



# DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm used in machine learning, particularly for identifying clusters of points in a dataset. It's especially effective when dealing with data that may not be well-suited for traditional clustering methods like K-means, such as data with irregular shapes or varying densities.

## Math Intuition:

Let's break down the key concepts behind DBSCAN:

1. **Density:** DBSCAN operates based on the concept of density. It defines two parameters:
  - $\epsilon$  (epsilon): The maximum distance between two points for them to be considered as part of the same neighborhood.
  - minPts: The minimum number of points required to form a dense region.
2. **Core Points, Border Points, and Noise Points:**
  - **Core Points:** A point is considered a core point if there are at least minPts points (including itself) within distance  $\epsilon$  of it.
  - **Border Points:** A point is considered a border point if it's not a core point but lies within the  $\epsilon$ -neighborhood of some core point.
  - **Noise Points:** Points that are neither core nor border points are considered noise points.
3. **Clustering Process:**
  - DBSCAN starts by randomly selecting a point from the dataset.
  - It then checks if this point is a core point. If it is, it forms a cluster by including all reachable points within  $\epsilon$  distance.
  - If the selected point is not a core point but a border point, it is assigned to the cluster of some core point within its  $\epsilon$ -neighborhood.
  - This process continues until all points have been assigned to clusters or labeled as noise.

## Real Example:

Let's say we have a dataset of GPS coordinates representing locations of restaurants in a city. We want to use DBSCAN to identify clusters of restaurants.

### **1. Density Analysis:**

- We define  $\epsilon$  (epsilon) to be 0.1 kilometers and minPts to be 5.
- This means that for a point to be considered a core point, it must have at least 5 other restaurants within a 0.1 km radius.

### **2. Clustering Process:**

- We start by randomly selecting a restaurant location.
- If there are at least 5 restaurants within 0.1 km of this location, it's labeled as a core point, and a cluster is formed including all reachable points within 0.1 km.
- If the selected location is not a core point but lies within 0.1 km of a core point, it's labeled as a border point and assigned to the cluster of the nearby core point.
- This process continues, forming clusters of restaurants based on their proximity and density.

### **3. Output:**

- The output of DBSCAN will be clusters of restaurants, where each cluster represents a group of restaurants that are densely located together.
- Additionally, some restaurants that are isolated from dense clusters may be labeled as noise points.