UNIT – V

Clustering: Introduction to Clustering, Partitioning of Data, Matrix Factorization Clustering of Patterns, Divisive Clustering, Agglomerative Clustering, Partitional Clustering, K-Means Clustering, Soft Partitioning, Soft Clustering, Fuzzy C-Means Clustering, Rough Clustering, Rough K-Means Clustering Algorithm, Expectation Maximization-Based Clustering, Spectral Clustering.

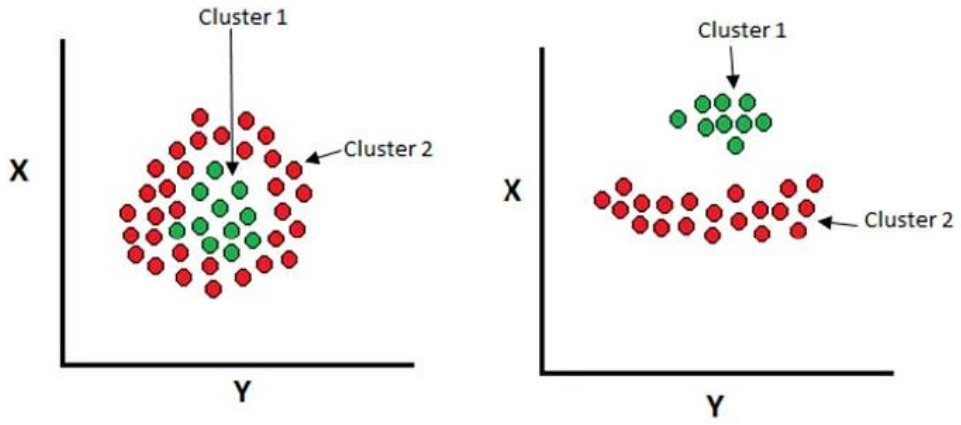
1. Introduction to Clustering

- Clustering is an unsupervised learning technique that groups a set of objects such that objects in the same group (cluster) are more similar to each other than to those in other groups.
- Clustering is a technique in machine learning that involves grouping similar data points together based on their characteristics or features.
- The goal of clustering is to identify patterns and similarities in data sets without any prior knowledge of the groups or categories that exist within the data.
- In clustering, the algorithm tries to find similarities between data points and creates groups, or clusters, of similar data points.
- The similarity between data points is based on a set of features or attributes that describe each data point.
- The algorithm aims to minimize the distance between the data points within each cluster while maximizing the distance between different clusters.
- Clustering has various applications, including customer segmentation, image segmentation, anomaly detection, and recommendation systems.
- The process of grouping objects based on similarities is called clustering. Described as an unsupervised learning problem with the objective of producing training data using a specific set of inputs but without any predetermined goal values.
- In order to make a collection of unlabelled data more comprehensible and manipulable, it is the process of looking for comparable structural features. A cluster is a group of data points that are connected to one another through their connections to nearby data points. Two uses for clustering are pattern analysis and feature engineering.

Types of clustering algorithms

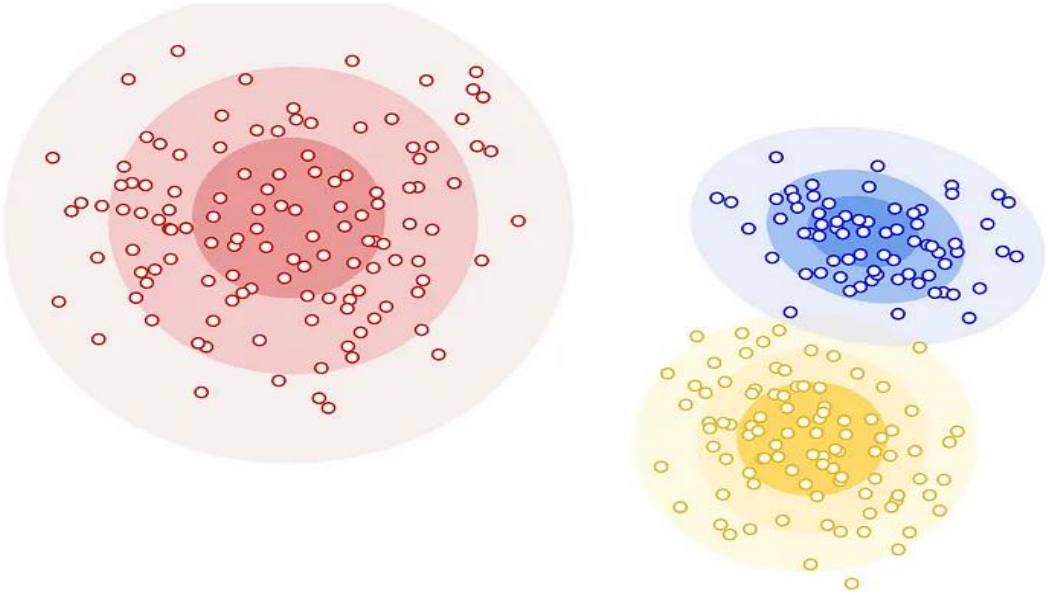
Different clustering techniques exist that can handle various sorts of data.

Density-based



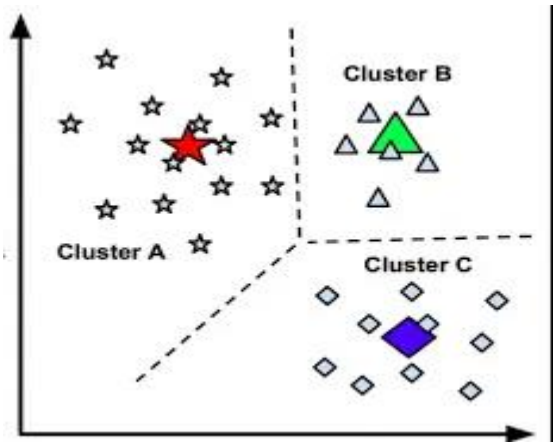
Data is organised into clusters that have high data point densities surrounded by low data point densities. In essence, the algorithm identifies areas with a high density of data points and designates those areas as clusters.

Distribution-based



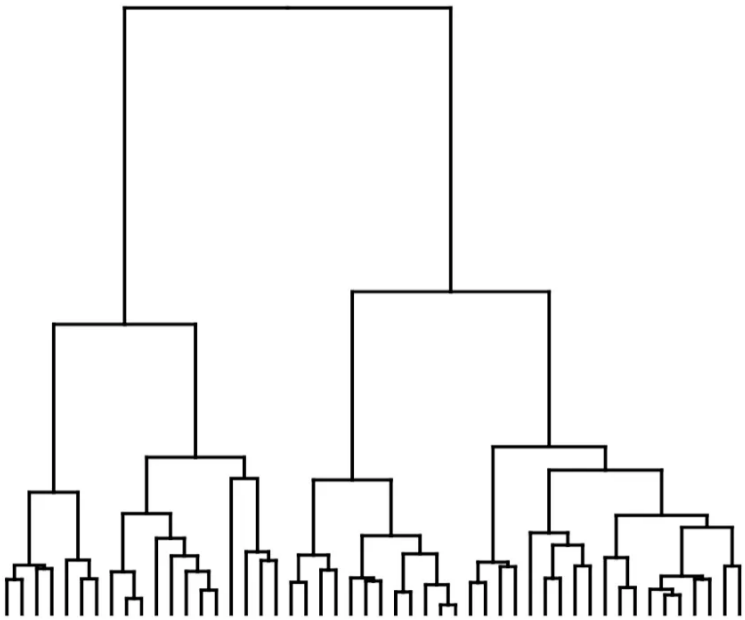
Based on the likelihood that each data point belongs to a certain cluster, all of the data points are regarded as components of that cluster.

Centroid-based



These methods divide data points depending on several centroids present in the data. In accordance with its squared distance from the centroid, each data point is grouped into a cluster. The most popular kind of clustering is this one.

Hierarchical-based



Everything is arranged top-down by creating a tree of clusters.

Popular clustering algorithms

- K-means clustering algorithm
- Gaussian Mixture Model algorithm
- DBSCAN clustering algorithm
- BIRCH algorithm
- Mean-Shift clustering algorithm
- OPTICS algorithm

Advantages of Clustering

Clustering brings several advantages to the field of data analysis:

1. Pattern Discovery:

Clustering helps uncover hidden patterns and structures within data. By grouping similar data points together, it becomes easier to identify trends and relationships that might not be apparent when examining individual data points.

2. Data Summarization:

Large datasets can be overwhelming to analyze. Clustering allows data scientists to summarize complex datasets into a smaller number of representative clusters, making it simpler to understand and interpret the data.

3. Anomaly Detection:

Outliers and anomalies can have a significant impact on analysis results. Clustering can help detect such anomalies by identifying data points that do not fit well into any cluster.

4. Customer Segmentation:

In marketing, clustering can be used to segment customers into groups based on their purchasing behavior, preferences, and demographics. This information is invaluable for targeted marketing strategies.

Applications of Clustering

The applications of clustering span various industries and domains:

1. Marketing:

Clustering assists in market segmentation, allowing companies to tailor their marketing strategies to specific customer groups, resulting in more effective campaigns and personalized experiences.

2. Healthcare:

In medical research, clustering can be used to group patients based on similar symptoms or genetic traits, aiding in disease diagnosis and treatment customization.

3. Image Segmentation:

In computer vision, clustering helps segment images into distinct regions, enabling object recognition and scene understanding in fields such as autonomous vehicles and medical imaging.

4. Natural Language Processing (NLP):

In NLP, clustering can be applied to group similar documents or words, assisting in topic modeling, sentiment analysis, and text summarization.

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:

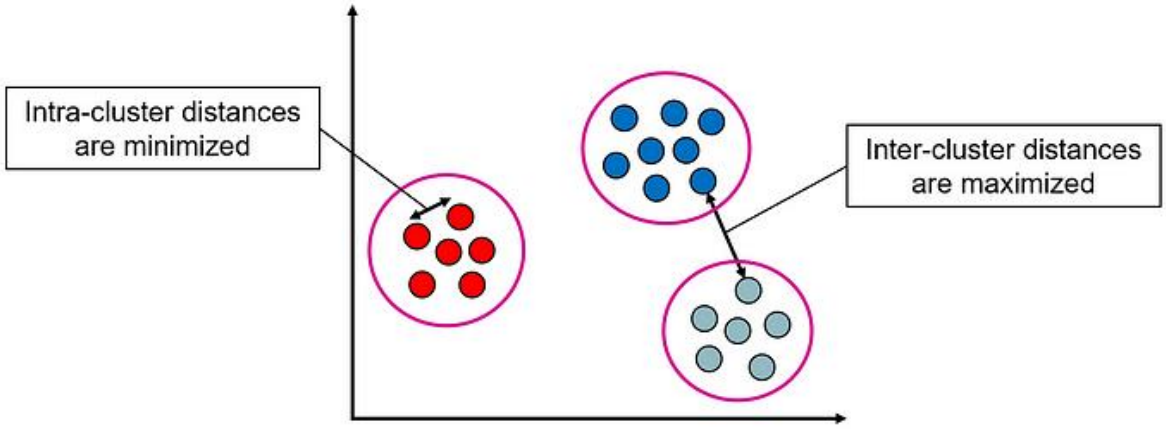
By reducing the data to clusters, dimensionality is effectively reduced, simplifying subsequent analysis tasks.

4. Foundation for Further Analysis:

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

1. **Intra-cluster distances** are minimized: data points within the same cluster are as close as possible to one another.
2. **Inter-cluster distances** are maximized: data points in different clusters are as far as possible from one another.

The entire set of clusters $\{C_1, \ldots, C_k\}$ is referred to as a **clustering**.



Objectives of clustering (image by author)

Divisive clustering

Divisive clustering is a hierarchical clustering method that involves dividing every cluster into smaller subsets, starting with each object in a single cluster, until the desired number of clusters is achieved.

Hierarchical clustering is a popular unsupervised machine learning technique used to group similar data points into clusters based on their similarity or dissimilarity. It is called "hierarchical" because it creates a tree-like structure of clusters known as a **dendrogram** where each node represents a cluster that can be divided into smaller sub-clusters

There are two types of hierarchical clustering techniques:

1. Agglomerative (Bottom-up approach)
2. Divisive clustering (Top-down approach)

Difference between agglomerative clustering and Divisive clustering:

Parameters	**Agglomerative Clustering**	**Divisive Clustering**
Approach	Bottom-up: Starts with individual points and merges them.	Top-down: Starts with all data in one cluster and splits.
Complexity Level	More computationally expensive due	Less computationally

Parameters	Agglomerative Clustering	Divisive Clustering
	to pairwise distance calculations.	expensive but requires careful cluster splitting.
Handling Outliers	Better at handling outliers, as outliers can be absorbed into larger clusters.	Outliers may lead to inefficient splitting and suboptimal results.
Interpretability	More interpretable due to clear cluster merging in the dendrogram.	Can be harder to interpret due to recursive splitting decisions.
Implementation	Scikit-learn provides multiple linkage methods such as “ward,” “complete,” “average,” and “single.”	Not widely implemented in major libraries like Scikit-learn and SciPy.
Example Applications	Image segmentation, customer segmentation, document clustering, etc.	Less common but can be used in hierarchical data analysis.

Feature	Agglomerative	Divisive
Starting point	Individual points	One large cluster
Process	Merges clusters	Splits clusters
Best for	Small to medium datasets	Large datasets
Outlier handling	Better	Can create separate clusters
Interpretability	More intuitive	Can be challenging

Key points:

- Both create a tree-like structure (dendrogram) showing data relationships
- Choice depends on data size, structure, and analysis goals
- Agglomerative is more common and often easier to interpret
- Divisive can be faster for large datasets

References

https://utsavdesai26.medium.com/the-beginners-guide-to-clustering-in-machine-learning-331987a7ceaf
https://www.linkedin.com/pulse/clustering-machine-learning-types-advantages-applications/
https://medium.com/@sainikhilesh/brief-introduction-to-clustering-and-different-methods-of-clustering-f9d6ec80907c
https://www.scaler.com/topics/data-mining-tutorial/partitioning-methods-in-data-mining/
https://medium.com/ai-made-simple/introduction-to-clustering-2ffc22673b5a
https://builtin.com/machine-learning/agglomerative-clustering