Density Based Clustering : DBSCAN

Density-based clustering

Density-based clustering in data mining

Density-based clustering refers to a method that is based on local cluster criterion, such as density connected points.

What is Density-based clustering?

Density-Based Clustering refers to one of the most popular unsupervised learning methodologies used in model building and machine learning algorithms.

The data points in the region separated by two clusters of low point density are considered as noise.

The surroundings with a **radius ε** of a given object are known as the **ε neighborhood of the object.**

If the **ε neighborhood** of the object comprises at least a minimum number, **MinPts** of objects, then it is called a **core object.**

Density-based clustering

Important parameters

There are two different parameters to calculate the density-based clustering.

E_{PS}: It is considered as the maximum radius of the neighborhood.

MinPts: MinPts refers to the minimum number of points in an Eps neighborhood of that point.

NEps (i) : { k belongs to D and dist (i,k) < = Eps}

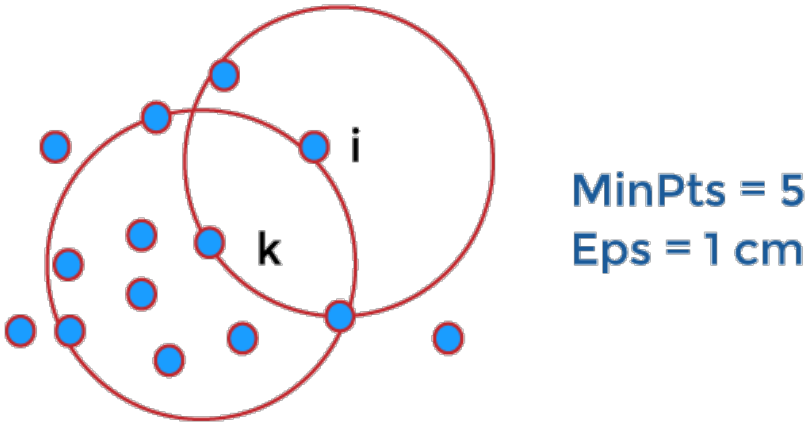
Density-based clustering

Directly density reachable:

A point i is considered as the directly density reachable from a point k with respect to Eps, MinPts if i belongs to NEps(k)

Core point condition:

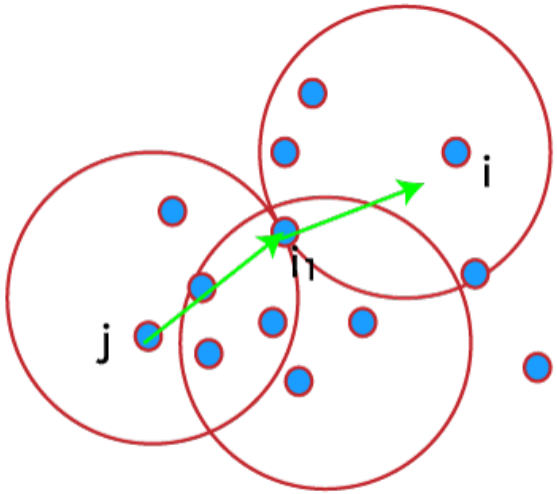
NEps (k) >= MinPts



Density-based clustering

Density reachable:

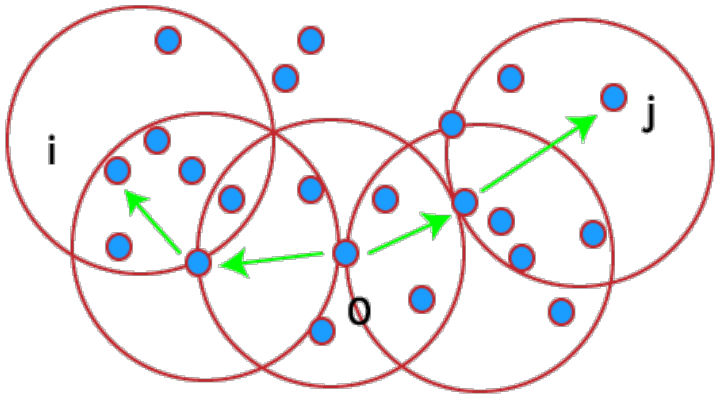
A point denoted by i is a density reachable from a point j with respect to Eps, MinPts if there is a sequence chain of a point i1,…., in, i1 = j, pn = i such that i_i + 1 is directly density reachable from i_i.



Density-based clustering

Density connected:

A point i refers to density connected to a point j with respect to Eps, MinPts if there is a point o such that both i and j are considered as density reachable from o with respect to Eps and MinPts.



DBSCAN

Density connected:

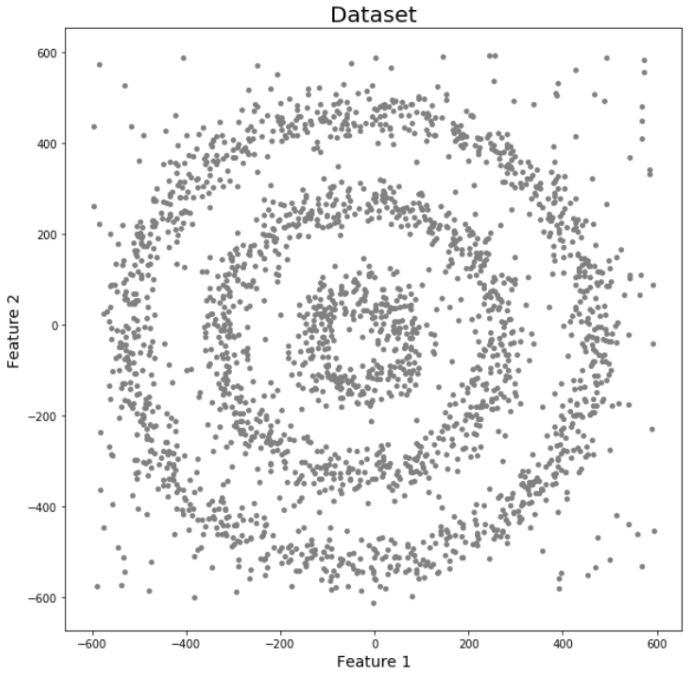
K-Means and Hierarchical Clustering both fail in creating clusters of arbitrary shapes. They are not able to form clusters based on varying densities. That's why we need DBSCAN clustering.

Example.

Here we have data points densely present in the form of concentric circles:

We can see three different dense clusters in the form of concentric

circles with some noise here.



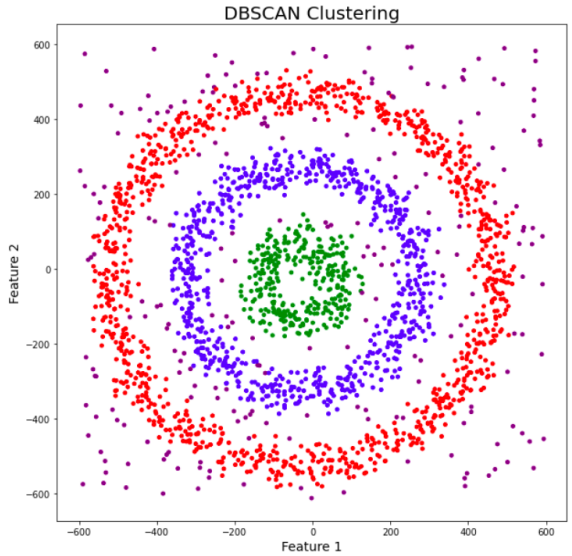
DBSCAN

If we run K-Means and Hierarchical clustering algorithms they will cluster these data points as shown below

DBSCAN

Sadly, both of them failed to cluster the data points. Also, they were not able to properly detect the noise present in the dataset.

The results from DBSCAN clustering are shown below



DBSCAN is not just able to cluster the data points correctly, but it also perfectly detects noise in the dataset.

DBSCAN

DBSCAN

DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density.

It groups ‘densely grouped’ data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points.

The most exciting feature of DBSCAN clustering is that it is robust to outliers.

It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

DBSCAN

DBSCAN

DBSCAN requires only two parameters: epsilon and minPoints. Epsilon is the radius of the circle to be created around each data point to check the density and minPoints is the minimum number of data points required inside that circle for that data point to be classified as a Core point.

In higher dimensions the circle becomes hypersphere, epsilon becomes the radius of that hypersphere, and minPoints is the minimum number of data points required inside that hypersphere.

DBSCAN

Steps of DBSCAN Implementation

1) Find all the neighbor points within eps and identify the core points or visited with more than MinPts neighbors.

2) For each core point if it is not already assigned to a cluster, create a new cluster.

3) Find recursively all its density connected points and assign them to the same cluster as the core point.

A point a and b are said to be density connected if there exist a point c which has a sufficient number of points in its neighbors and both the points a and b are within the eps distance. This is a chaining process. So, if b is neighbor of c, c is neighbor of d, d is neighbor of e, which in turn is neighbor of a implies that b is neighbor of a.

4) Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.


Thank you!