HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise

What is it?

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is an advanced clustering algorithm that automatically discovers clusters of varying densities in data without requiring you to specify the number of clusters beforehand.

The Big Idea

Unlike traditional clustering methods like K-means, HDBSCAN finds clusters based on density - it looks for regions where data points are tightly packed together and can identify clusters of different shapes and sizes. It's "hierarchical" because it builds a tree of possible clusterings at different density levels and intelligently selects the most stable clusters.

How It Works (Simple Version)

1. **Build a graph:** Connect each point to its k nearest neighbors
2. **Calculate distances:** Measure how far each point is from dense regions (mutual reachability distance)
3. **Create hierarchy:** Build a tree showing how clusters merge as you decrease density requirements
4. **Extract stable clusters:** Identify clusters that persist across multiple density levels
5. **Label noise:** Points that don't belong to any stable cluster are marked as noise (-1)

Key Formula

The core concept is **mutual reachability distance**:

$$d_{mreach}(a,b) = \max(\text{core\_distance}(a), \text{core\_distance}(b), d(a,b))$$

Where:

- core_distance(a) = distance to the kth nearest neighbor of point a
- d(a,b) = actual distance between points a and b

Pros & Cons

Pros

- ✓ No need to specify number of clusters
- ✓ Handles varying density clusters
- ✓ Robust to noise and outliers
- ✓ Discovers clusters of arbitrary shapes
- ✓ More stable than DBSCAN

Cons

- ✗ Computationally expensive for large datasets
- ✗ Sensitive to parameter choices (min_cluster_size)
- ✗ Less interpretable than simple methods
- ✗ May struggle with very high-dimensional data

When Should You Use It?

- When you don't know how many clusters exist
- When clusters have different densities
- When you need to identify outliers
- When clusters have irregular shapes
- When you have noisy data

Common Uses

- Anomaly detection in cybersecurity
- Customer segmentation with outlier identification
- Geographic data clustering
- Astronomical data analysis
- Gene expression analysis
- Image segmentation