DBSCAN

Introduction

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised learning algorithm used for density-based clustering. Unlike K-Means, DBSCAN does not require specifying the number of clusters in advance and can discover arbitrarily shaped clusters while identifying noise. DBSCAN groups points that are closely packed together while marking points that lie alone in low-density regions as outliers. It is particularly effective when the data contains clusters of varying shapes and sizes.

Core Concepts and Definitions

- ε (epsilon): Radius parameter defining a neighborhood around a point.
- MinPts: Minimum number of points required to form a dense region.
- Core Point: A point with at least MinPts neighbors within ε.
- Border Point: A point within ε of a core point with fewer than MinPts neighbors.
- Noise Point (Outlier): A point neither a core nor a border point.

Algorithm Steps

1. For each unvisited point:
    - Mark it as visited.
    - Retrieve its ε-neighborhood.
2. If the point is a core point:
    - Create a new cluster and recursively include all density-reachable points.
3. If the point is not a core point and not density-reachable from another core point, mark it as noise.
4. Continue until all points have been visited.

Mathematical Notation

Let $D(x, y)$ be the distance between points x and y. The ε-neighborhood of point x is defined as:

$$N_\varepsilon(x) = \{y \in X \mid D(x, y) \leq \varepsilon\} \tag{1}$$

A point x is a core point if:

$$|N_\varepsilon(x)| \geq \text{MinPts} \tag{2}$$

Key Characteristics

- Type: Unsupervised density-based clustering algorithm.
- No need to specify the number of clusters.
- Distance Metric: Typically uses Euclidean distance but can be generalized.
- Robust to Noise: Naturally identifies outliers.

Strengths

- Can discover clusters of arbitrary shape and size.
- Automatically identifies outliers.
- Requires minimal domain knowledge to set parameters.

Weaknesses

- Sensitive to the choice of ε and MinPts.
- Struggles with clusters of varying density.
- Performance degrades in high-dimensional spaces.

Applications of DBSCAN

- Geospatial Data Analysis: Identifying geographic clusters or regions.
- Anomaly Detection: Detecting outliers in network traffic or financial transactions.
- Image Processing: Grouping pixels or image segments.
- Astronomy: Clustering stars or celestial objects.

Conclusion

DBSCAN is a robust clustering algorithm that is well-suited for discovering complex structures in data and identifying noise. Its ability to detect non-convex clusters without requiring the number of clusters as input makes it a valuable tool in exploratory data analysis.