Clustering Algorithms

Introduction

Clustering algorithms are unsupervised learning techniques that group similar data points into clusters based on their features. The goal is to ensure that data points within the same cluster are more similar to each other than to those in other clusters.

What is Clustering?

Clustering is the process of dividing a dataset into groups such that data points in the same group share similar characteristics.

```
Data Points → Similarity Measure → Clusters
```

Types of Clustering

1. Partition-based clustering
2. Hierarchical clustering
3. Density-based clustering
4. Model-based clustering

1. K-Means Clustering

K-Means is a partition-based clustering algorithm that divides data into K clusters by minimizing the distance between data points and their cluster centroids.

1. Choose number of clusters K
2. Assign points to nearest centroid
3. Update centroids
4. Repeat until convergence

Example: Customer segmentation.

```
Data → Assign to Nearest Mean → Update Mean → Clusters
```

2. Hierarchical Clustering

Hierarchical clustering builds a tree-like structure of clusters called a dendrogram. It does not require specifying the number of clusters in advance.

Types: Agglomerative and Divisive.

```
Individual Points → Merge → Tree Structure (Dendrogram)
```

3. DBSCAN

DBSCAN groups points that are closely packed together and identifies noise or outliers as points that lie alone in low-density regions.

1. Density-based approach
2. Does not require number of clusters
3. Detects outliers

Example: Anomaly detection.

Clustering Evaluation Metrics

1. Silhouette Score
2. Davies–Bouldin Index
3. Inertia (within-cluster sum of squares)

Advantages

1. No labeled data required
2. Useful for data exploration
3. Helps discover hidden patterns

Disadvantages

1. Choosing optimal clusters can be difficult
2. Sensitive to noise and outliers
3. Results may vary with algorithm choice

Real-Life Example

Streaming platforms use clustering algorithms to group users with similar interests to recommend movies and shows.

Summary

Clustering algorithms are powerful tools for grouping data without labeled outputs. They are widely used in customer segmentation, anomaly detection, and recommendation systems.