Data Science with Python : Clustering Techniques #373

NAME: DEEPTHI M

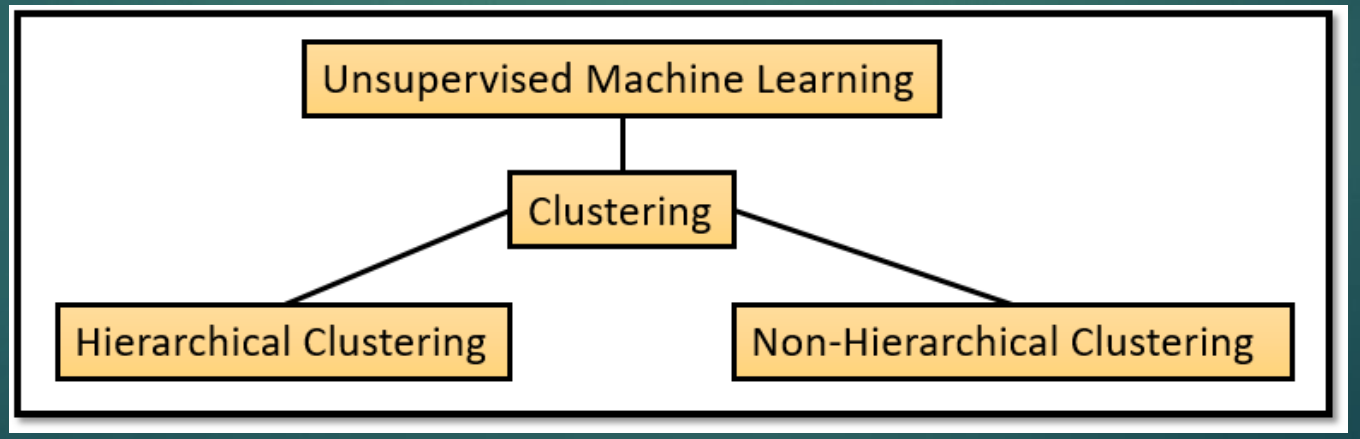
SERIAL NUMBER: 172

BATCH NUMBER: 05

Clustering Techniques

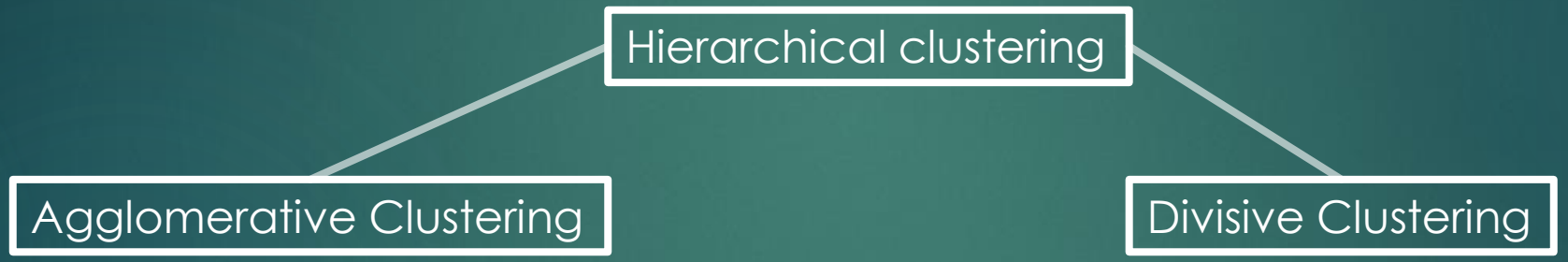
Clustering is an unsupervised machine learning model which is used to group the data based on their similarities or differences. It is broadly used to group the homogeneous data points. Similarities and differences between the data points can be measured by the distance measures.

Clustering can be categorized into hierarchical clustering and non-hierarchical clustering. Hierarchical clustering is further classified into agglomerative clustering and divisive clustering. Non-hierarchical clustering is further classified into k-means, k-modes, k-medians, k-medoids, clarans.



Hierarchical clustering

- Hierarchical clustering follows a hierarchy.
- It can be categorized into two types: agglomerative clustering and divisive clustering

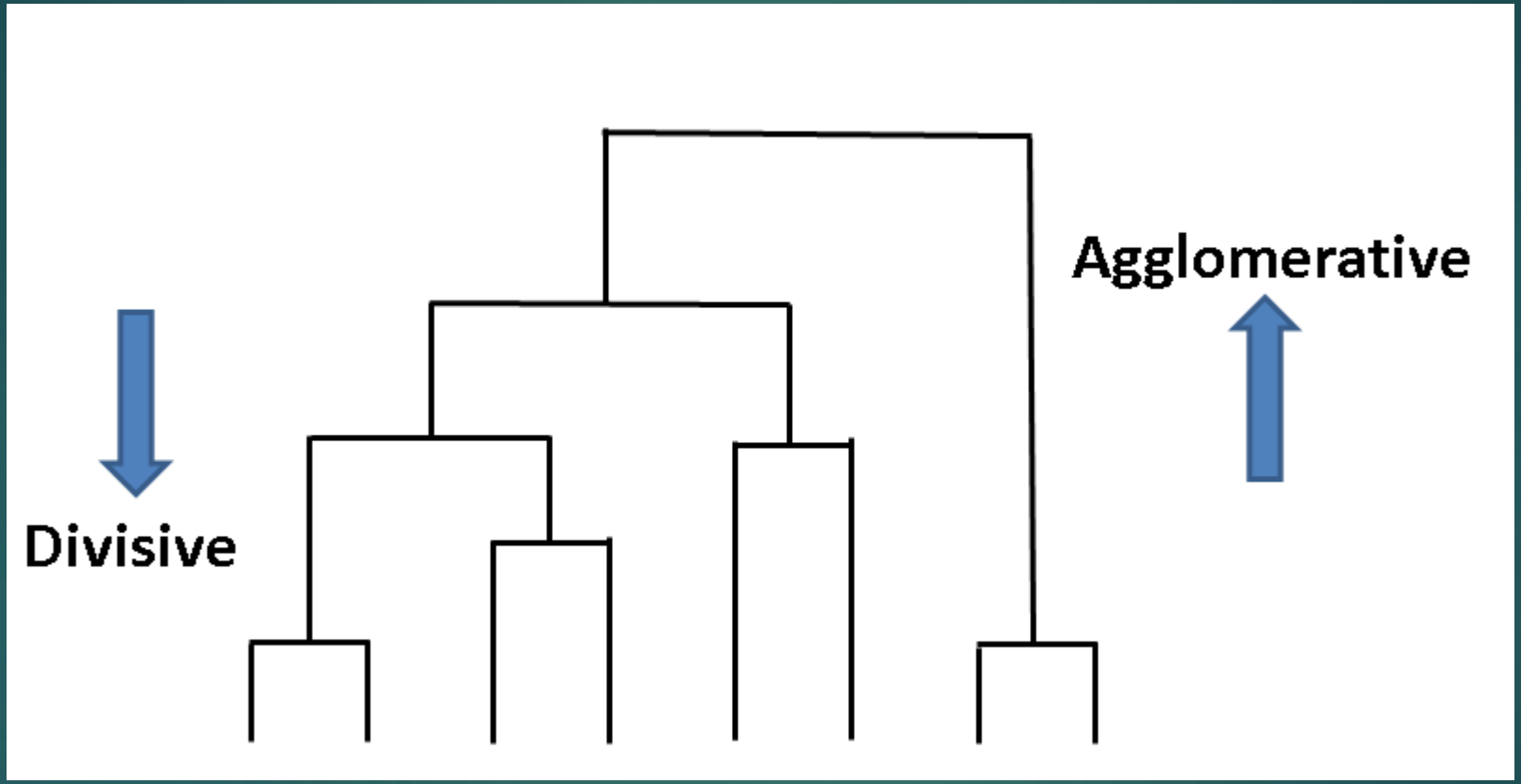


Agglomerative Clustering:

- Bottom-up approach.
- First data points are grouped separately and merged into a single cluster iteratively based on similarity.
- Distance used to measure the similarity between data points.

Divisive Clustering:

- Top-down approach.
- First data points are grouped into a single cluster and separated into several clusters iteratively based on similarity.
- Distance used to measure the similarity between data points.


Agglomerative
Divisive

What is k-means clustering?

K-means clustering is a non-hierarchical clustering. It is non-hierarchical because it does not follow any hierarchy.

K-means clustering used to group the homogeneous data points.

How do k-means clustering algorithm works?

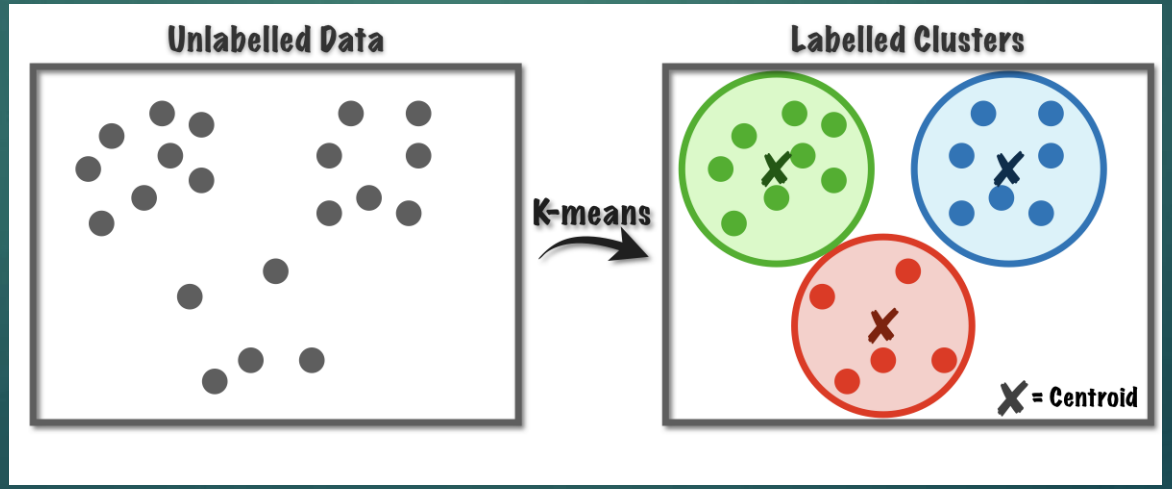
K-means clustering usually forms the clusters based on the number of clusters we pass While building the model.

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.

Selection of k in k-mean clustering:

- Randomly assigned
- Odd number of k is chosen
- Large number of k is not preferred as it forms large number of cluster which might lose the homogeneity nature of clusters.
- Too small is not chosen as it has more prone to outliers.



k-medians clustering

***k*-medians clustering** is a cluster analysis algorithm. It is a variation of *k*-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters

Advantages:

- Dendogram are great for visualization.
- Provides Hierarchical relations between clusters.
- Shown to be able to capture concentric clusters

Disadvantages:

- Not easy to implement
- Experiment showed that other clustering techniques outperforms Hierarchical clustering.
- K-means is more prone to outliers
- Non-hierarchical clustering changes on every execution.