Clustering

Clustering is an unsupervised machine learning technique that involves grouping similar data points together. It's a way to discover hidden patterns or structures within unlabeled data.

Types of Clustering Algorithms

a. Partitioning Clustering:

- Divides data into non-overlapping partitions.

- Examples: K-means, K-medoids

b. Hierarchical Clustering:

- Creates a hierarchical structure of clusters, either agglomerative (bottom-up) or divisive (top-down).

- Examples: Agglomerative Hierarchical Clustering, Divisive Hierarchical Clustering

c. Density-Based Clustering:

- Groups data points based on density in regions of space.

- Examples: DBSCAN, OPTICS

d. Model-Based Clustering:

- Assumes a probabilistic model for the data and clusters based on the model's parameters.

- Examples: Gaussian Mixture Models

- Gaussian Mixture Models: Assumes data is generated from a mixture of Gaussian distributions.

Hard clustering and soft clustering

1. Hard Clustering:

Hard clustering, also known as crisp clustering, assigns each data point to exactly one cluster.

Examples of hard clustering algorithms:

- K-Means
- Hierarchical Clustering
- DBSCAN

2. Soft Clustering:

Soft clustering, also known as fuzzy clustering, assigns each data point a probability or degree   of belonging to each cluster.

Examples of soft clustering algorithms:

- Fuzzy C-Means
- Gaussian Mixture Models
- Soft K-Means

Applications

- Customer segmentation in marketing
- Image segmentation
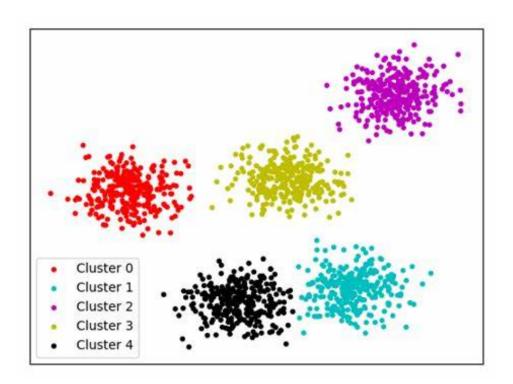- Anomaly detection
- Document clustering
- Biological taxonomy

Challenges

- Determining the optimal number of clusters

- Handling high-dimensional data

- Dealing with outliers and noise

- Interpreting and validating cluster results

Evaluation Metrics

- Silhouette Score

- Calinski-Harabasz Index

- Davies-Bouldin Index