Data Science with Python : Cluster Analysis #5850
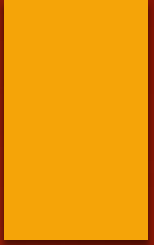
NAME: DEEPTHI M

BATCH NUMBER: 05

SERIAL NUMBER: 172

- Cluster analysis or clustering is a unsupervised machine learning model.
- It is used to group the homogeneous groups together or similar groups together.
- Homogeneity or similarity is measured by distance measures.
- It is of two types:
- Hierarchical clustering
- Non-hierarchical clustering.
- Algorithm which follows a hierarchy is a hierarchical clustering and the algorithm which does not follow a hierarchy is a non-hierarchical clustering.
- Hierarchical clustering further classified into agglomerative clustering and divisive clustering.
- Non-hierarchical clustering further classified into k-means, k-medians, and many more.

Agglomerative Clustering:
• it is a Bottom-up approach.
• In this algorithm, first data points are grouped separately and merged into a single cluster iteratively based on similarity.
• In this algorithm, Distance used to measure the similarity between data points.

Divisive Clustering:
• It is a Top-down approach.
• In this algorithm, First data points are grouped into a single cluster and separated into several clusters iteratively based on similarity.
• In this algorithm, Distance used to measure the similarity between data points.
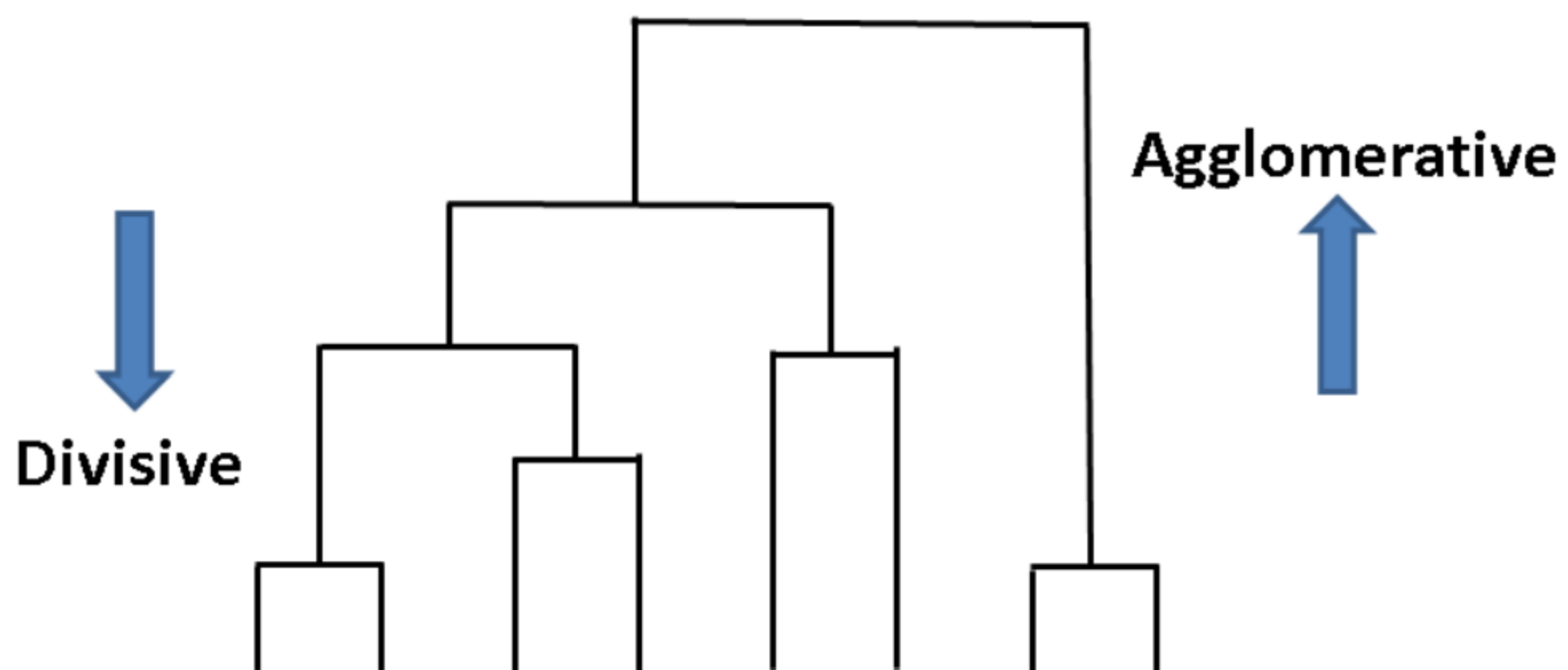
What is k-means clustering?

- It is a non-hierarchical clustering.
- It does not follow any hierarchy.
- It is used to group the homogeneous data points.

How do k-means clustering algorithm works?

- It forms the clusters on the number of clusters we pass while building the model. It randomly chooses centroid and forms a cluster by grouping the nearest data points. It is a non-deterministic model which changes on every execution.
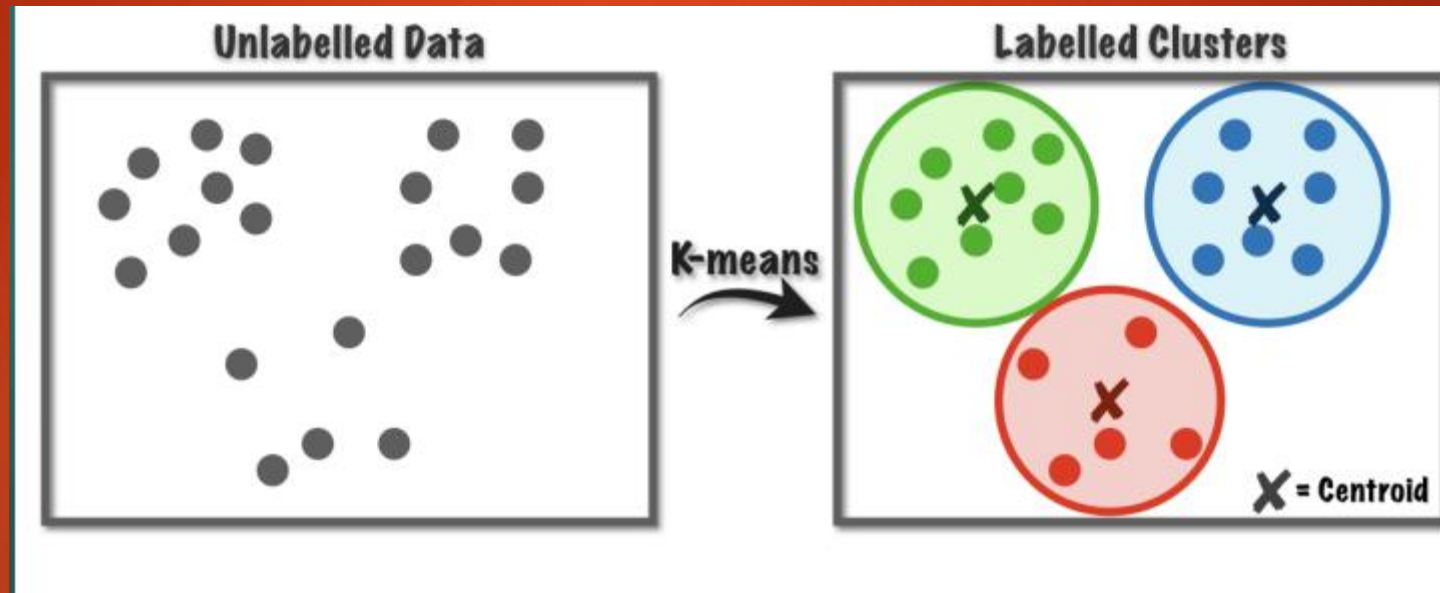
k-medians clustering

- It is a cluster analysis model.
- It chooses a median on the number of clusters we pass and then group the data points which are nearest to the meadian. This has the effect of minimizing error over all clusters

Advantages:
- Clustering has the dendogram visualization.
- It provides Hierarchical relations between clusters.
- It is used to group the similar groups.

Disadvantages:
- It is not easy to implement
- K-means is more prone to outliers
- Non-hierarchical clustering changes on every execution.
- It requires human interpretation for the analysis of the output.