DBSCAN

Vraj Patel

Density based Clustering

This algorithm divides your entire datasets into dense regions seperated by sparse regions

Examples: DBSCAN
OPTICS

DBSCAN: Density Based spatial clustering of Application with Noise.

Minpts and Epsilon

Minpts: stands for "minimum points" is a parameter that specifies the minimum number of points required to form a dense region, which is considered a cluster

Epsilon (ε) is a key parameter that defines the radius of the neighbourhood around a given data point. Specifically, ε is the maximum distance between two points for them to be considered as a part of the same neighbourhood. This parameter is crucial in determining whether points are close enough to be included in a cluster.

A point is considered a "Core point" if it has a minimum no. of other points (specified by Minpts) within a given radius ε of itself.

- A border point is defined as follows:
  Not a core point: A border point does not meet the criteria to be a core point. It has a fewer than Minpts within its ε-neighbourhood.
- Neighbour of a core point: A border point is within the ε distance of one or more core points. In other words, it lies on the edge of a cluster, within the radius ε of at least one core point.

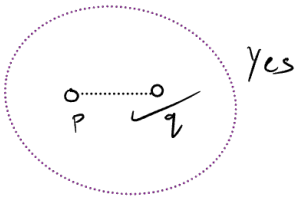
A noise point is a data point which is neither a core point nor a border point.

Density Connected Points

Directly density Reachable:

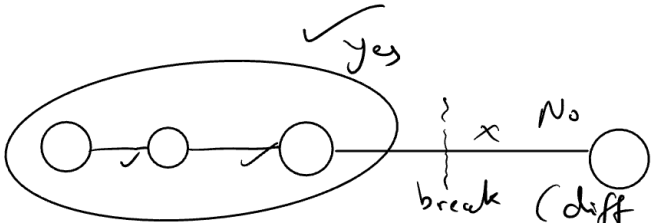
A point is directly density reachable from a point Q given Eps, MinPts if:

1. P is the Eps - neighbourhood of Q
2. Both P and Q are core points



Density Connected Points

A point P is density connected to Q, given MinPts if there is a chain of points $P_1, P_2, P_3 \dots P_n$, $P_1 = P$ and $P_n = Q$ such that P_{i+1} is directly density reachable from P_i



Intuition :

Step 1 - Identify all points as either core point, border point or noise point

Step 2 - For all of the unclustered core points

Step 2a - Create a new cluster

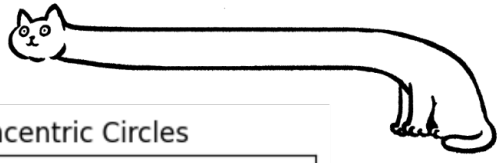
Step 2b - add all the points that are unclustered and density connected to the current point into this cluster

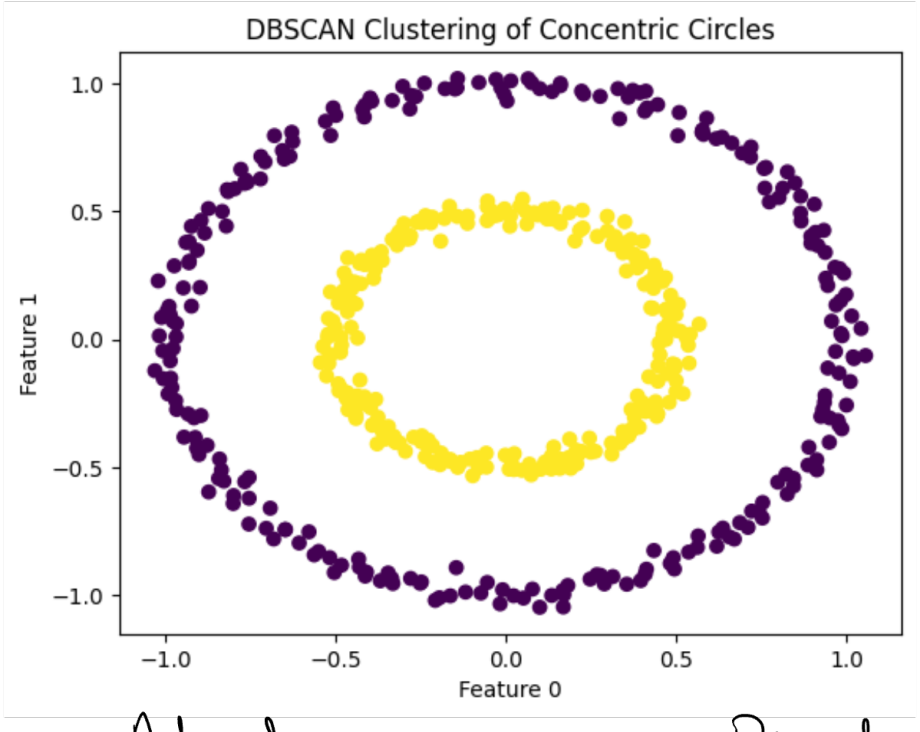
Step 5 - For each unclustered border point assign it to the cluster of nearest core point

Step 4 - Leave all the noise points as it is.

DBSCAN is one of the only algorithm that does not have a predict function. Therefore, it can only be used for clustering an existing data. (No prediction)

This algorithm was developed by software engineering domain engineers which is why, this does not have any mathematical equation.





Advantages

1. Robust to outliers
2. No need to specify clusters
3. Can find arbitrary shaped clusters
4. Only 2 hyperparameters

Disadvantages

1. Sensitive to hyperparamet
   (very hard on outliers)
   (noise)
2. Difficult with varying density clusters
3. Does not predict

Applications of DBSCAN

1. **Spatial Data Analysis**: DBSCAN is particularly well-suited for spatial data clustering due to its ability to find clusters of arbitrary shapes, which is common in geographic data. It's used in applications like identifying regions of similar land use in satellite images or grouping locations with similar activities in GIS (Geographic Information Systems).

2. **Anomaly Detection**: The algorithm's effectiveness in distinguishing noise or outliers from core clusters makes it useful in anomaly detection tasks, such as detecting fraudulent activities in banking transactions or identifying unusual patterns in network traffic.

3. **Image Processing:** In image analysis, DBSCAN can be used for tasks like object recognition and image segmentation, where the goal is to group pixels or features that form meaningful structures.

4. **Bioinformatics:** DBSCAN is applied in bioinformatics for tasks such as gene expression data analysis, where it helps to identify groups of genes with similar expression patterns, which might indicate a functional relationship.

5. **Customer Segmentation:** In marketing and business analytics, DBSCAN can be used for customer segmentation by identifying clusters of customers with similar buying behavior or preferences.

6. **Astronomy**: The algorithm is employed in astronomy for tasks like star cluster identification, where it groups stars based on their physical proximity or other attributes.

7. **Environmental Studies:** DBSCAN can be used in environmental monitoring, for example, to cluster areas based on pollution levels or to identify regions with similar environmental characteristics.

8. **Traffic Analysis:** In traffic and transportation studies, DBSCAN is useful for identifying hotspots of traffic congestion or for clustering routes with similar traffic patterns.

9. **Machine Learning and Data Mining:** More broadly, in the fields of machine learning and data mining, DBSCAN is employed for exploratory data analysis, helping to uncover natural structures or patterns in data that might not be apparent otherwise.

10. **Social Network Analysis:** The algorithm can be used to detect communities or groups within social networks based on interaction patterns or shared interests.