

DBSCAN CLUSTERING

Sourav Karmakar

souravkarmakar29@gmail.com

Density Based Clustering Methods

Density based clustering is a clustering technique that employs density parameter (local cluster criterion) or explicitly constructed density functions.

The major features of Density Based Clustering is following:

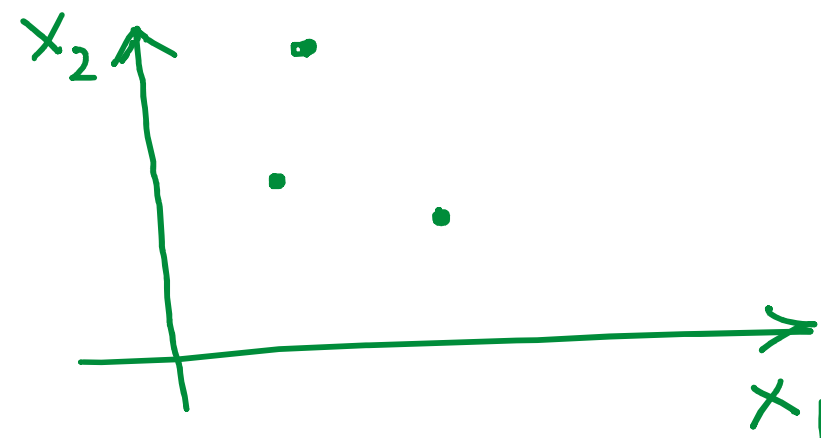
- It can discover clusters of arbitrary shape
- It can handle noisy data points
- Only single scan/iteration is required to do the clustering.
- It needs density parameter

There are several density based clustering techniques developed. Among which we shall focus on **DBSCAN**.

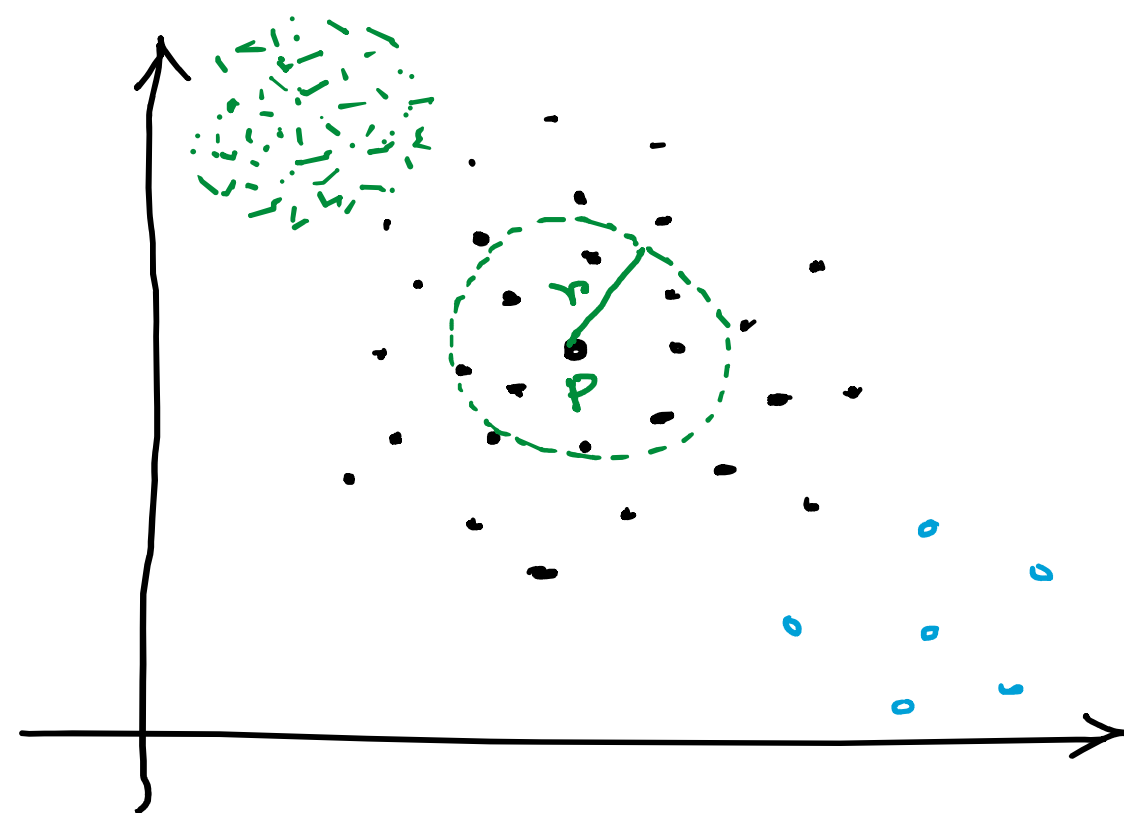
DBSCAN stands for *Density-Based Spatial Clustering of Applications with Noise*. It can find arbitrary shaped clusters and clusters with noise (i.e., outliers).

Density :-

x_1	x_2
0.5	0.3
0.4	0.7
0.2	0.4



Each instance of the data can be represented as points in feature space.



$$\text{Volume} \propto r^n$$

$n \rightarrow$ dimension of the data.

Number of points ' N '

$$\text{Density (d)} = \frac{\text{number of points}}{\text{volume}}$$

$$d \propto \left(\frac{N}{r^n} \right)$$

DBSCAN

The main idea behind **DBSCAN** is that a point belongs to a cluster if it is close to many points from that cluster.

There are two key parameters of **DBSCAN**:

eps: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to ϵ .

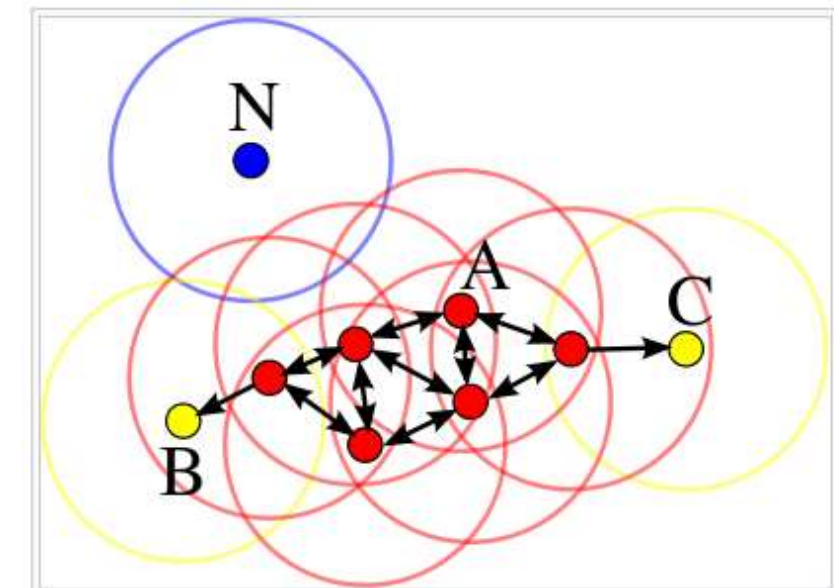
minPts: Minimum number of data points to define a cluster.

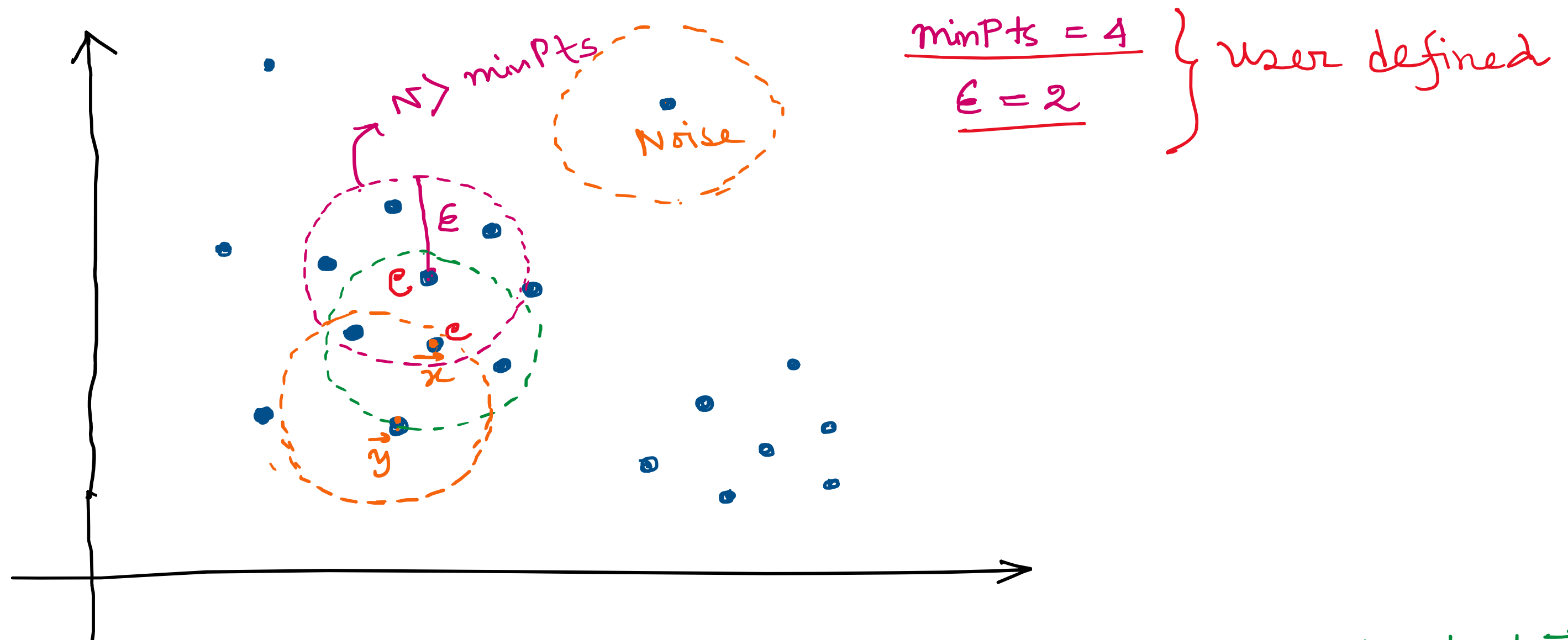
Based on these two parameters, points are classified as core point, border point, or outlier:

Core point: A point is a core point if there are at least **minPts** number of points (including the point itself) in its surrounding area with radius **eps**.

Border point: A point is a border point if it is reachable from a core point and there are less than **minPts** number of points within its surrounding area.

Outlier or Noise: A point is an outlier/noise if it is not a core point and not reachable from any core points.



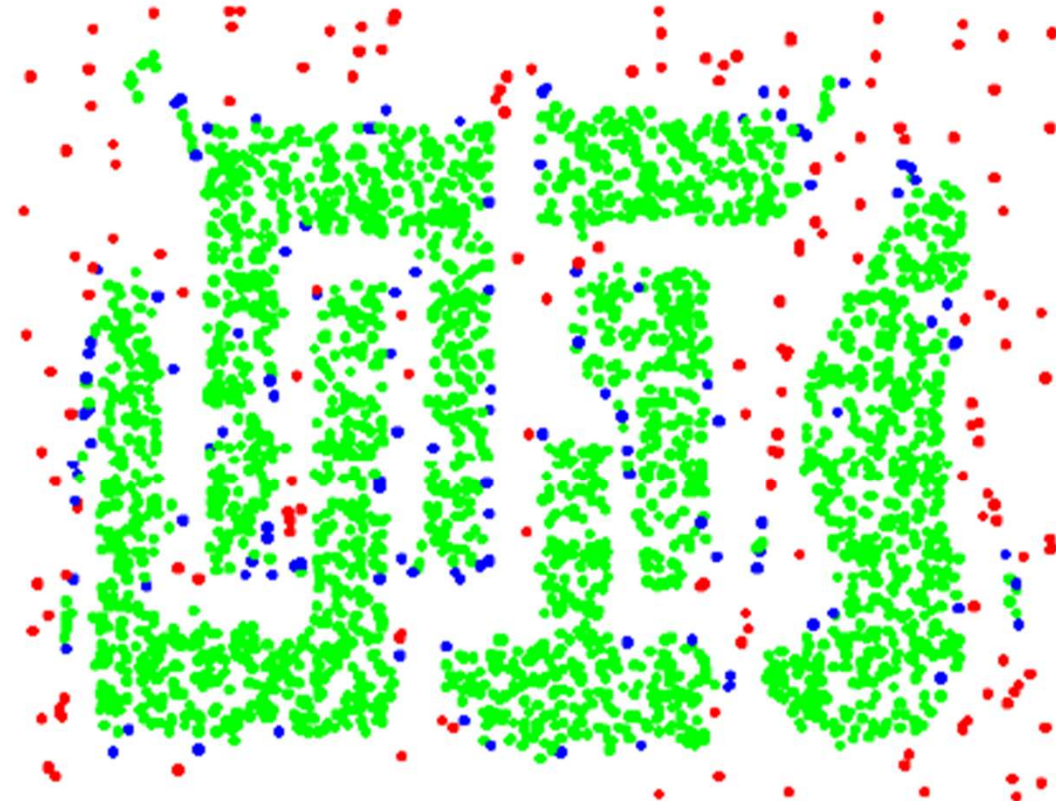


$N \rightarrow$ number of points which is ϵ reachable from the point \vec{x}
 if $N \geq \text{minPts}$ then \vec{x} is a core point
 $N < \text{minPts}$ but \vec{x} is ϵ reachable from a core point then \vec{x}
 is border point
 else \vec{x} is noise / outlier.

DBSCAN



Original Points



Point types: core,
border and noise

Eps = 10, MinPts = 4

DBSCAN

Following is the algorithm of DBSCAN:

Suppose the dataset is denoted by D and x is a point / object in the dataset.

For each $x \in D$

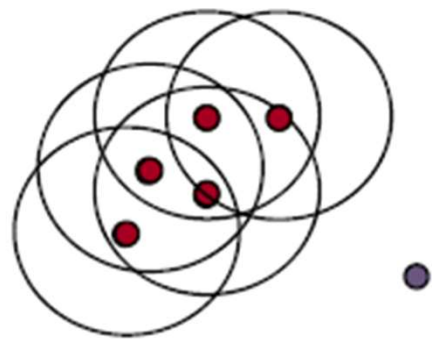
If x is not yet assigned to any cluster

If x is a **core point** **then**

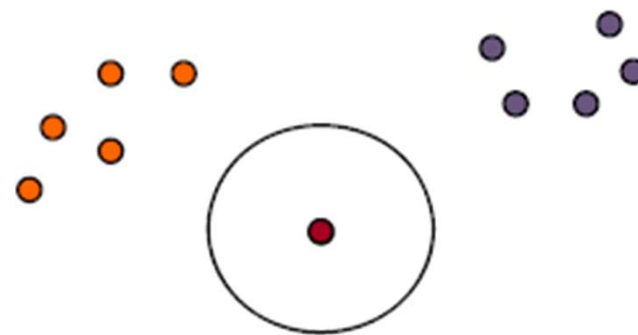
Collect all the points which are density-reachable (i.e. within radius ϵ) from x and assign them to a new cluster.

Else assign x Noise

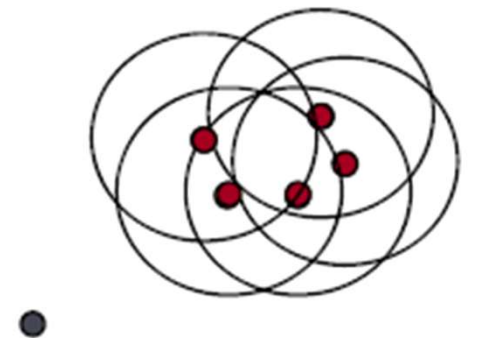
Cluster-A



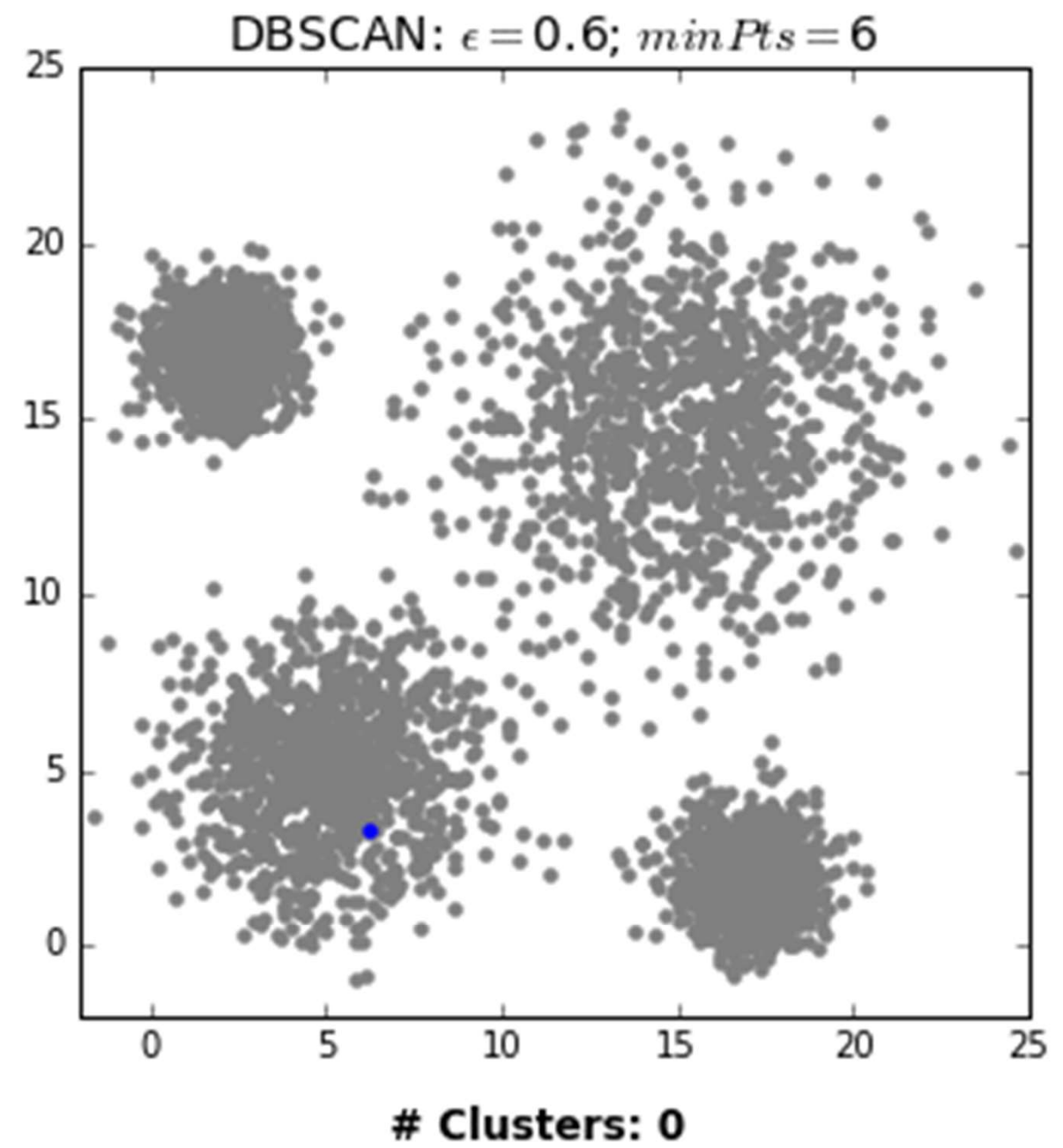
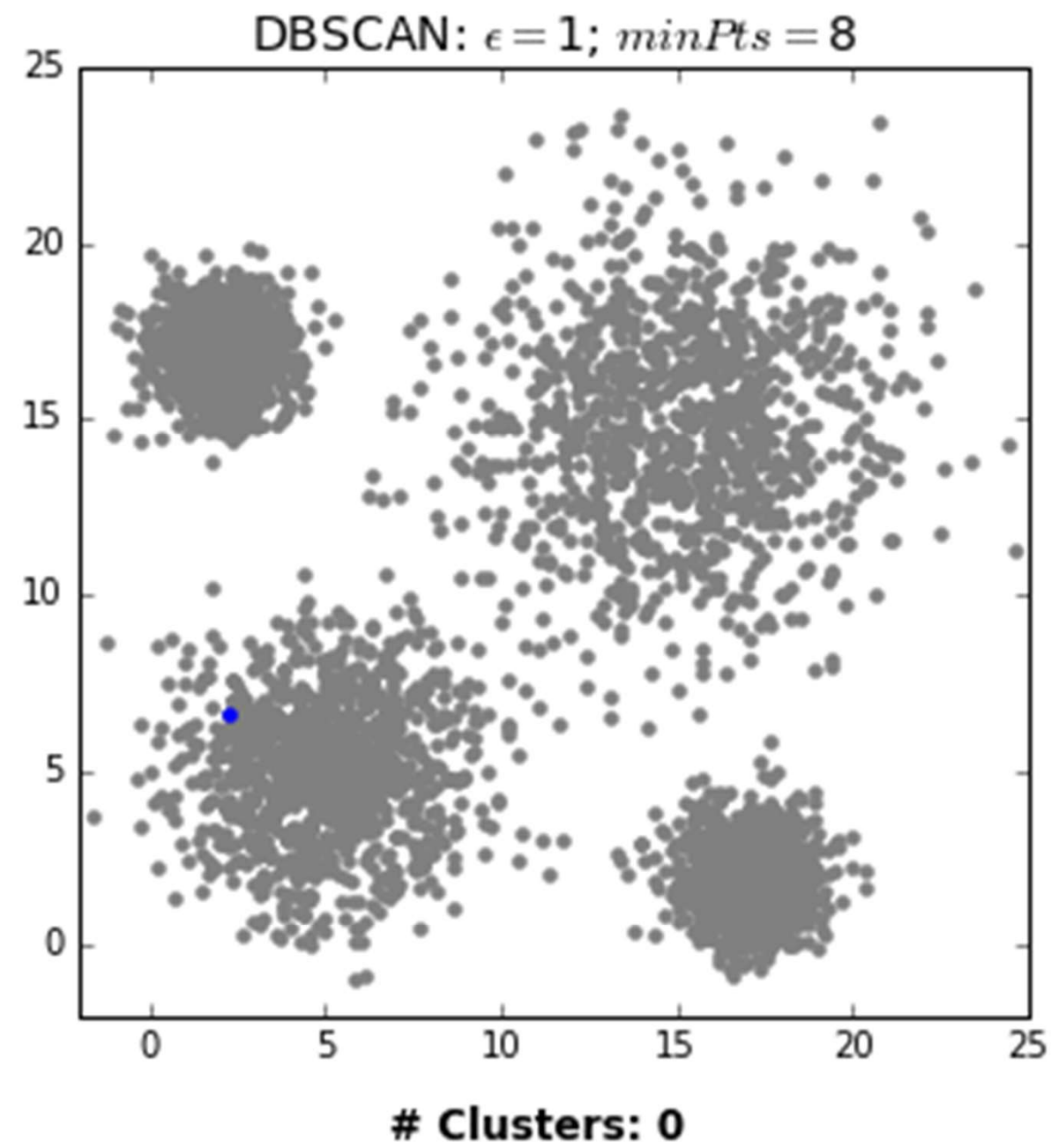
Noise Point



Cluster-B



DBSCAN



DBSCAN

Pros:

- Does not require to specify number of clusters beforehand.
- Performs well with arbitrary shapes clusters.
- DBSCAN is robust to outliers and able to detect the outliers.

Cons:

- In some cases, determining an appropriate distance of neighborhood (eps) is not easy and it requires domain knowledge.
- If clusters are very different in terms of in-cluster densities, DBSCAN is not well suited to define clusters. The characteristics of clusters are defined by the combination of eps-minPts parameters. Since we pass in one eps-minPts combination to the algorithm, it cannot generalize well to clusters with much different densities.

Thank You