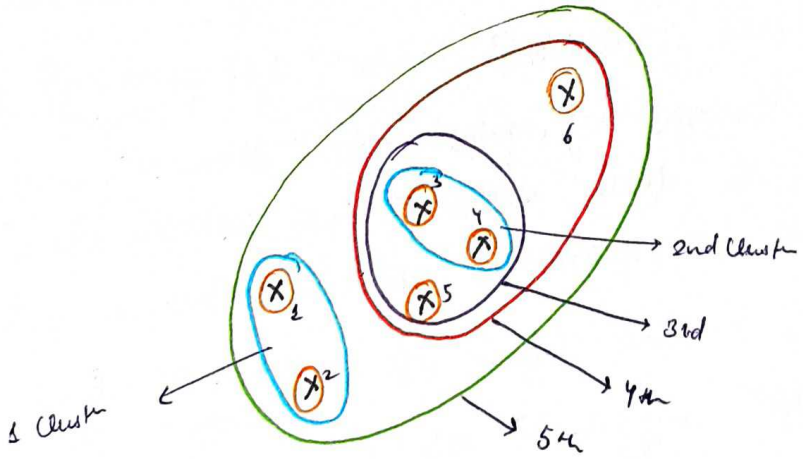
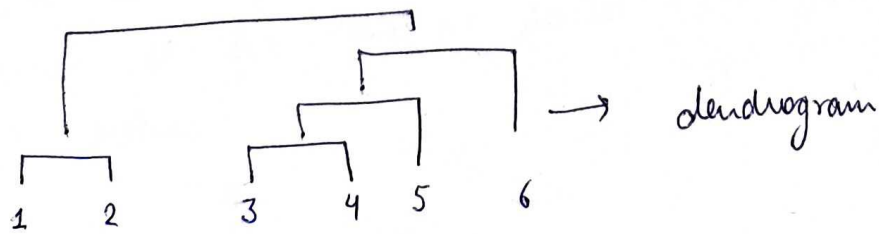
Hierarchical Clustering

Hierarchical Clustering is a method of cluster analysis used in data mining. It seeks to build a hierarchical cluster in a step-by-step manner. There are two main types of hierarchical clustering.

1. **Agglomerative (Bottom-Up Approach):**
   - **Initial Step:** Starts by treating each data point as a seprate cluster. So, if there are N data points you begin with N clusters.
   - **Clustering Process:** In each step, the algo merges the two clusters that are closest to each other until all the clusters are merged into one big cluster containing all data points.
   - **Dendrogram:** The result can be represented in a tree-like structure called a dendrogram, which shows the arrangement of the cluster and their proximity.



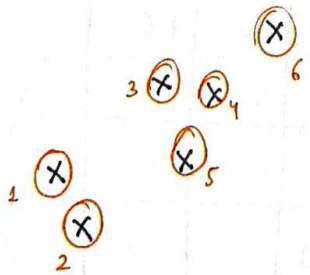


2. Divisive (Top-Down Approach):

- Initial Step: Begins with all data points in a single cluster.
- Clustering Process: At each step, the algo splits the cluster until each cluster contains only one data point.
- Top-Down Splitting: This is less common compared to agglomerative clustering and is computationally more intensive.

Agglomerative Algorithm

1. Initialization: Treat each Data point as a separate cluster. Thus, if you have N data points, you start with N cluster, each containing just one data point.

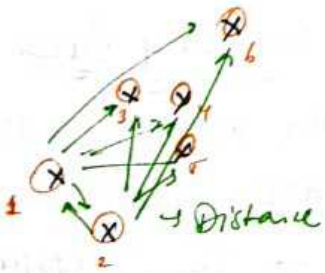


2. Compute Distance Matrix: • Calculate the distance both each pair of clusters. Common distance metrics include Euclidean, Manhattan and Cosine distances • The choice of distance metrics can significantly affect the outcome of the clustering.

- This results in an N * N distance matrix, where the distance between a cluster and itself is zero.

	P_1	P_2	P_3	P_4	P_5	
P_1	0					
P_2	1	0				
P_3	2.5	2.5	0			
P_4	3	3.5	1.2	0		
P_5	2	2.3	1.5	1.2	0	
P_6	5	6	2.6	2.1	3.2	

Closest point (P_2, P_1)



3. Find the Closest Cluster: Identify the two clusters that are closest to each other based on the distance matrix.

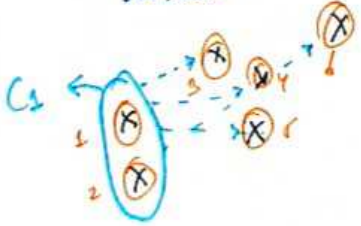
4. Merge Clusters: • Combine the two closest clusters into a single cluster.

- This step reduces the total number of cluster by one.

	C_1	P_3	P_4	P_5	P_6
C_1	0				
P_3	2.2	0			
P_4	2.0	1.2	0		
P_5	1.8	1.5	1.7	0	
P_6	6	2.8	2.1	3.2	0

→ already know distance

Update distance



5. Update Distance Matrix : • Recalculate the distance between the new cluster and all the existing cluster.

• The method of recalculating the distance depends on the linkage criterion used.

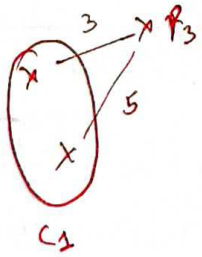
Common linkage criteria include:

- Single Linkage: Distance between two cluster is defined as the shortest distance between any two points in the cluster.
- Complete Linkage: Distance is the longest distance between any two points in the clusters.
- Average Linkage: Distance is the average distance bet^n all pairs of points in the clusters.
- Ward's Method: Distance is calculated as the increase in the total within cluster variance after merging the cluster.

Linkage criterion → Distance between 2 cluster

→ Distance bet^n 1 cluster and 1 core pts

- min
- max
- avg
- Ward

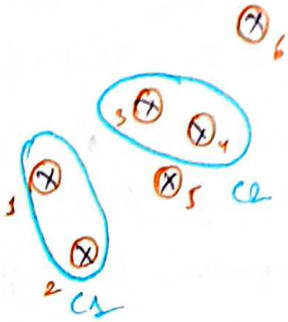


→ using min method

→ Acc to min method lowest or minimum dist^n choose.

→ 3 is Dist^n bet^n C_1 and P_3

	C_1	P_3	P_4	P_5	P_6
C_1	0				
P_3	2.2	0			
P_4	2.8	1.2	0		
P_5	1.8	1.5	1.7	0	
P_6	6	2.6	2.1	3.2	0

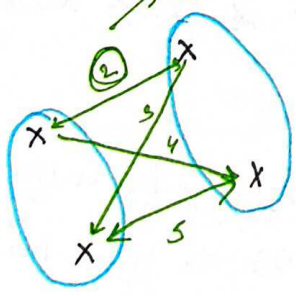


	C_1	C_2	P_5	P_6
C_1	0			
C_2	2.7	0		
P_5	1.8	1.4	0	
P_6	6	2.1	3.2	0

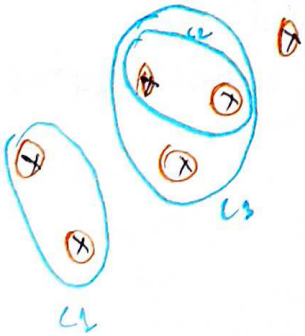
→ small distⁿ

6. Repeat :

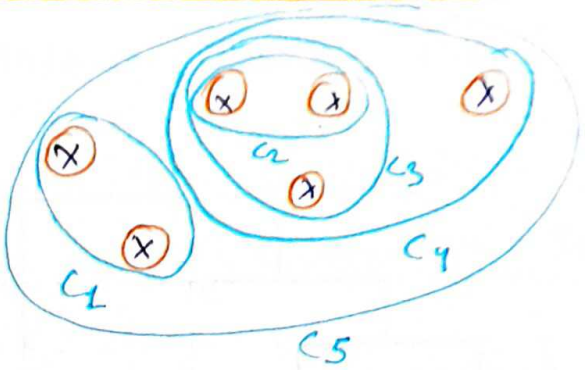
- Repeat steps 3 to 5 until all data points are merged into single cluster

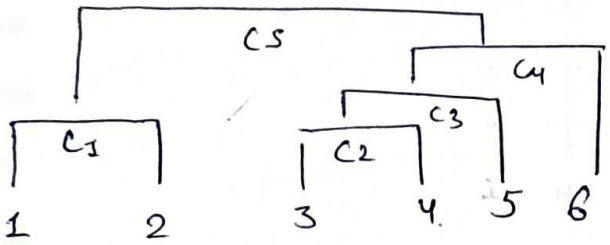


	C_1	C_3	P_6
C_1	0		
C_3	2	0	
P_6	6	1.8	0

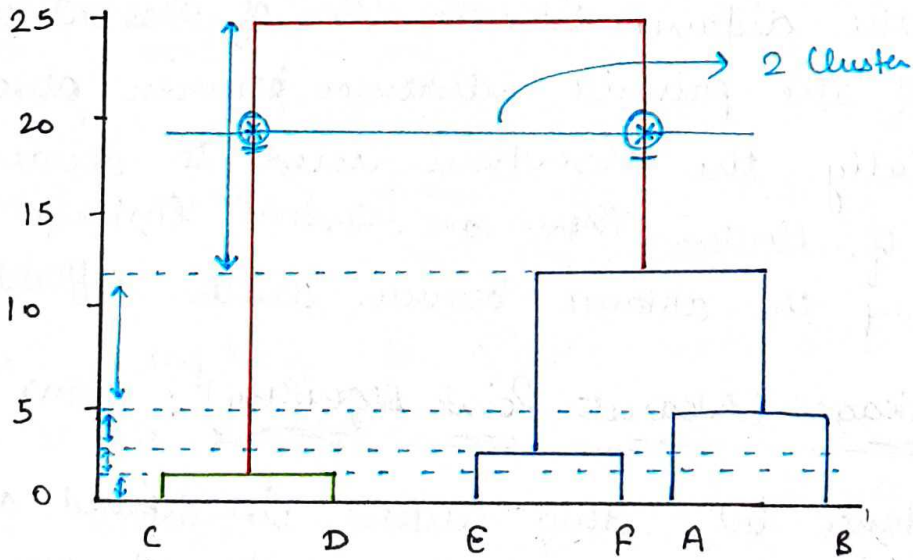


	C_1	C_4
C_1	0	
C_4	2	0



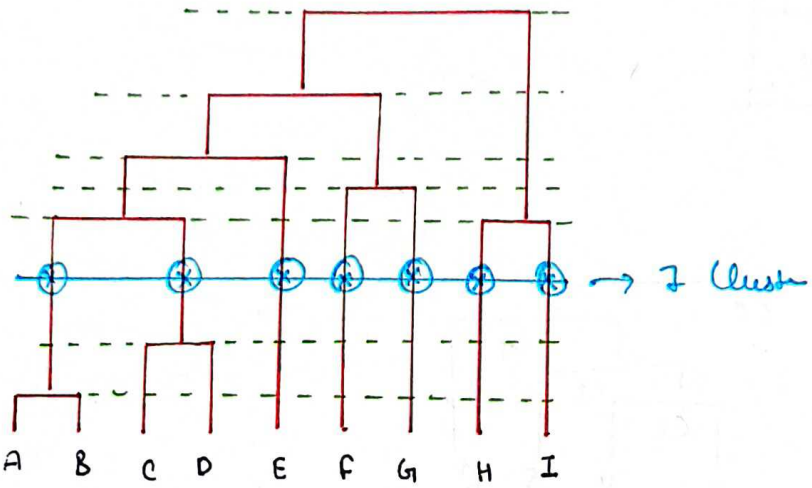


Finding n Cluster



dendrogram

Find the biggest distance and intercept the point is the number of Cluster.



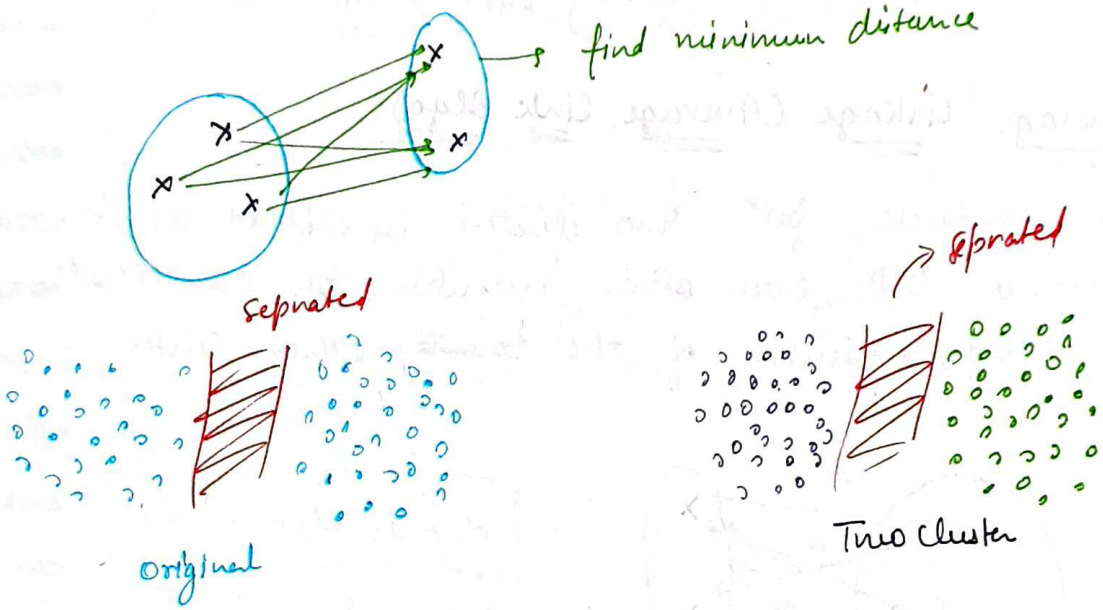
Linkage

In hierarchical clustering, linkage is the criterion that determines the distance between sets of observations as a function of the pairwise distances between observations. It's essentially the algorithm used to decide the proximity of clusters. There are several linkage methods, each defining the distance between clusters differently:

1. Single Linkage (Nearest Point Algorithm) : (Min)

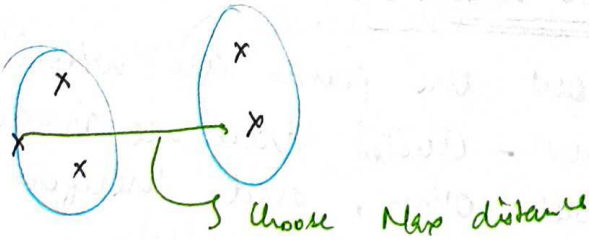
- The distance betⁿ two clusters is defined as the shortest distances from any member of one cluster to any member of the other cluster.
- Capable of detecting non-elliptical shapes in the data.

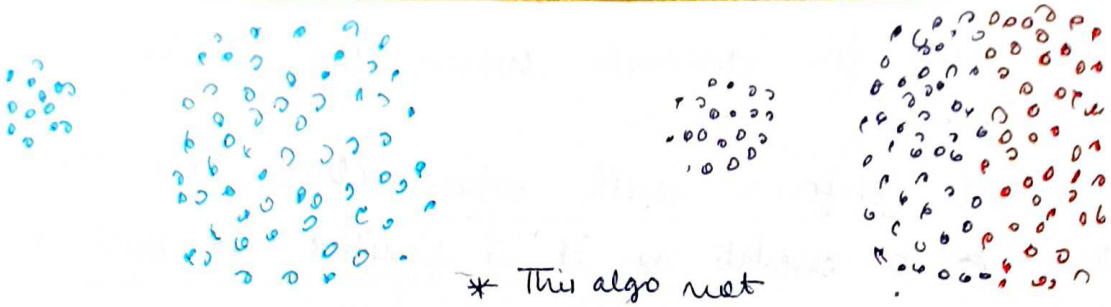
- Works well for datasets where the clusters are well-separated.
- May not perform well when clusters are close together or overlap as it is sensitive to outliers.



2. Complete Linkage (Farthest Point Algo): (Max)

- The distance betⁿ two clusters is defined as the longest distance from any member of one cluster to any member of the other cluster.
- Less susceptible to noise and outliers compared to single linkage.
- Can struggle with elongated cluster or non-convex shapes.

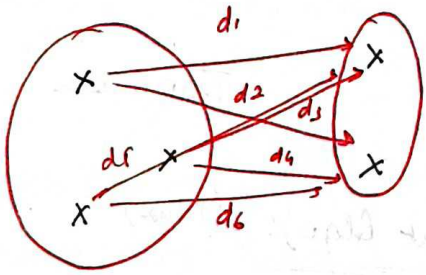




* This algo not working with unequal size of cluster

3. Average Linkage (Average link Algo):

- The distance betn two clusters is defined as the avg distance betn each ~~other~~ member of one cluster to every member of the ~~country~~ other cluster.



$$\frac{d_1 + d_2 + d_3 + \dots d_6}{6}$$

↳ proximity

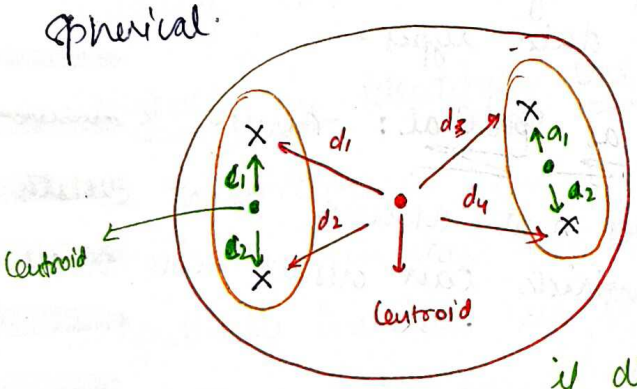
4. Ward's Method:

- **Objective:** The main goal of ward's method is to find the pair of cluster that, when merged, will increase the total within-cluster variance as possible. This is like trying to keep the clusters as compact as possible.

- **Within-Cluster Variance:** This is a measure of how spread out the points are within a cluster. A lower within-cluster variance means the points are closer to each other, and therefore, the cluster is more compact.

- **How it works:** At each step of the algo, Ward's method looks at all possible pair of clusters and all calculated how much the within-cluster variance would increase if those two cluster were merged. It then merges the two cluster that result in the smallest increase in variance.

- **Resulting Clusters:** Because Ward's method tries to keep the within-cluster variance low, it tends to create clusters that are compact and roughly spherical in shape. This can be particularly effective if the natural groups in your data are also compact and spherical.



$$\left[d_1^2 + d_2^2 + d_3^2 + d_4^2 - c_1^2 - c_2^2 - a_1^2 - a_2^2 \right]$$

↑ distance

if distance → is higher → Cluster → Spread
→ is lower → Cluster → Shrink

Time Complexity

- Time Complexity → $O(n^3)$
* Big dataset take lot of time
* This algo perform good with small dataset

Space Complexity

Space Complexity → $O(N^2)$

This algo perform good with small dataset

Advantages and Disadvantages

Advantages

1) **Discovery of Hierarchical Structure:** The algo reveals the hierarchical and nested structure within the data, which can be informative for understanding complex relationships.

2) **Useful for any Distance Measure:** The method can be used with any distance measure, which is beneficial for different types of data, such as genomic data or mixed data types.

3) **Does Not Assume Cluster as Spherical:** Unlike k-means, agglomerative clustering does not assume that clusters are spherical in shape, which can result in more natural cluster shapes.

4) **Easy to Implement and understand:** The algorithm is conceptually simple and can be easily implemented making it accessible for users with varying levels of expertise.

5) **Robust to Noise and Outliers:** With the appropriate choice of linkage criteria (such as Ward's method), hierarchical clustering can be relatively robust to noise and outliers, as these will typically be merged into cluster at later stages of the process.

Disadvantage

1. **Computational Complexity:** One of the biggest drawbacks is its computational cost. The algo has a time complexity of $O(n^3)$ and space complexity of $O(n^2)$ for the simplest implementations, making it impratical for large dataset.

2. **Sensitivity to Noise and Outliers:** Certain linkage criteria, such as single linkage, can be highly sensitive to noise and outliers, which can lead to misleading results. Outliers can cause clusters to merge permaturely, distorting the true structure of the data.

3. **Difficulty in Identifying the Number of Clusters:** While the dendrogram can provide insights into the potential number of clusters, there is often subjectivity involved in interpreting where to 'cut' the dendrogram to define the clusters.

4. **Arbitrary Decision in Linkage Criteria:** The choice of linkage criteria (single, complete, average, ward etc.) can significantly affect the results and there is no definite rule for choosing the best method, which can make the process somewhat arbitary.

5. **No Global Objective Function:** Unlike K-means, which minimizes within-cluster variance, there's no clear global objective in hirearchical clustering which can make it difficult to asses the quality of the resulting clusters.