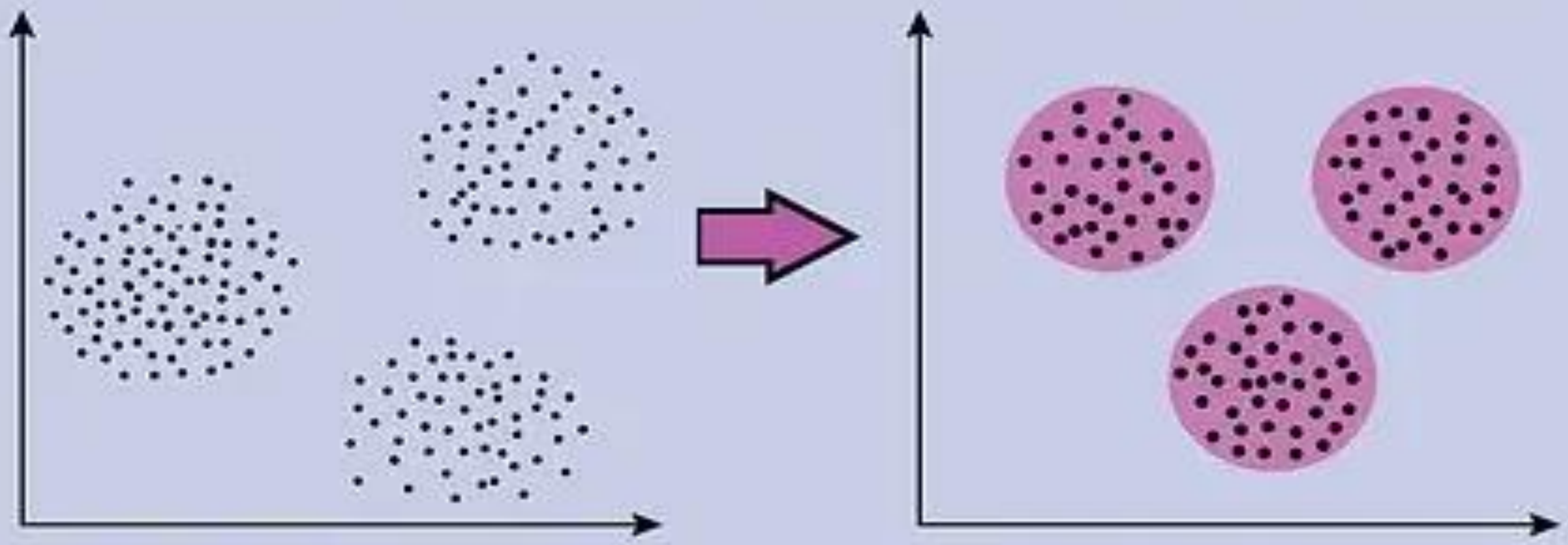
Clustering Algorithms in Machine Learning



ProjectPro

What is Clustering ?

The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method is defined under the branch of Unsupervised Learning, which aims at gaining insights from unlabelled data points, that is, unlike supervised learning we don’t have a target variable.

Clustering aims at forming groups of homogeneous data points from a heterogeneous dataset. It evaluates the similarity based on a metric like Euclidean distance, Cosine similarity, Manhattan distance, etc.

Types of Clustering Algorithms

Clustering is the process of determining how related the objects are based on a metric called the similarity measure.
Similarity metrics are easier to locate in smaller sets of features. It gets harder to create similarity measures as the number of features increases. Depending on the type of clustering algorithm being utilized in data mining, several techniques are employed to group the data from the datasets.

What is the best clustering method?

The **Top 10** clustering algorithms are:

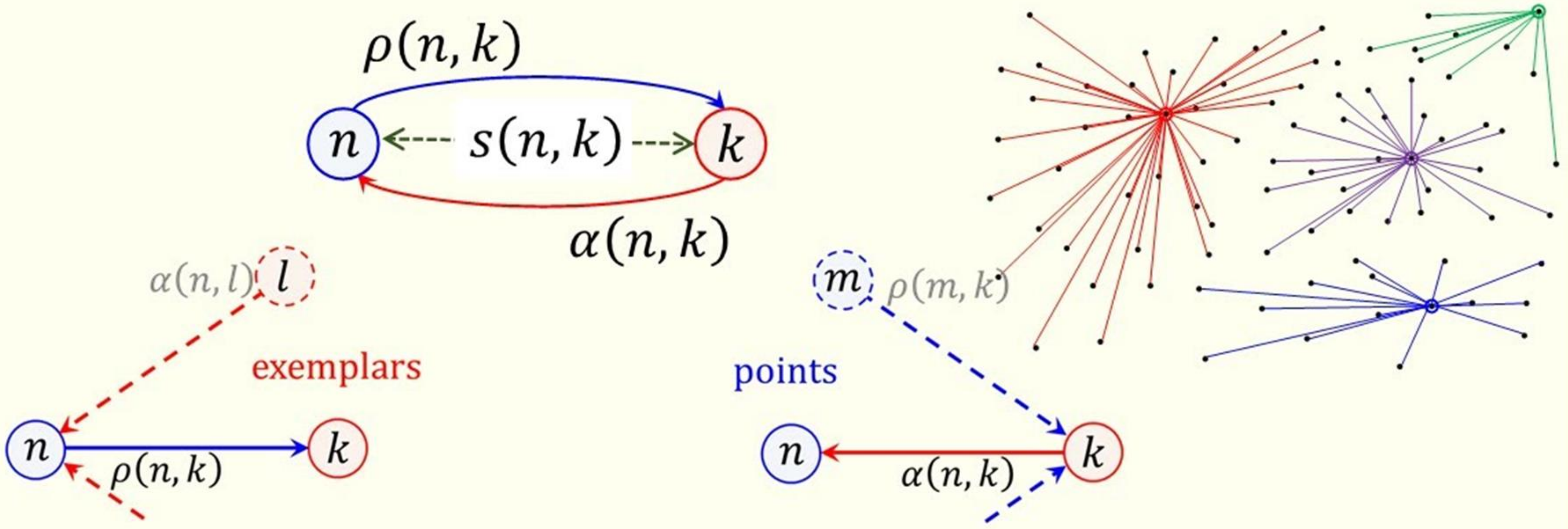
- K-means Clustering
- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Gaussian Mixture Models (GMM)
- Agglomerative Clustering
- Spectral Clustering
- Mean Shift Clustering
- Affinity Propagation
- OPTICS (Ordering Points To Identify the Clustering Structure)
- Birch (Balanced Iterative Reducing and Clustering using Hierarchies)

In the above **Top 10** Algorithms, We discuss few of them.

Affinity Propagation Clustering Algorithm

- Initialize each data point as a potential exemplar
- Measure similarity between data points Euclidean distance, similarity measures etc.
- Iteratively update responsibiliry and availability metrices.
- Excemplar selection
- Cluster asssignments
- Convergence check and finalization

1



responsibility $n \to k$

$$\rho(n,k) \leftarrow s(n,k) - \max_{l \neq k}\big(\alpha(n,l) + s(n,l)\big)$$

availability $k \to n$

$k \neq n$ (cross-availabilities)

$$\alpha(n,k) \leftarrow \min\left\{0, \rho(k,k) + \sum_{m \neq n,k} \max\big(0, \rho(m,k)\big)\right\}$$

$k = n$ (self-availabilities)

$$\alpha(k,k) \leftarrow \sum_{n \neq k} \max\big(0, \rho(n,k)\big)$$

Advantages, Disadvanges & Applications of Affinity Propagation

Applications

Social Network Analysis
Marketing Fields(Analysing)
Medical Image fields
Speech Recognition

Advantages

- Automatic cluster determination
- Discovery of exemplars
- Handles non linear relationships

Disadvantages

- Computationally expensive
- In High dimensional datas Its not suitable

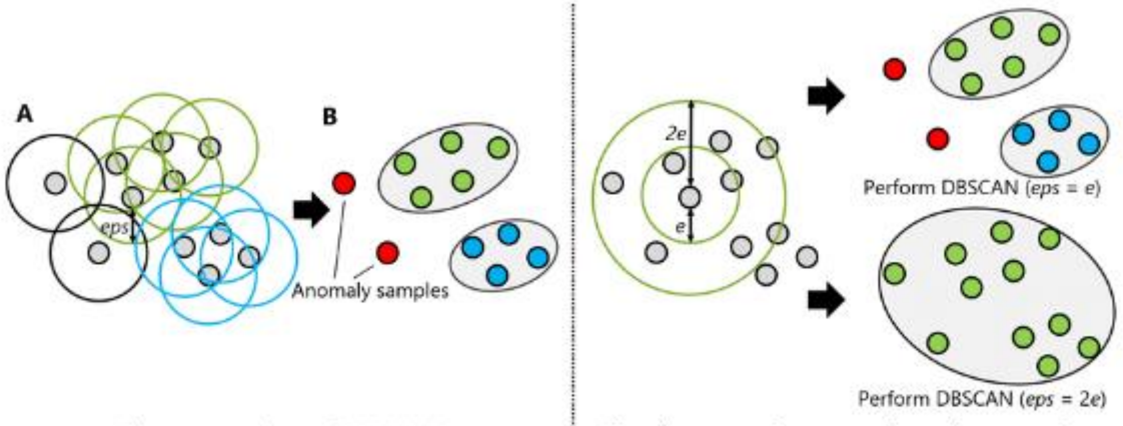
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- Start with an unvisited point
- Identify nearby points within a certain distance (**Epsilon**)
- Mark it as core point if enough neighbours(minpts)
- Expand clusters
- Include density reachable points
- Detect Noise and label it (not enough neighbours)

Applications

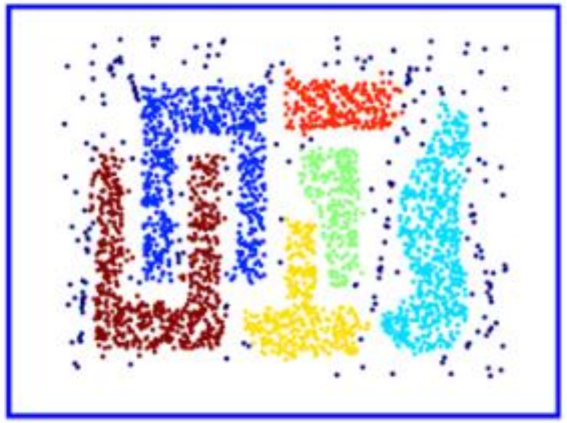
- Anomaly Detection
- Image segmentation
- Fraud Detection
- Customer Segmentation

DBSCAN identifies clusters based on the density of data points in their vicinity. It can discover clusters of arbitrary shapes and is robust to noise and outliers, making it suitable for datasets with varying cluster densities.



The processing of DBSCAN

The impact of *eps* on clustering results



DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Advantages

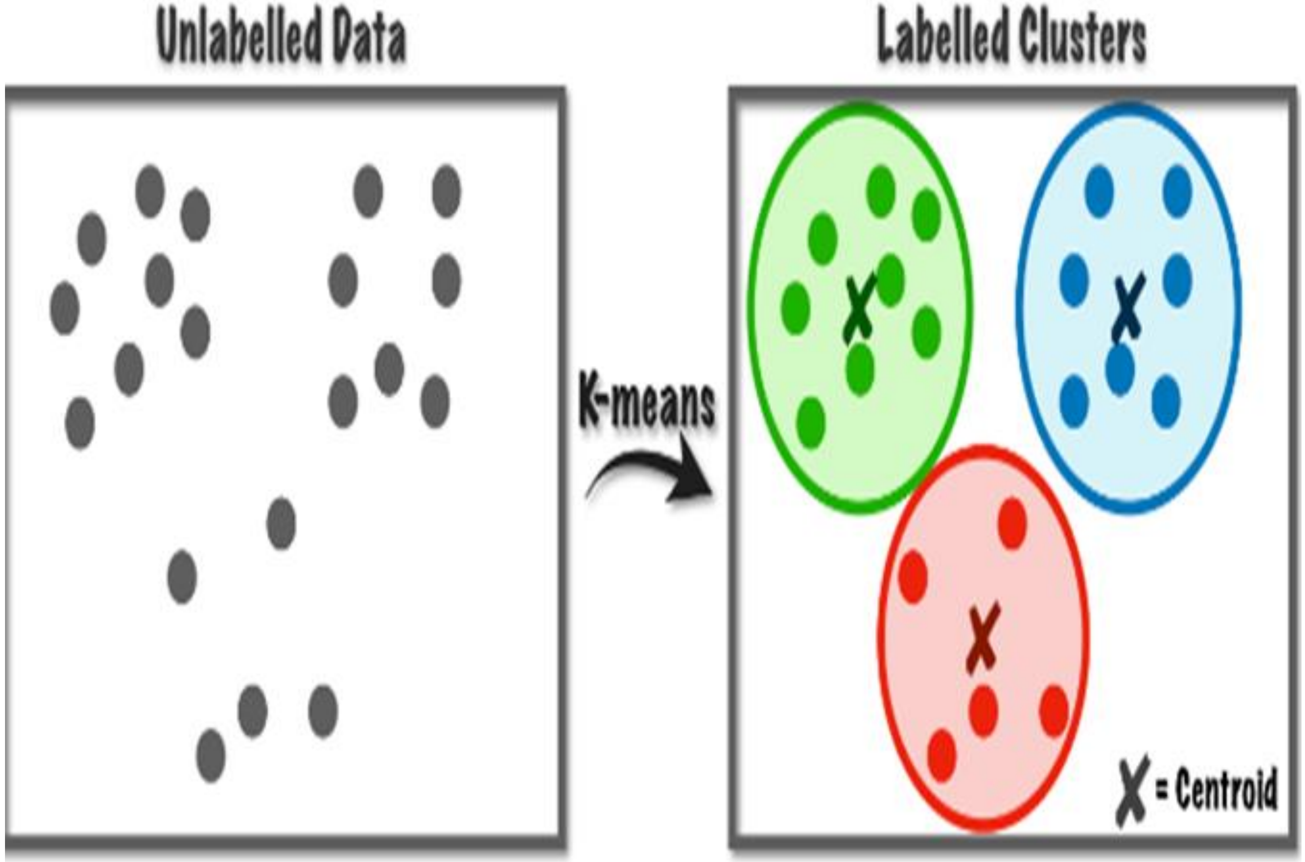
- Robust to Noise
- Identify arbitrary shapes
- Efficient

Disadvantages

- Parameters sensitivity
- Struggles in identifying clusters of different densities

K-Means Clustering

K-Means is one of the most popular clustering algorithms. It aims to partition the data into a predefined number of clusters (K) by iteratively adjusting cluster centroids to minimize the distance between data points and their assigned centroids. It is efficient and works well when clusters are roughly spherical and evenly sized.



Applications of K-Means Clustering

K-Means clustering is used in a variety of examples or business cases in real life, like:

1. Academic performance
2. Diagnostic systems
3. Search engines
4. Wireless sensor networks

Advantages

- Simple and easy to implement
- Fast and efficient
- Scalability
- Flexibility

Disadvantages

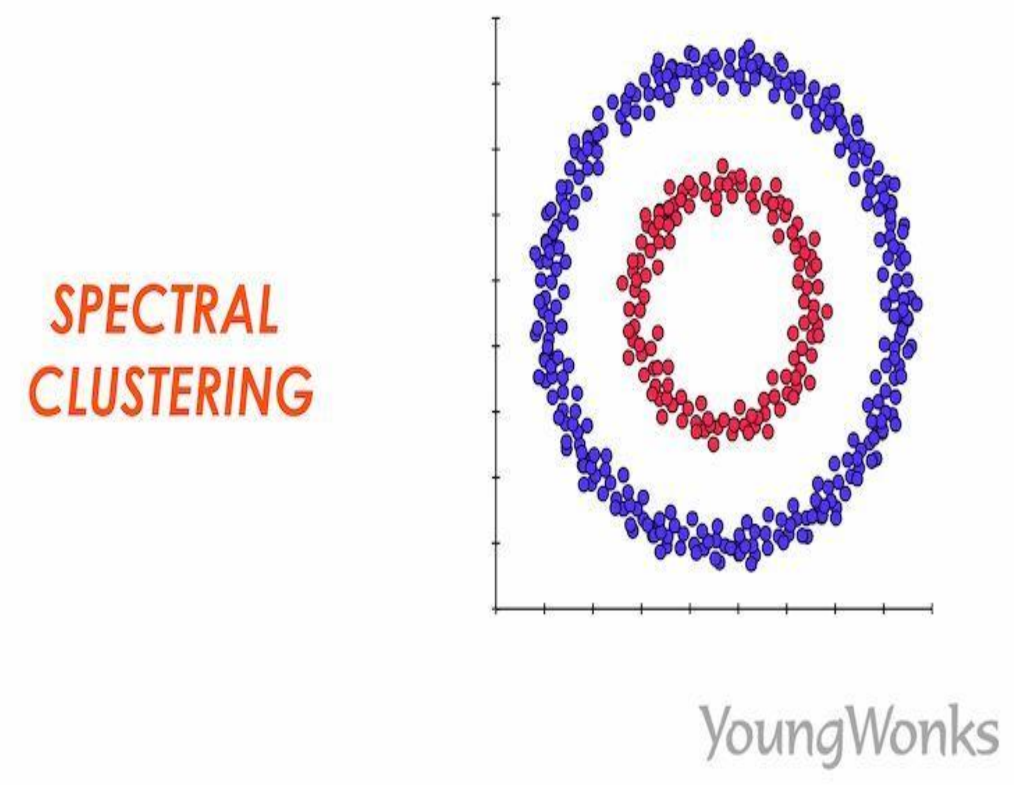
- Sensitivity to initial centroids
- Requires specifying the number of clusters
- Sensitive to outliers

Spectral Clustering

- Create similarity **matrix**
- **Graph** construction
- Compute **Laplacian** matrix

L=D(diagonal matrix)-W(adjacency matrix)

- Compute **eigenvectors** and **eigenvalues**
- Cluster assignments



Spectral Clustering Applications

1. Image segmentation
2. Document clustering
3. Anomaly Detection
4. Graph-based clustering

Advantages

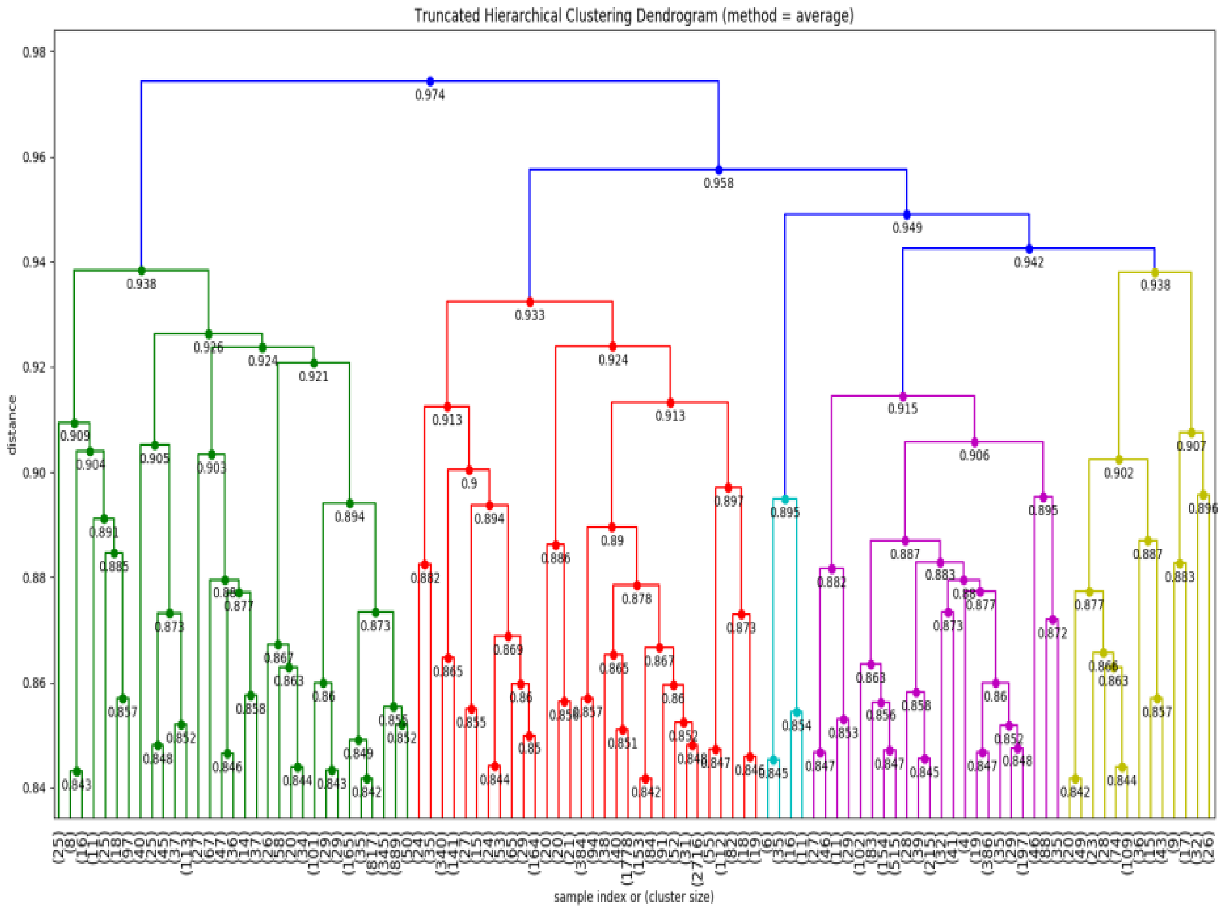
- No Assumptions in the shape or size
- Handles Non-linear data

Disadvantages

- Computaionally expensive

Hierarchical Clustering

Hierarchical clustering builds a tree-like structure of clusters, where each data point starts as its own cluster and then progressively merges with other clusters based on their similarity. This results in a dendrogram, a visual representation of the clustering hierarchy. Hierarchical clustering is particularly useful when the number of clusters is unknown.



Significance of Clustering

Clustering plays a pivotal role in data analysis for several reasons:

1. Insights from Unlabeled Data:

Unlike supervised learning, clustering can provide insights from unlabeled data, making it valuable for exploratory analysis and hypothesis generation.

2. Decision-Making:

Clusters can guide decision-making by revealing patterns that can inform strategic choices and resource allocation.

3. Dimensionality Reduction:

By reducing the data to clusters, dimensionality is effectively reduced, simplifying subsequent analysis tasks.

4. Foundation for Further Analysis:

Clustered data serves as a foundation for more advanced analysis techniques, such as classification and anomaly detection.

Conclusion

In conclusion, clustering is a powerful technique in the realm of Machine Learning that enables the identification of patterns, relationships, and structures within complex datasets. Its versatility and broad range of applications make it an indispensable tool for various industries, facilitating informed decision-making, personalized experiences, and deeper insights into the underlying data. As data continues to grow in complexity, clustering remains a fundamental approach to unlocking its hidden treasures.

ThankYou