DBSCAN

Density-Based Spatial Clustering of Applications with Noise

- ✓ The **DBSCAN algorithm** is based on this intuitive notion of **clusters** and **noise**.
- ✓ Clusters are dense regions in the data space, separated by regions of the lower density of points.
- ✓ DBSCAN does not require specifying the number of clusters in advance, making it particularly useful for datasets where the number of clusters is not known beforehand or where clusters have irregular shapes.

Parameters Required For DBSCAN Algorithm

The algorithm defines two parameters:

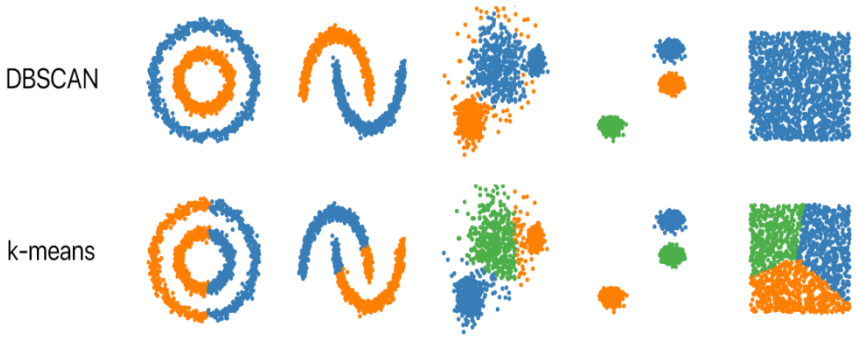
- **Epsilon (ε)**: Also known as the neighbourhood radius, ε specifies the maximum distance between two points for them to be considered as neighbours. Points within ε distance of each other are considered directly reachable.

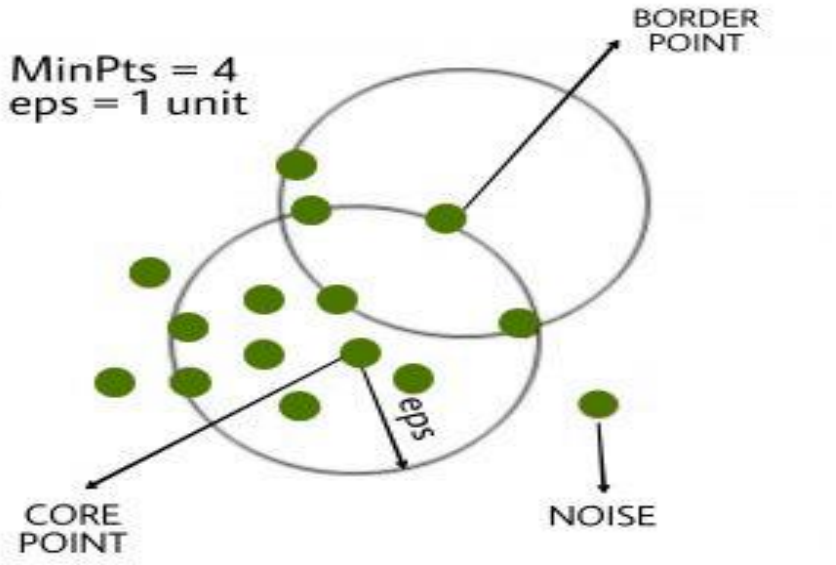
- **MinPts**: MinPts specifies the minimum number of points required to form a dense region. A point is considered a core point if it has at least MinPts points (including itself) within its ε-neighbourhood.

DBSCAN categorizes points

Using the required parameters, DBSCAN categorizes points into three categories:

- **Core points**: Points that have at least MinPts points within their ε-neighborhood.
- **Border points**: Points that are not core points but lie within the ε-neighborhood of a core point.
- **Noise points**: Points that are neither core points nor border points.





How Does DBSCAN work?

Intialization

- Choose two parameters: ε (epsilon) and MinPts.

Point Classification

- If the number of points within the ε-neighborhood (including the point itself) is greater than or equal to MinPts, classify the point as a core point.
- If the number of points within the ε-neighborhood is less than MinPts, but the point lies within the ε-neighborhood of some core point, classify it as a border point.
- If the point does not meet the criteria to be a core or border point, classify it as a noise point.

Cluster Formation

- DBSCAN iterates through each core point and identifies all points that are reachable from it within the ε-neighborhood.
- If a reachable point is also a core point, its neighbors are recursively visited and added to the same cluster.

Continue…

- If a reachable point is a border point, it is added to the cluster as well.

Cluster Expansion

- Once all points reachable from a core point have been added to the cluster, DBSCAN selects another core point that has not been assigned to any cluster and repeats the process.
- This process continues until all core points have been explored and all reachable points have been assigned to clusters.

Noise Handling

- Points classified as noise points remain unassigned to any cluster.

Output

- The output of DBSCAN is a set of clusters, where each cluster contains a group of data points that are closely packed together based on density.

Advantages, Disadvantages and Application

Advantages

- Robustness to noise
- Ability to identify arbitrary-shaped clusters
- Parameterization independence
- Examples or case studies showcasing the advantages of DBSCAN

Disadvantages

- Sensitivity to the choice of parameters
- Difficulty in clustering data of varying densities
- Scalability issues with large datasets
- Examples or case studies showcasing the disadvantages of DBSCAN

Application

- Spatial data analysis
- Image segmentation
- Anomaly detection
- Customer segmentation