DBSCAN

Density-Based Spatial Clustering of Applications with Noise

What is DBSCAN?

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm that groups together points that are closely packed together while marking points in low-density regions as outliers or noise.

Key Concept: Unlike K-Means which requires you to specify the number of clusters beforehand, DBSCAN discovers clusters based on the density of data points. It can find arbitrarily shaped clusters and automatically identifies outliers.

The Big Idea

DBSCAN operates on a simple but powerful principle: **clusters are dense regions in the data space, separated by regions of lower density**.

The algorithm identifies three types of points:

Core Points: Points that have at least a minimum number of neighbors within a specified radius.

Border Points: Points that are within the radius of a core point but don't have enough neighbors to be core points themselves.

Noise Points: Points that are neither core nor border points (outliers).

How It Works (Simple Version)

1. **Pick a random unvisited point** from your dataset.
2. **Find all points within epsilon (ε) distance** from this point.
3. **Check if there are enough neighbors**: If the number of neighbors is greater than or equal to minPts, mark this as a core point and start a new cluster.
4. **Expand the cluster** by visiting all reachable points (neighbors of neighbors) and add them to the cluster.
5. **If not enough neighbors**, mark the point as noise (it might become a border point later).
6. **Repeat** until all points have been visited.

Key Formula & Parameters

```
Two Main Parameters:

ε (epsilon): The maximum distance between two points for them to
be considered neighbors

  Distance(p, q) ≤ ε


minPts: The minimum number of points required to form a dense
region (cluster)

  |Nε(p)| ≥ minPts
```

Density Reachability: A point q is density-reachable from point p if there is a chain of points p_1, p_2, ..., p_n where $p_1 = p$ and $p_n = q$, such that each p_{i+1} is within ε distance of p_i.

Pros & Cons

Advantages

- ✓ No need to specify number of clusters beforehand
- ✓ Can find arbitrarily shaped clusters
- ✓ Robust to outliers (identifies noise)
- ✓ Works well with spatial data
- ✓ Only requires two parameters

Disadvantages

- ✗ Struggles with clusters of varying densities
- ✗ Sensitive to parameter selection (ε and minPts)
- ✗ Not suitable for high-dimensional data
- ✗ Can be computationally expensive for large datasets
- ✗ Border points can be ambiguous

When Should You Use It?

Use DBSCAN when:

- You don't know how many clusters exist in your data
- Your clusters have non-spherical shapes
- You need to identify outliers/anomalies
- Your clusters have similar density
- You're working with spatial or geographic data

Avoid DBSCAN when:

- Your data has varying densities across clusters
- You're working with very high-dimensional data
- You need a deterministic clustering result
- Your dataset is very large (consider HDBSCAN instead)

Common Uses

Anomaly Detection

Identifying fraudulent transactions, network intrusions, or manufacturing defects

Geographic Analysis

Finding crime hotspots, disease outbreak clusters, or customer location patterns

Image Segmentation

Separating objects in images based on color or texture similarity

Market Segmentation

Grouping customers with similar behaviors while identifying outlier segments

Astronomy

Identifying galaxy clusters and cosmic structures in space

Traffic Analysis

Detecting traffic congestion patterns and identifying unusual traffic flows