, 17, 20, 20, 21, 22, 23, 29, 36, 41.

Initial cluster

$k = 2$

Centroid (C1) = 16

centroid (C2) = 22.

C1 & C2 two points are choosen randomly from the dataset.

Iteration - 1

$C1 = \{16, 16, 17\}$

$\text{mean} = \frac{16+16+17}{3}$

$C1 = 16.33$

$C2 = \{20, 20, 21, 22, 23, 29, 36, 41\}$

$= \frac{212}{8} = 26.5.$

Iteration - 2.

$C1 = \{16, 16, 17, 20, 20, 21\}$

$= \frac{110}{6} = 18.33$

$C2 = \{22, 23, 29, 36, 41\}$.

$= \frac{151}{5} = 30.2.$

Iteration - 3.

$C1 = \{16, 16, 17, 20, 20, 21, 22, 23\}$

$= \frac{155}{8} = 19.375$

$C2 = \{29, 36, 41\}$

$= 35.33.$

Iteration - 4.

$C1 = \{16, 16, 17, 20, 21, 21, 22, 23\}$

$= \frac{155}{8} = 19.37$

$C2 = \{29, 36, 41\}$

$= \frac{106}{3} = 35.33.$

No change Between iteration 3 and 4, so we Stop. Therefore, we get the clusters (16 - 23) and (29-41) as 2 clusters we get using K-mean algorithm.

Clustering Method :-

- Partitioning Clustering Method
- Hierarchical clustering Method.
- Density Based clustering Method.
- Grid-based clustering method
- constraint Based method
- model Based clustering method

PAM Algorithm (Partitioning Around Medoid)

K-medoids (also called partitioning-Around Medoid) algorithm was proposed in 1987 by Kaufman and Rousseeuw.

A medoid can be difined as a point in the cluster, whose dissimilarities with all the other points in the cluster are minimum The dissimilarity of the medoid (c_i) and object (P_i) is calculated by using

$$E = |P_i - C_i|$$

The cost in k-medoids algorithm is given as

$$C = \sum |P_i - C_i|$$

Algorithm:-

1. initialize: Select k random points out of the n datapoints as the medoids
2. Associate: Each datapoint to the closest medoid by using any common distance metric methods.
3. While the cost decreases: For each medoid m, for each data O point which is not a medoid.
   - Swap m and O, associate each data point to the closest medoid, and recompute the cost.
   - If the total cost is more than that in the previous step, undo the swap.

Example:

i	x	y	C1	C2	cluster.
x1	2	6	3	7	C1
x2	3	4	0	4	C1
x3	3	8	4	8	C1
x4	4	7	4	6	C1
x5	6	2	5	3	C2
x6	6	4	3	1	C2
x7	7	3	5	1	C2
x8	7	4	4	0	C2
x9	8	5	6	2	C2
x10.	7	6	6	2	C2

Step 1:- Select two medoids initially

C1 = (3,4)

C2 = (7,4)

Manhattan Distance = $|x_1 - x_2| + |y_1 - y_2|$

Mdist |(2,6); (3,4)| = |2-3| + |6-4| = 3

mdist |(3,4), (3,4)| = |3-3| + |4-4| = 0.

Mdist |(3,8), (3,4)| = |3-3| + |8-4| = 4

mdist |(4,7), (3,4)| = |4-3| + |7-4| = 4

Step 2 :-

Clusters are

C1 : {(2,6), (3,4), (8,5), (7,6)}

Mdist | (6,2), (3,4) | = |6-3| + |2-4| = 5

Mdist | (6,4), (3,4) | = |6-3| + |4-4| = 3

Mdist | (7,3), (3,4) | = |7-3| + |3-4| = 5

Mdist | (7,4), (3,4) | = |7-3| + |4-4| = 4

Mdist | (8,5), (3,4) | = |8-3| + |5-4| = 6

Mdist | (7,6), (3,4) | = |7-3| + |6-4| = 6.

For C2.

Mdist | (2,6), (7,4) | = |2-7| + |6-4| = 7

Mdist | (3,4), (7,4) | = |3-7| + |4-4| = 4

Mdist | (3,8), (7,4) | = |3-7| + |8-4| = 8

Mdist | (4,7), (7,4) | = |4-7| + |7-4| = 6

Mdist | (6,2), (7,4) | = |6-7| + |2-4| = 3

Mdist | (6,4), (7,4) | = |6-7| + |4-4| = 1

Mdist | (7,3), (7,4) | = |7-7| + |4-3| = 1

Mdist | (7,4), (7,4) | = |7-7| + |84-4| = 0

Mdist | (8,5), (7,4) | = |8-7| + |5-4| = 2

Mdist | (7,6), (7,4) | = |7-7| + |6-4| = 2.

Step 2:-

Clusters are

C1: {(2,6), (3,4), (3,8), (4,7)}

C2: {(6,2), (6,4), (7,3), (7,4), (8,5), (7,6)}

Calculate the total cost:

cost (C, x) = $\Sigma_i |C_i - x_i|$

3-2, 4-6
1+2

Total cost = | cost ((3,4), (2,6)) + cost ((3,4), (3,8)) +
cost ((3,4), (4,7)) + cost ((7,4), (6,2)) + cost ((7,4), (6,4))
+ cost ((7,4), (7,3)) + cost ((7,4), (8,5)) + cost ((7,4), (7,6))

Total cost = 3+4+4+2+3+1+1+2.

= 20.

Step 3:-

New medoids:

C1 = (3,4) and O(7,3).

Manhattan Distance = $|x_1 - x_2| + |y_1 - y_2|$

C1: {(2,6), (3,4), (3,8), (4,7)}

C2: {(6,2), (6,4), (7,3), (7,4), (8,5), (7,6)}

Calculate the total cost

cost (C, x) = $\Sigma_i |C_i - x_i|$

For O,

Mdist | (2,6), (7,3)| = |2-7| + |6-3| = 8

Mdist | (3,4), (7,3)| = |3-7| + |4-3| = 5

Mdist |(3,8), (7,3)| = |3-7| + |8-3| = 9

Mdist | (4,7), (7,3)| = |4-7| + (7-3| = 7

Mdist |(6,2), (7,3)| = |6-7| + [2-3| = 2

Mdist | (6,4), (7,3)| = |6-7| + |4-3| = 2

Mdist | (7,3), (7,3)| = |7-7| + |3-3| = 0

Mdist |(7,4), (7,3)| = |7-7| + |4-3| = 1

Mdist |(8,5), (7,3)| = |8-7| + |5-3| = 3

Mdist |(7,6), (7,3)| = |7-7| + |6-3| = 3

i	x	y	C1	O	cluster.
x1	2	6	3	8	C1
x2	3	4	0	5	C1
x3	3	8	4	9	C1
x4	4	7	4	7	C1
x5	6	2	5	2	0
x6	6	4	3	2	0
x7	7	3	5	0	0
x8	7	4	4	1	0
x9	8	5	6	3	0
x10	7	6	6	3	0.

Calculate the total cost

cost (C, x) = Σ_i | Ci − xi |

Current Total cost = { cost ((3,4), (2,6)) + cost ((3,4), (3,8)) + cost ((3,4), (4,7)) + cost ((7,3), (6,2)) + cost ((7,3), (6,4)) + cost ((7,3), (7,4)) + cost ((7,3), (8,5)) + cost ((7,3), (7,6))

Current Total cost = 3+4+4+2+2+1+3+3

= 22.

Step 4:-

Cost of swapping of medoid C2 with O.

S = Current total cost − Previous Total cost.

S = 22 − 20 = 2 > 0.

Hence swapping C2 with O is not a good idea

Final Medoids are C1 = (3,4) and C2 = (7,4)

Clusters are

C1 = { (2,6), (3,4), (3,8), (4,7) }

C2 = { (6,2), (6,4), (7,3), (7,4), (8,5), (7,6) }.

Limitations of PAM:-

Hi Time complexity: $O(k*(n-k)^2)$

Possible combinations for every node: $k*(n-k)$

Lec Cost for each computation: $(n-k)$

or Total cost: $k*(n-k)^2$

Hence, PAM is suitable and recommended to be used for small data sets.

H

c Advantages:-

1. Deals with noise and outliers data effectively.

2. Easily implementable and simple to understand.

3. Faster compared to other partitioning algorithms.

Disadvantages.

1. Not suitable for clustering arbitarily shaped groups of data points.

2. As the initial medoids are choosen randomly the results might vary based on the choice in different runs.

Hierarchical clustering:-

Hierarchical clustering refers to an unsupervised Learning procedure that determines successive clusters based on previously defined clusters.

It works via grouping data into a tree of clusters. Hierarchical clustering starts by treating each data points as an individual clusters.

The endpoint refers to a different set of clusters, where each cluster is different from the other cluster, and the objects within cluster are the same as one another.

There are two types of hierarchical clustering.

1. Agglomerative Hierarchical clustering.
2. Divisive clustering.

1. Agglomerative Hierarchical clustering:-

Agglomerative Hierarchical clustering is one of the most common types of hierarchical clustering used to group similar objects in clusters.

Agglomerative clustering is also known as AGNES (Agglomerative Nesting) in agglomerative clustering, each data point act as a individual cluster and at each step, data objects are grouped in a bottom up method.

Th Initially, each data objects is in its cluster. At each iteration the clusters are combined with different clusters until one cluster is formed.

Agglomerative Hierarchical clustering algorithm:

Step1: Determine the Similarity b/w individuals and all other clusters.

Step2: consider each data point as an individual cluster.

Step3: Combine Similar clusters.

Step4: Recalculate the Proximity matrix for each cluster.

Step5: Repeat Step 3 and Step 4 until you get a Single cluster.

Agglomerative clustering (Single link approach).

	A	B	C	D	E	F
A	0					
B	5	0				
C	14	9	0			
D	11	20	13	0		
E	18	15	6	3	0	
F	10	16	8	10	11	0

Pair (D, E) - 3

(DE, A) - min (D: (D, A), D: (E, A))

= min (11, 18) = 11

(DE, B) - min (20, 15) = 15

(DE, C) - min (13, 6) = 6

(DE, F) - min (10, 11) = 10.

	A	B	C	DE	F	
A	0					
B	5	0				
C	14	9	0			
DE	11	15	6	0		
F	10	16	8	10	0	

Pair (A, B) - 5

TK

Pair(AB, C) = min (14, 9) = 9

(AB, DE) = min (114, 15) = 11

(AB, F) = min (10, 16) = 10.

	AB	C	DE	F
AB	0			
C	9	0		
DE	11	6	0	
F	10	8	10	0

Pair (C, DE) = 6.

Pair (AB, CDE) = min (D: (AB, C), D: (AB, E)

= min (9, 11) = 9

Pair (CDE, F) = min (D: (C, F), D: (DE, F)

min (8, 10) = 8.

	AB	CDE	F
AB	0		
CDE	9	0	
F	10.	8	0

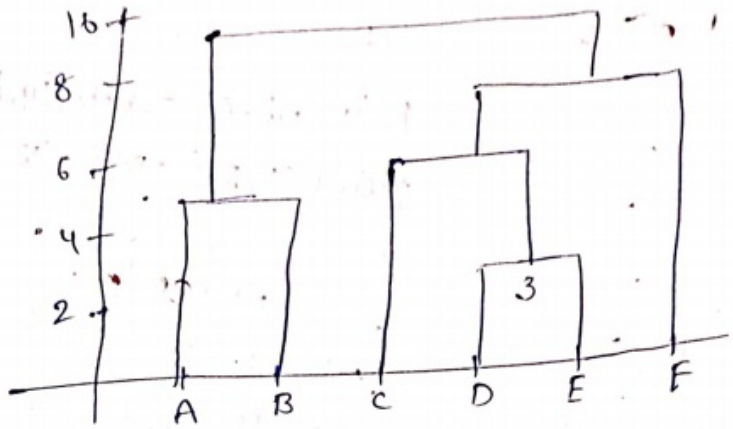
Pair (CDE, F) = 8.

Pair (AB, CDEF) = min (D: (AB, CDE), D: (AB, F)

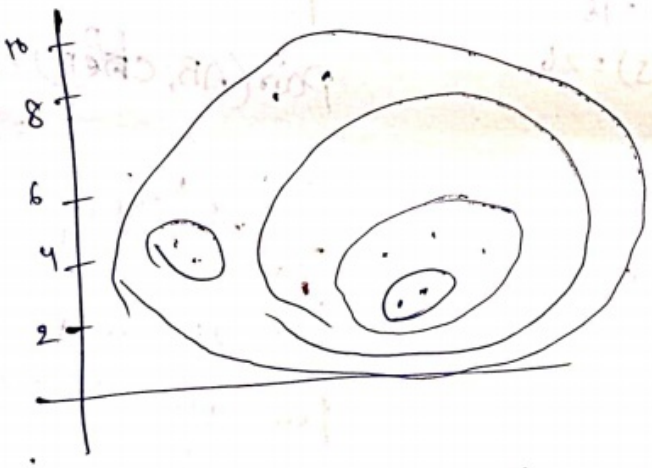
= min (9, 10) = 9.

	AB	CDEF
AB	0	
CDEF	9	0

Pair (AB, CDEF) = 9.



Dendrogram:



A

proach)

Agglomerative clustering (Complete link approach):

	A	B	C	D	E	F
A	0					
B	5	0				
C	14	9	0			
D	11	20	13	0		
E	18	15	6	3	0	
F	10	16	8	10	11	0

Pair (D, E) = 3

Pair (DE, A) = Max (D:(D,A), D:(E,A))
= Max (11, 18) = 18

Pair (DE, B) = MAX (20, 15) = 20.

	A	B	C	DE	F
A	0				
B	5	0			
C	14	9	0		
DE	18	20	13	0	
F	10	16	8	11	0

Pair (A, B) = 5

pair (AB, C) = MAX (14, 9)
= 14

Pair (AB, DE) = MAX (18, 20)
= 20

Pair (AB, F) = MAX (10, 16)
= 16.

	AB	C	DE	F
AB	0			
C	14	0		
	13	0		

Pair (C, F) = 8

Pair (AB, CF) = MAX (14, 16) = 16

Pair (CF, DE) = MAX (13, 11) = 13.

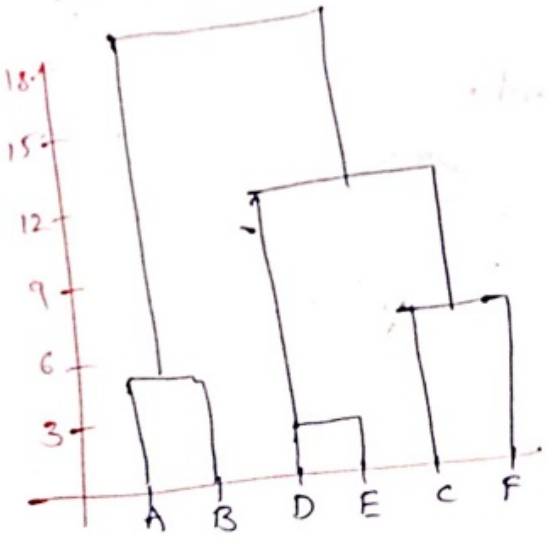
	AB	CF	DE
AB	0		
CF	16	0	
DE	20	13	0

Pair (DE, CF) = 13

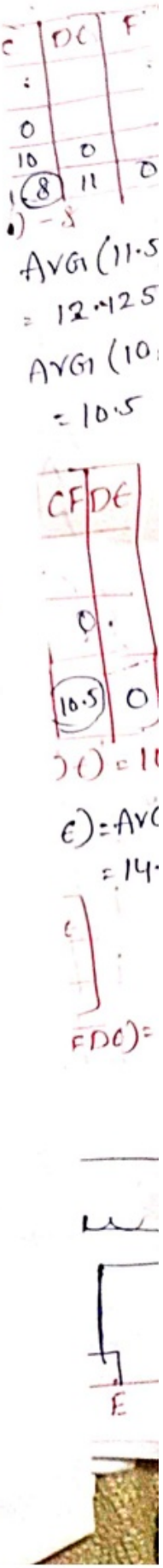
pair (AB, CFDE) = MAX (16, 20)
= 20.

	AB	CFDE
AB	0	
CFDE	20	0

Pair (AB, CFDE) = 20.



Dendrogram.



Agglomerative clustering (Average link approach)

	A	B	C	D	E	F
A	0					
B	5	0				
C	14	9	0			
D	12	20	13	0		
E	18	14	7	3	0	
F	10	16	8	10	12	0

Pair (D, E) = 3.

Pair (DE, A) - AVG (12, 18) = 15

Pair (DE, B) - AVG (20, 14) = 17

Pair (DE, C) - AVG (13, 7) = 10

Pair (DE, F) - AVG (10, 12) = 11

	A	B	C	DE	F
A	0				
B	(5)	0			
C	14	9	0		
DE	15	17	10	0	
F	10	16	8	11	0

Pair (A, B) = 5.

Pair (AB, C) - AVG (14, 9) = 11.5

Pair (AB, DE) - AVG (15, 17) = 16

Pair (AB, F) - AVG (10, 16) = 13.

	AB	C	DE	F
AB	0			
C	11.5	0		
DE	16	10	0	
F	13	(8)	11	0

Pair (C, F) - 8

Pair (AB, CF) = AVG (11.5, 13)
= 12.125

Pair (CF, DE) = AVG (10, 11)
= 10.5

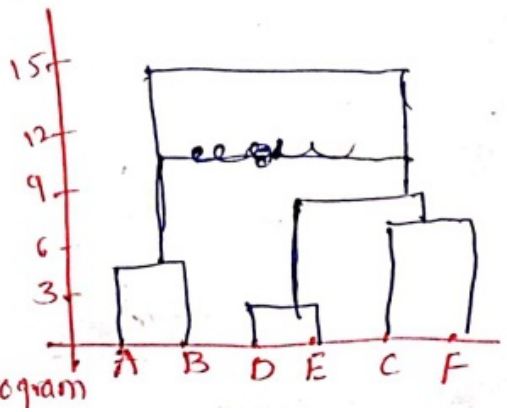
	AB	CF	DE
AB	0		
CF	12.125	0	
DE	16	(10.5)	0

Pair (AB, CFDE) = 10.5

Pair (AB, CFDE) = AVG (12.125, 16)
= 14.125.

	AB	CF DE
AB	0	
CFDE	14.125	0

Pair (AB, CFDE) = 14.125

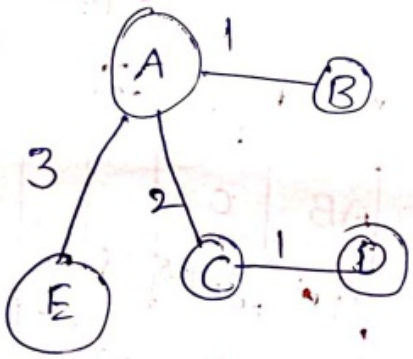


Dendrogram

Divisive method:-

	A	B	C	D	E
A	0				
B	1	0			
C	2	2	0		
D	2	4	1	0	
E	3	3	5	3	0

Edge (A-B) (C-D) (A-C) (A-D) (B-C) (A-E) (B-E) (D-E) (B-D) (C-E)
Cost 1 1 2 2 2 3 3 3 4 5



V = 5
E = 4
V-1 = E

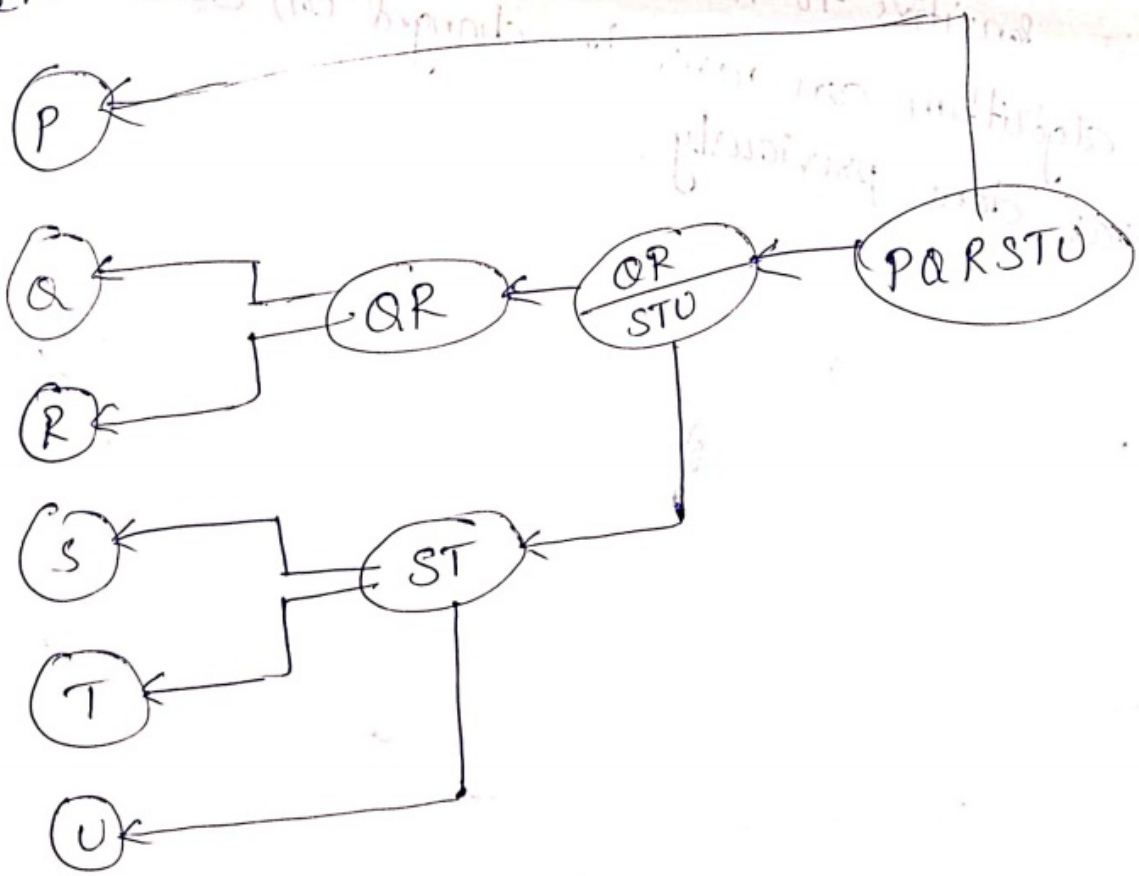
Apply Divisive - from largest to smallest

E / ABCD
E / AB / CD
E / AB / C / D
E / A / B / C / D.

Divisive Hierarchical clustering. It follows a top to bottom approach.

Divisive Hierarchical clustering is exactly the opposite of agglomerative hierarchical clustering.

∴ In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration the data points that are not similar are Separated from the cluster. The Separated data points are treated as an individual cluster. Finally, we are left with N clusters.



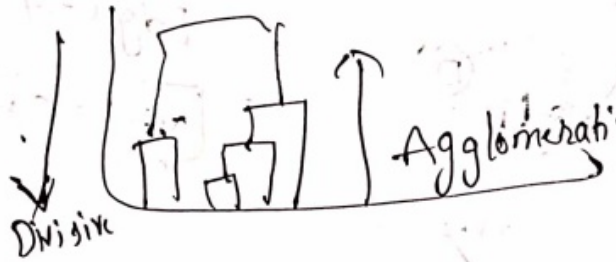
Advantages of divisive hierarchical clustering

1. It is simple to implement and gives the best output in some cases.
2. It is easy and results in a hierarchy, a structure that contains more information.
3. It does not need us to pre-specify the no. of clusters.

Disadvantages:-

1. It breaks the large clusters.
2. It is difficult to handle different sized clusters and convex shapes.
3. It is sensitive to noise and outliers.
4. The algorithm can never be changed (or) deleted once it was done previously.

Divisive cluster method:- It follows a top to bottom approach.



	A	B	C	D	E
A	0				
B	1	0			
C	2	2	0		
D	2	4	1	0	
E	3	3	5	3	0

Step1:- Compute a minimum spanning tree (MST) for the given adjacency matrix.

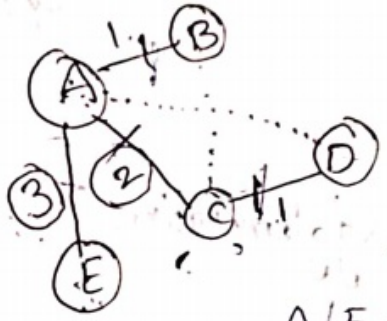
Step2:- Now create a new cluster by breaking the link corresponding to the largest distance.

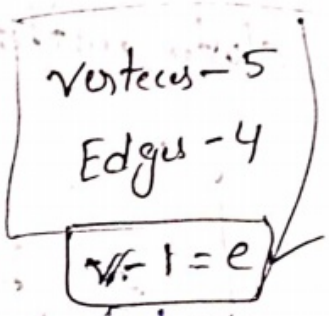
Step3:- until the condition if no condition until we end with only single cluster.

Edge (A-B) (C-D) (A-C) (A-D) (B-C) (A-E) (B-E) (D-E)
cost 1 1 2 2 2 3 3 3

(B-D) (C-E)
4 5

place low to high cost & avoid following loops





A/E = 3

E/ABCD

E/AB/CD

E/AB/C/D

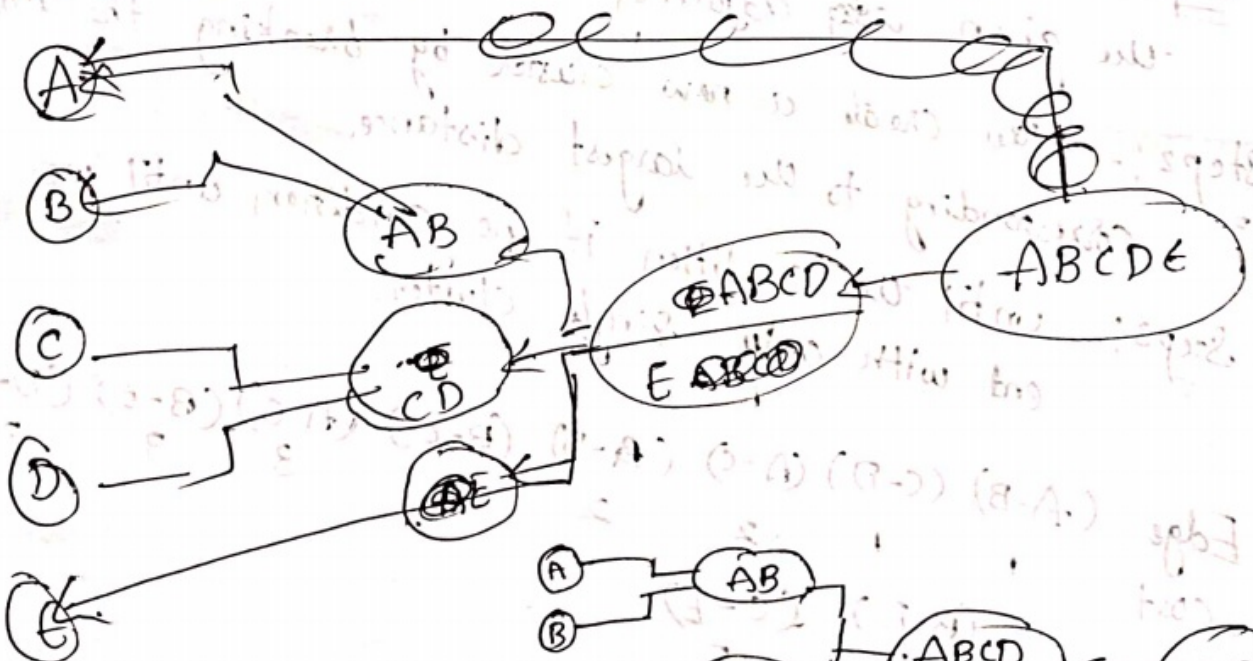
E/A/B/C/D.

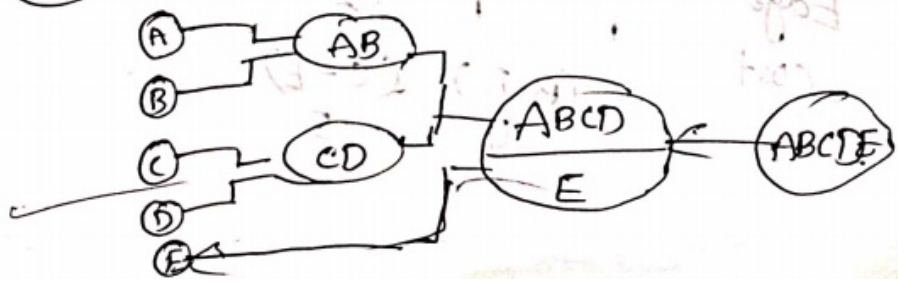
Largest edge between A & E

now we end with cluster

E & A,B,C,D

now next edge largest A & C





Key Issues in Hierarchical clustering:-

1. Lack of a Global objective function:-

Agglomerative hierarchical clustering techniques perform clustering on a local level and as such there is no global objective function like in the K-Means algorithm. This is actually an advantage of this technique because the time and space complexity of global functions tends to be very expensive.

Ability to Handle Different cluster Sizes:- we have to decide how to treat clusters of various sizes that are merged together.

Merging Decisions are final:- One downside of this technique is that once two clusters have been merged they cannot be split up at a later time for a more favorable union.

Outlier Analysis:

1. The data objects that are noncuniform (or) Inconsistent with the remaining data set is known as an outlier. outliers donot comply with the corresponding data model (or) behaviour. They may occur due to the execution error (or) inherent data variability. Usually, they are preferred to be minimized (or) eliminated by the data mining algorithms. But, this elimination may lead to certain problems, because the outliers may have some hidden information, which may be essential to some applications. An outlier may be defined as noise for one person could be a sight for another person.

Given a set of n data objects (or) data points and k no of expected outliers. outliers identifies the objects that are highly inconsistent (or) irrelevant with the dataset model. The process of analyzing and ~~dependency~~ detecting data for outliers is known as outlier mining. The outlier mining problem is divided into two subproblems

1. The data is that is irrelevant with respect to the given dataset is to be defined.
2. An efficient method that is used to mine the outliers is to be determined.

re:

The data objects that are noncuniform (or)

1. Inconsistent with the remaining dataset is known as an outlier. outliers donot comply with the corresponding data model (or) behaviour. They may occur due to the execution error (or) inherent data variability. Usually, they are preferred to be minimized (or) eliminated by the data mining algorithms. But, this elimination may lead to certain problems, because the outliers may have some hidden information, which may be essential to some applications.

An outlier may be defined as noise for one person could be a sight for another person.

Given a set of n data objects (or) data points and k no. of expected outliers. outliers identifies the objects that are highly inconsistent (or) irrelevant with the dataset model. The process of analyzing and ~~depending~~ detecting data for outliers is known as outlier mining. The outlier mining problem is divided into two subproblems

1. The data is that is irrelevant with respect to the given dataset is to be defined.
2. An efficient method that is used to mine the outliers is to be determined.

Applications of outlier Mining:-

1. Outlier Mining is applicable in fraud detection so as to indicate the illegal activities that are being performed.
2. It is also applicable in telecommunication networks, where the wrong usage of telecommunication services (or) credit cards is being detached.
3. It is applicable in customized marketing, where the customers behaviour in spending money with too low (or) extremly high income is determined.
4. It is applicable in the medical analysis, where unusual response to several medical treatments are found.

Outlier Methods:-

When time-series data and multidimensional data is analyzed, the detection of outliers become difficult. The difficulty can be avoided by the data visualization method. But, this method may also become weak because of the vision of the humans that are good at detecting numerical data of two (or) three dimensions but not at detecting more than three dimensions of data. Beside the data visualization method, there are four computer-based methods for outlier detection. They are:

1. Statistical Distribution - based outlier detection.
2. Distance - based outlier detection.
3. Density - based outlier detection.
4. Deviation - based outlier detection.

1. Statistical Distribution - Based Outlier Detection :-

A Statistical Distribution based outlier detection uses either Normal (or) poisson distribution model and discordancy test to identify outliers from the given dataset.

To conduct a discordancy test one must have the knowledge of dataset parameters (i.e., the assumed distribution), distribution parameters (i.e., mean and variance) and the no. of outliers expected from a given dataset.

Working of a Discordancy Test :-

There are two hypothesis analyzed by a discordancy test. They are:

1. Working Hypothesis :-

A working Hypothesis denoted by H is a statement whose complete dataset containing n no. of objects are taken from a single initial distribution model represented by F. The working hypothesis is expressed as,

$$H: O_i \in F$$

where, $i \rightarrow 1, 2, \ldots, n$.

If the significant evidence supporting the rejection of working hypothesis is not presented then the hypothesis is held back. This confirmation regarding rejection is verified by the discordancy test. It checks whether object O_i is significantly large (or) small with respect to the distributions model F. There are several test statistics to be used as a discordancy test. Significance probability is given by,

$$SP_i(V_i) = Prob(T > V_i)$$

where,

$V_i \rightarrow$ value of statistical's T for object O_i.

$T \rightarrow$ statistic.

For small values of $SP(V_i)$ which result in O_i as discordant, the working hypothesis is rejected. Therefore in this manner discordancy test evaluates the acceptance (or) rejection of a working hypothesis.

2. Alternative Hypothesis :-

An alternative hypothesis denoted by H states that another model G is considered, from where the object O_i is taken. This evaluation of object O_i relies on which model F is considered as O_i may appears as an outlier under one model but, can be a valid value under another model If O_i is an outlier in alternative distribution then a working hypothesis is rejected. Therefore, alternative distribution

than a working

Therefore, alternative hypothesis is very important to evaluate the power of the test.

Alternative Distributions:

There are different types of alternative distributions. They are,

i) Inherent Alternative Distribution:

If all the objects arrive from a distribution F_i, then the working hypothesis is rejected and alternative hypothesis takes its objects from a distribution G_i.

$$H: O_i \in G_i, \text{ for } i = 1, 2, \ldots, n.$$

The distributions, F and G_i can be different (or) may be same distributions with varying parameters.

It is necessary for a distribution G_i to have a capability of producing outliers.

i, Mixture