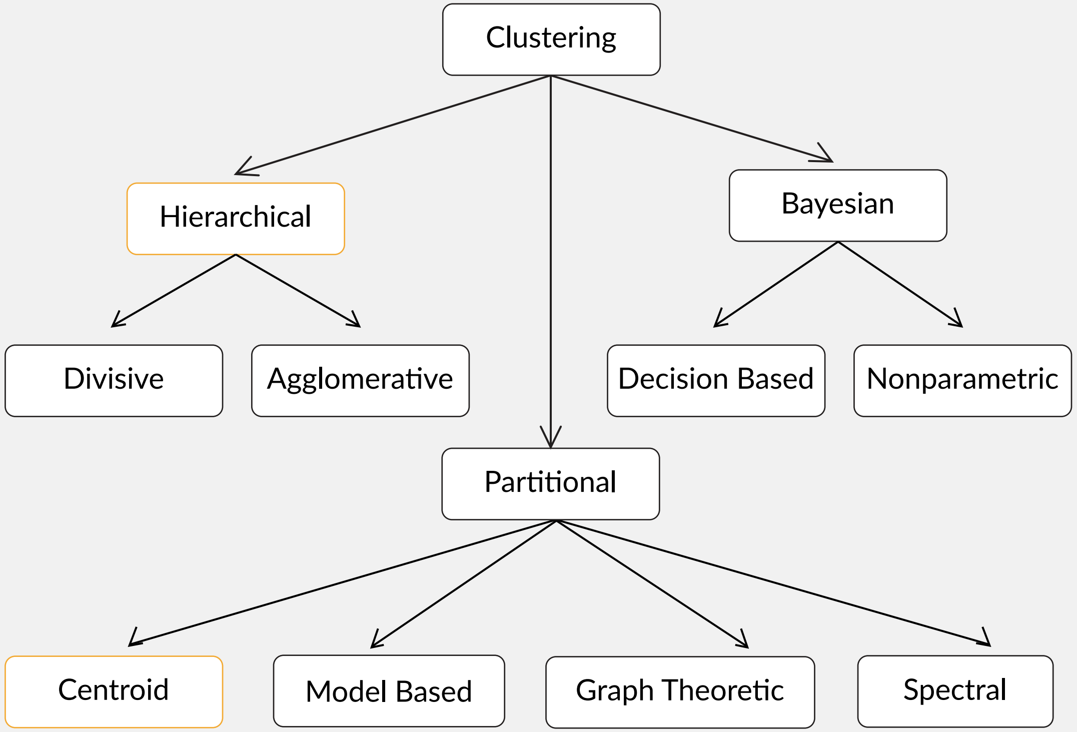
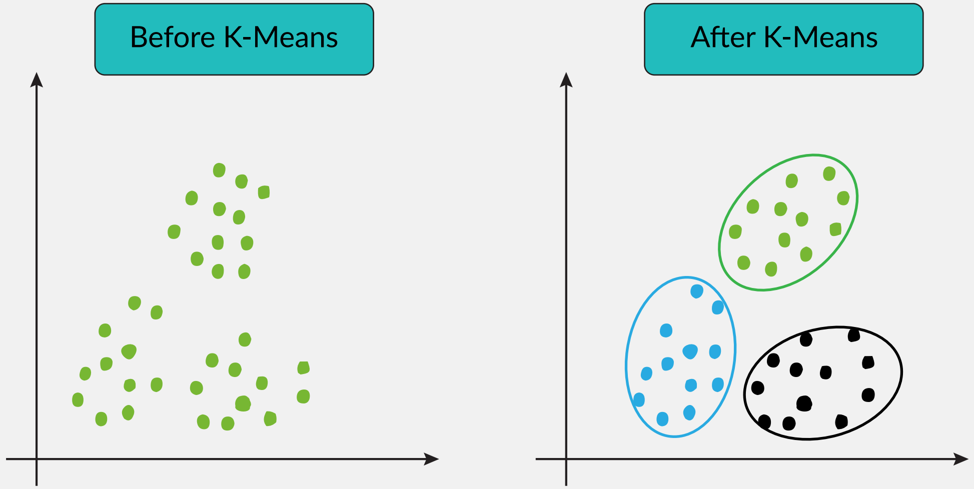
Clustering (Unsupervised Learning)

- Clustering techniques group data points into various clusters or segments based on some features available in a given data; since we use unlabelled data for clustering, it is a type of unsupervised learning.
- These techniques are commonly used for classification problems.
- These algorithms can be divided into various types based on the technique used (refer to the image on the right). K-Means and hierarchical clustering are the most commonly used techniques.
- The clustering algorithm requires data to be in a standard form; hence, data preprocessing and standardization are recommended.
- These algorithms are widely used for customer segmentation, medical imaging, and pattern recognition.



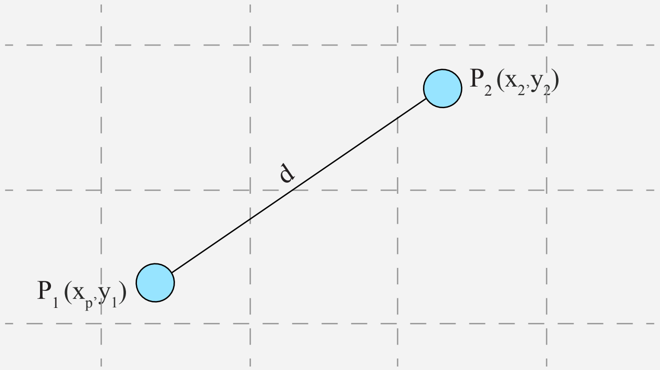


Common Interview Questions:

1. Explain the K-Means clustering algorithm.
2. Explain the hierarchical clustering algorithm.
3. Are all clustering methods type of unsupervised learning?
4. How to decide the optimal number of clusters in a data set?
5. Do all clustering algorithms use centroid?
6. List some applications of clustering algorithms.
7. Explain the concept of “elbow point” in clustering.
8. What is the difference between soft and hard clustering?
9. Compare hierarchical clustering and K-Means clustering.
10. In which situations should K-Means clustering not be used?
11. How does dimensionality impact the use of clustering algorithms?

Clustering (Unsupervised Learning)

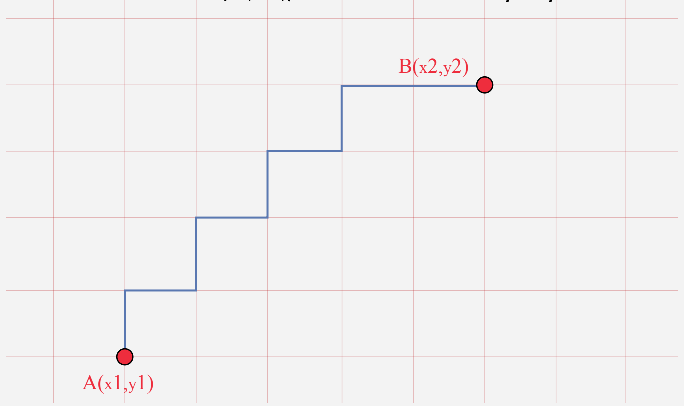
Euclidean distance



$$\text{Euclidean distance (d)} = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

Manhattan distance

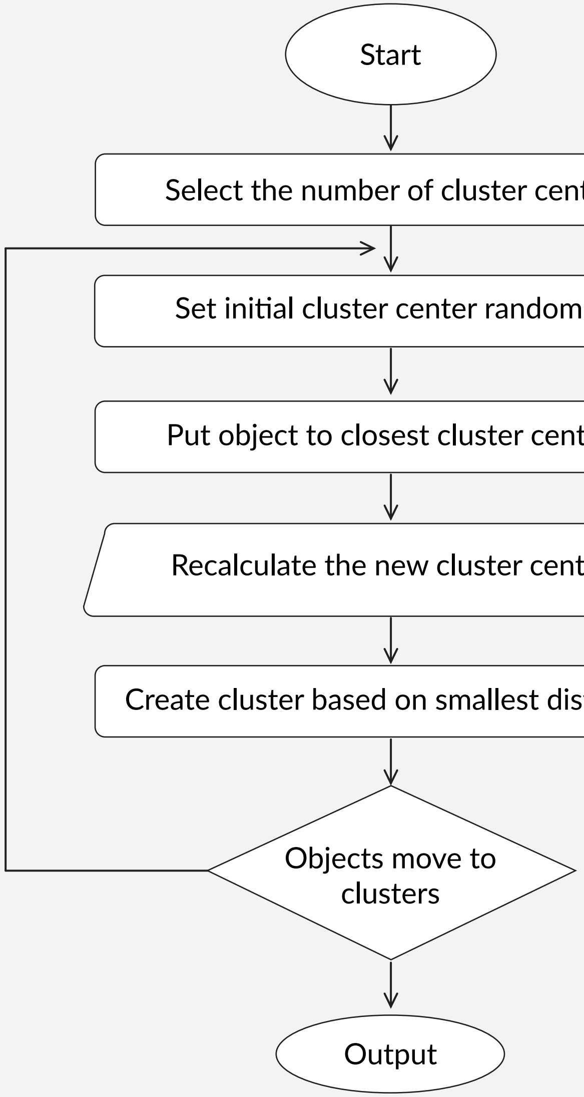
Manhattan(A, B,) = I x1-x2 I + Iy1-y2I



K - Means Clustering

- K-Means clustering is the most commonly used clustering algorithm.
- It uses a centroid to decide the optimum clusters. A centroid is a geometric center of a cluster (mean of the coordinates of all the cluster points).

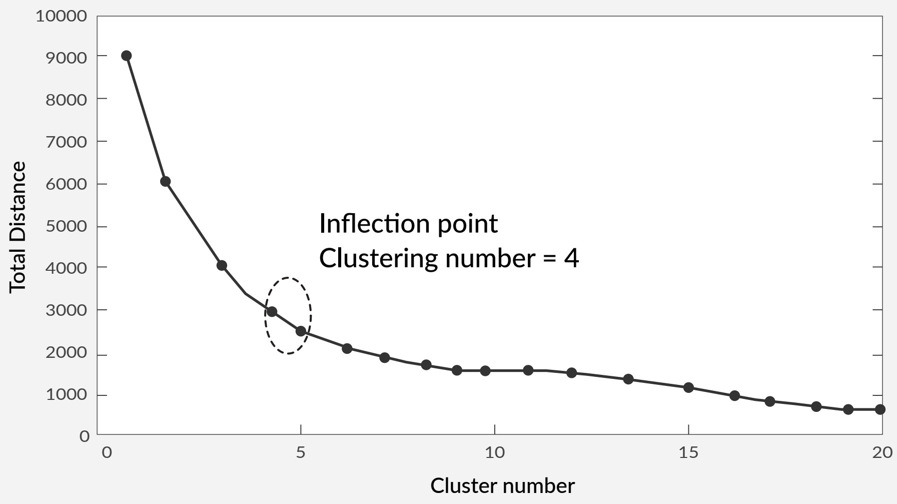
- The algorithm cannot decide the “number of clusters.”, We need to determine that, and once we give that number to the algorithm, it creates the optimum clusters.
- The optimum number of clusters can be found using the elbow point method.


Start
Select the number of cluster center
Set initial cluster center randomly
Put object to closest cluster center
Recalculate the new cluster center
Create cluster based on smallest distance
Objects move to clusters
Output

Clustering (Unsupervised Learning)

Elbow Point

- This technique is used to decide the optimum number of clusters for a data set.
- The elbow method is a graphical representation of finding the optimal “K value” in K-Means clustering. It works by finding the Within-Cluster Sum of Square (WCSS), i.e., the sum of the square distance between points in a cluster and the cluster centroid.
- All the possible numbers of clusters are created, and then, the sum of the distance is plotted against it.



Inflection point
Clustering number = 4

- The point of a sudden change in the slope is considered to be the “number of clusters.”

Code

```
sklearn.cluster.KMeans (n_clusters).fit(X)
#where, n_clusters = number of cluster required
```

Pros of K- Means Clustering

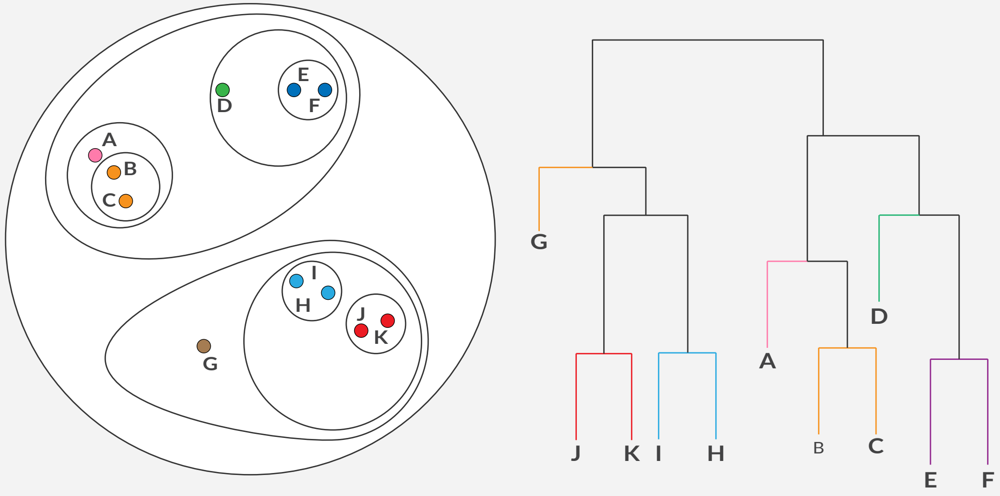
- It is simple, fast, and intuitive.
- TIt works well with large databases as well.

Cons of K- Means Clustering

- Information about the number of clusters is required.
- It does not work well with a large number of features.
- It works well only for convex clusters.

Dendrogram

- A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar data sets
- It shows the relations as sub-branch, which are easier to interpret.
- Here is an example of the same relationship in the set and dendrogram format.

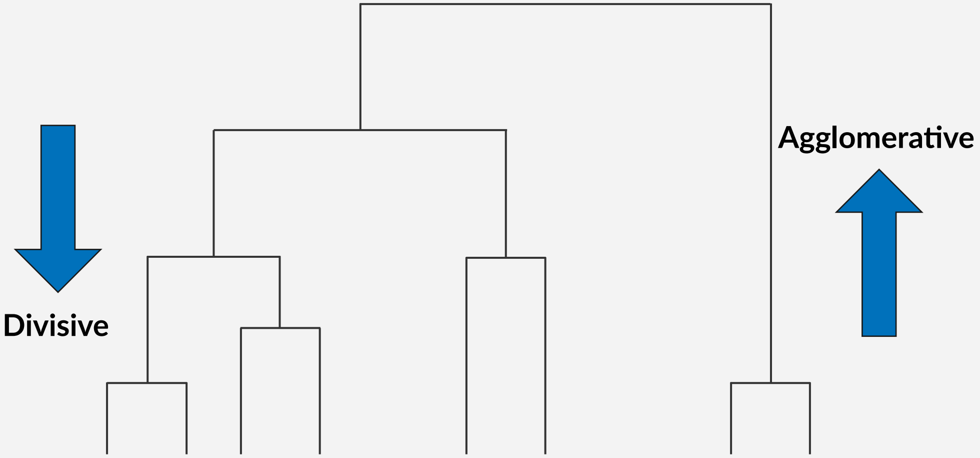


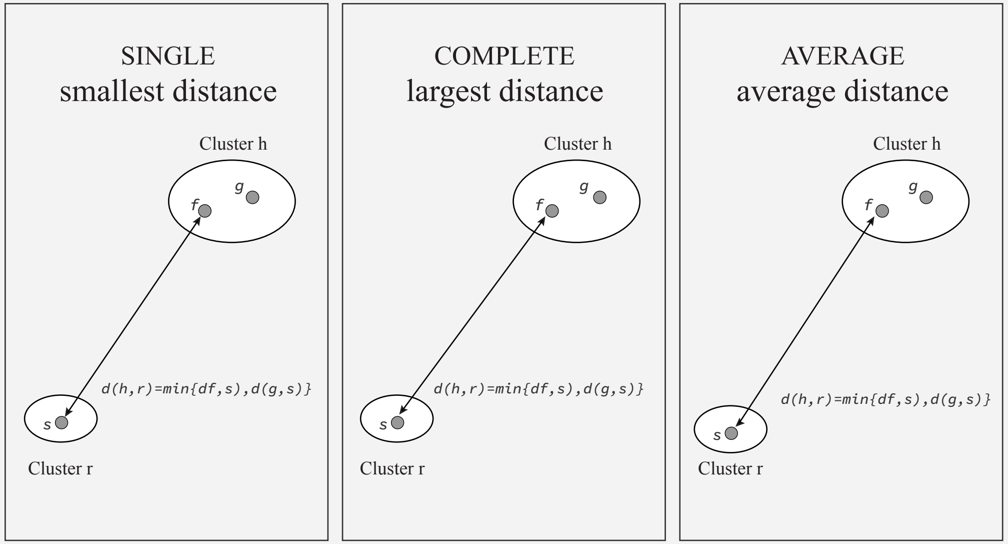
Classification Using Decision Trees

Hierarchical Clustering

- It builds a hierarchy of clusters.
- It is a family of algorithms (broadly divided into devisive and agglomerative algorithms).
- Various metrics are used to decide the optimum clusters, for example, simple linkage, centroid linkage, complete linkage, and average linkage. Ward's distance or Ward's linkage is the most commonly used metric.

The two types of hierarchical clustering are divisive (top-down) and agglomerative (bottom-up).





Pros of Hierarchical Clustering

- Knowledge about the number of clusters is not necessary.
- It can work well if the input data set has a hierarchical structure.

Cons of Hierarchical Clustering

- It uses the greedy algorithm, which might not always give the best results.
- It works well mostly on convex and homogeneous clusters.