Density-based Clustering

Density-based Clustering

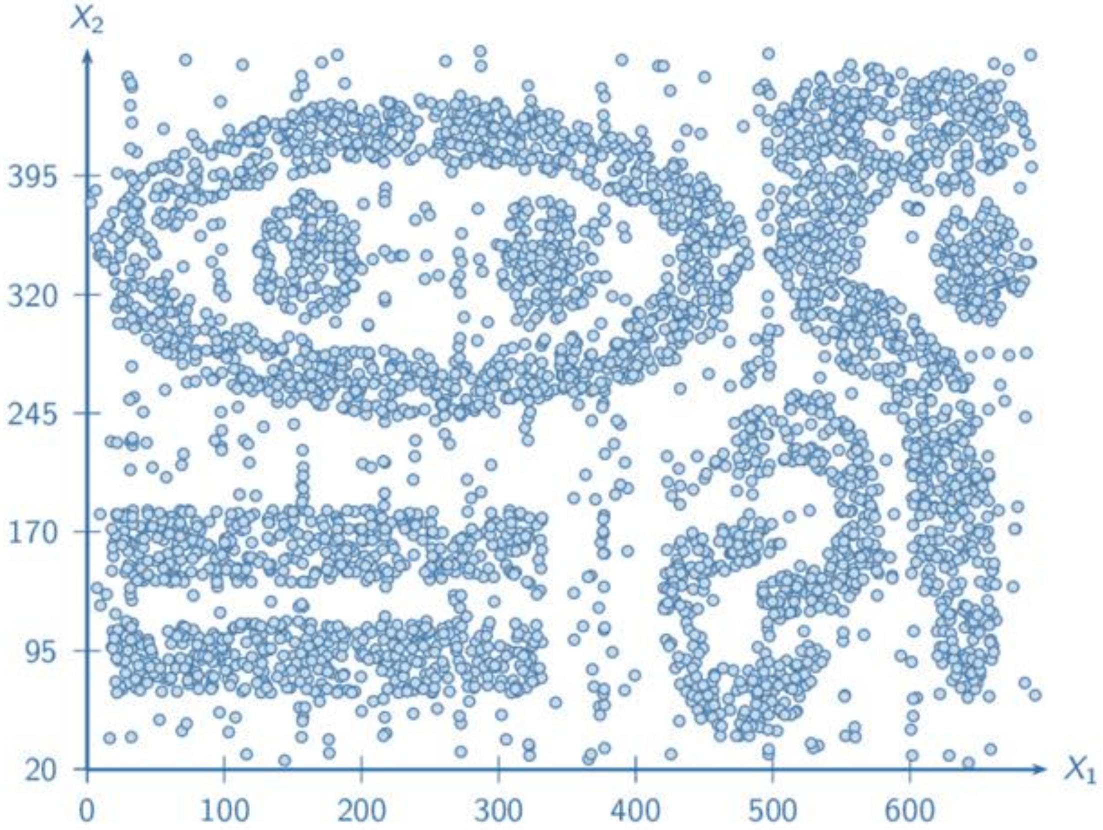
- **[1] Data Mining and Machine Learning: Fundamental Concepts and Algorithms**, Second Edition, *Mohammed J. Zaki and Wagner Meira*, *Jr*, Cambridge University Press, March 2020,ISBN: 978-1108473989.

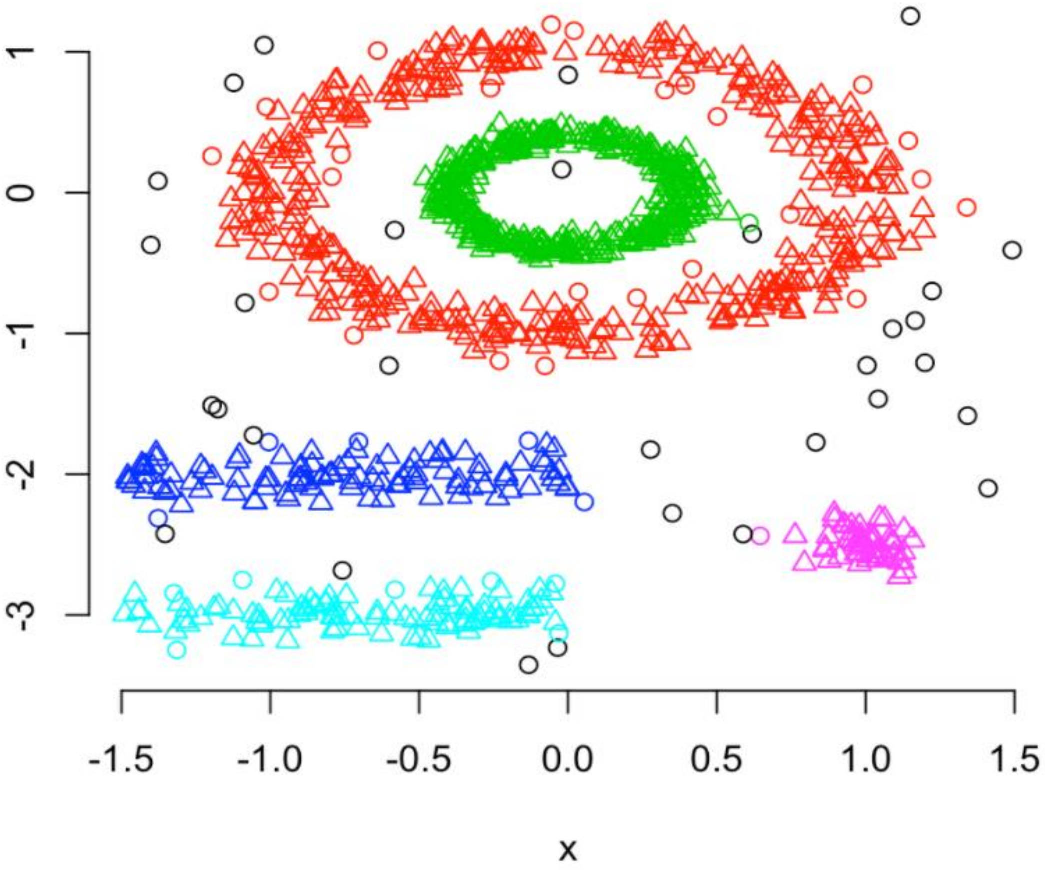
Density-based Clustering

- The representative-based clustering methods like K-means and expectation maximization are suitable for finding ellipsoid-shaped clusters, or at best convex clusters.
- However, for nonconvex clusters, K-means and expectation maximization methods have trouble finding the true clusters, as two points from different clusters may be closer than two points in the same cluster.

Density-Based Clustering

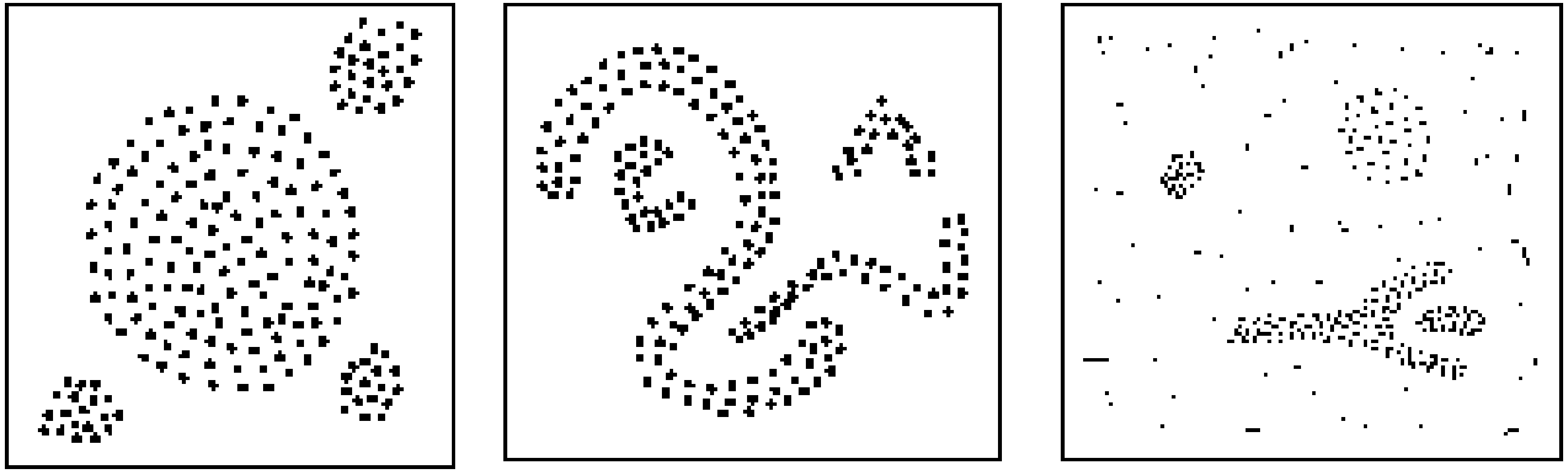
- Density-based methods are able to mine nonconvex clusters, where distance-based methods may have difficulty.





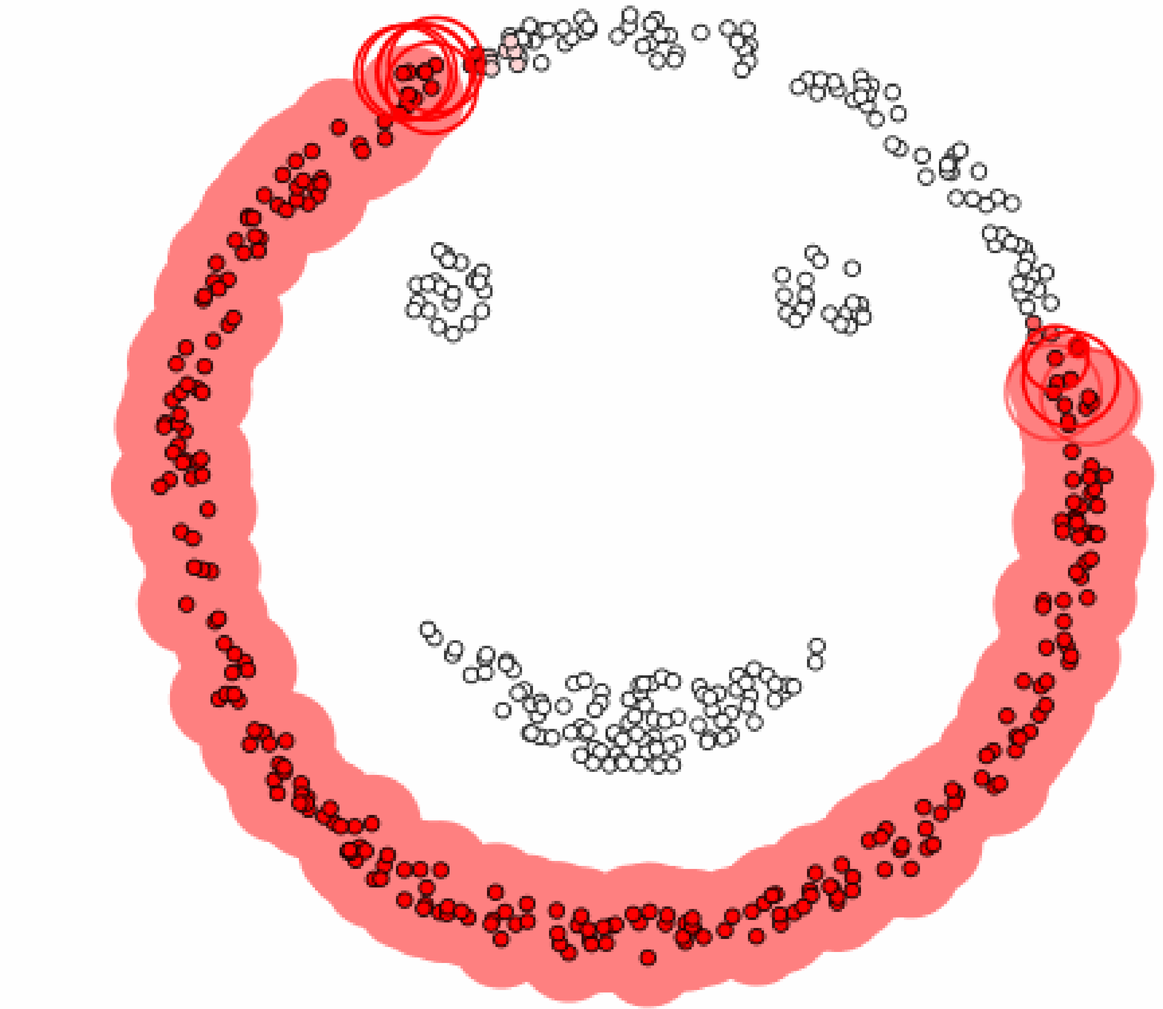
Density-Based Clustering

- Clustering based on density (local cluster criterion), such as density-connected points
- Each cluster has a considerable higher density of points than outside of the cluster



Density-Based Clustering Methods

1. Discover clusters of arbitrary s
2. Handle noise
3. One scan
4. Need density parameters

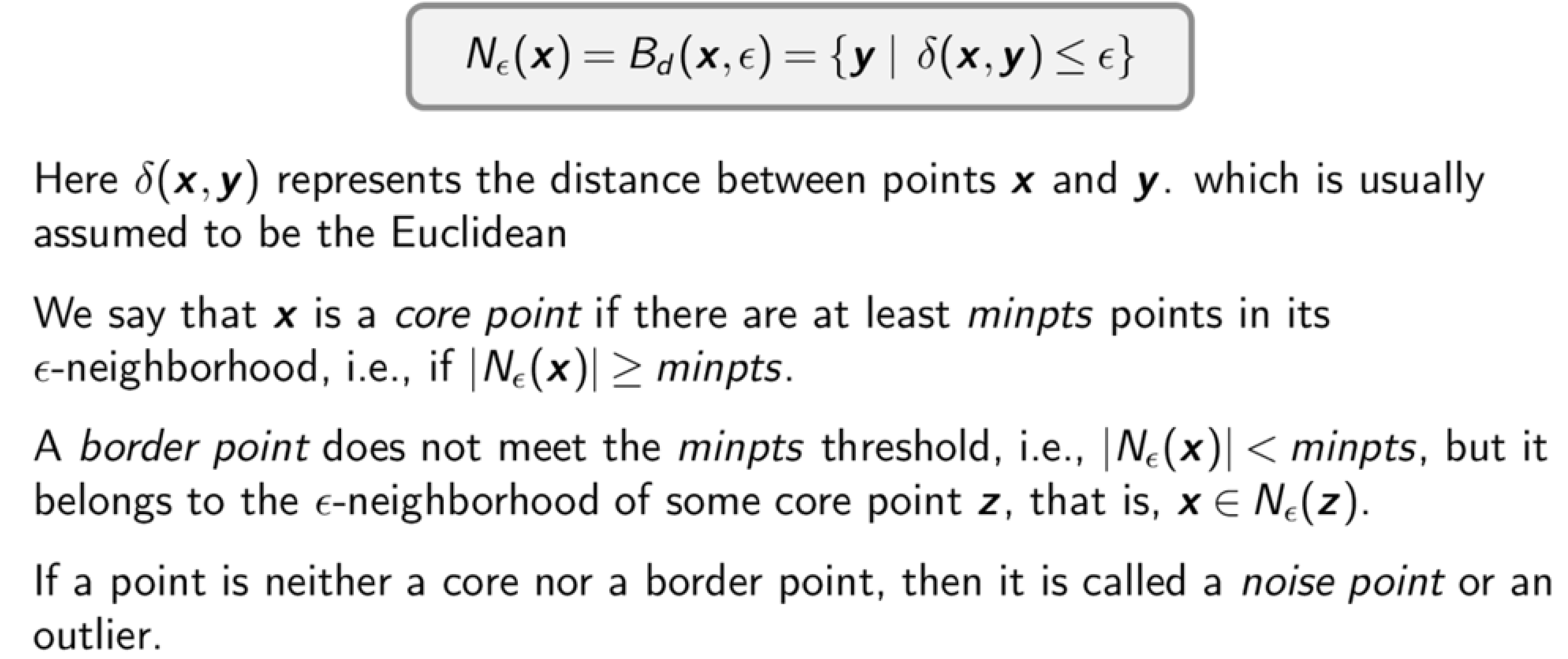


DBSCAN

- Density-Based Spatial Clustering of Applications with Noise

DBSCAN

- Density-based clustering uses the local density of points to determine the clusters, rather than using only the distance between points.
- We de fine a ball of radius ϵ around a point $x \in \mathbb{R}^d Rd$, called the ϵ -neighborhood of x w



$$\boxed{N_\epsilon(\boldsymbol{x}) = B_d(\boldsymbol{x}, \epsilon) = \{\boldsymbol{y} \mid \delta(\boldsymbol{x}, \boldsymbol{y}) \leq \epsilon\}}$$

Here $\delta(\boldsymbol{x}, \boldsymbol{y})$ represents the distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$. which is usually assumed to be the Euclidean

We say that $\boldsymbol{x}$ is a *core point* if there are at least *minpts* points in its ϵ-neighborhood, i.e., if $|N_\epsilon(\boldsymbol{x})| \geq minpts$.

A *border point* does not meet the *minpts* threshold, i.e., $|N_\epsilon(\boldsymbol{x})| < minpts$, but it belongs to the ϵ-neighborhood of some core point $\boldsymbol{z}$, that is, $\boldsymbol{x} \in N_\epsilon(\boldsymbol{z})$.

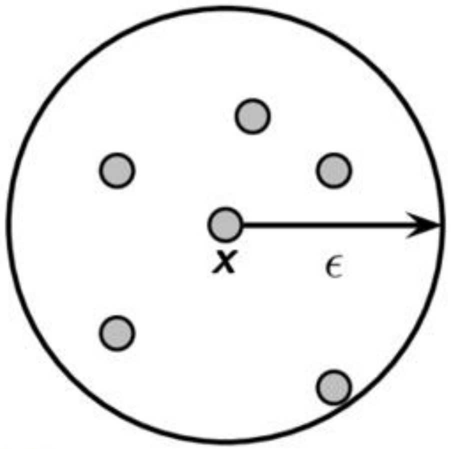
If a point is neither a core nor a border point, then it is called a *noise point* or an outlier.

[1]

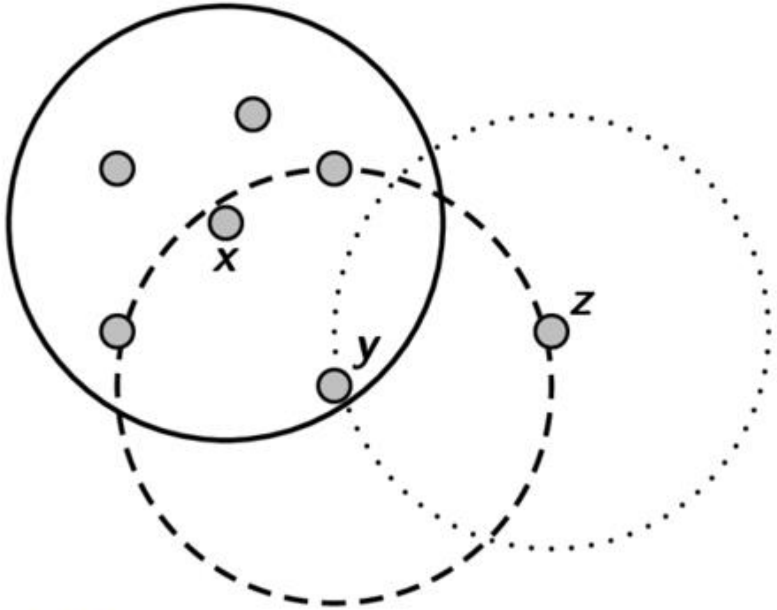
Core Border and Noise Points

- *minpt s* = 6.
    - $\boldsymbol{x}$ is a core point because $|N_{\epsilon}\ (\boldsymbol{x})|\ =\ 6$,
    - **y** is a border point because $|N_{\epsilon}\epsilon(\boldsymbol{y})| < minpts$, but it belongs to the ϵ -neighborhood of the core point $\boldsymbol{x}$,

      $\mathbf{y} \in N_{\epsilon}\ (\boldsymbol{x})$
    - **z** is a noise point.

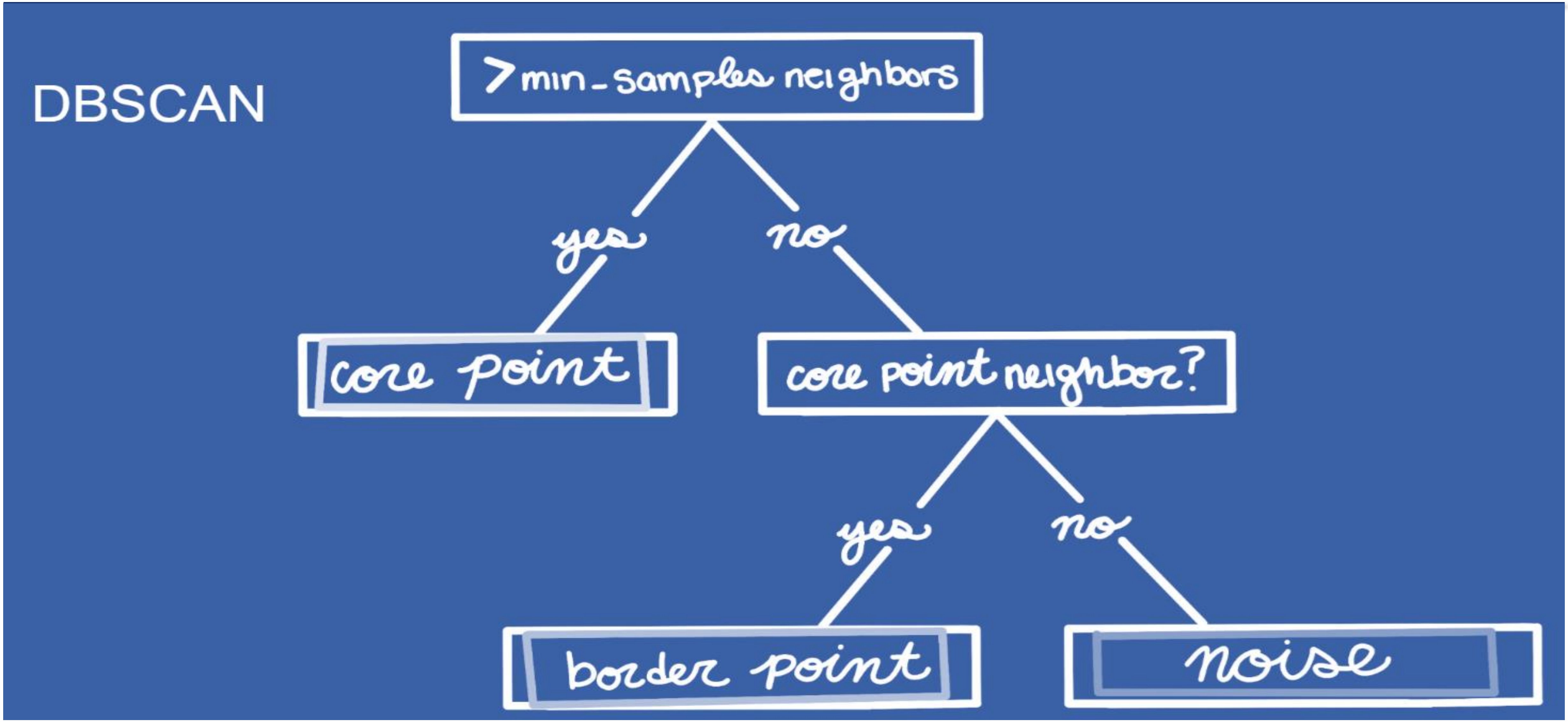


(a) Neighborhood of a Point



(b) Core, Border, and Noise Points

Core Border and Noise Points



https://github.com/cmparlettpelleriti/CPSC392ParlettPelleriti/blob/master/InClass/LectureSlides/Lecture%2018%20_%20DBSCAN.pdf

DBSCAN Reachability

A point $\boldsymbol{x}$ is *directly density reachable* from another point $\boldsymbol{y}$ if $\boldsymbol{x} \in N_\epsilon(\boldsymbol{y})$ and $\boldsymbol{y}$ is a core point.

A point $\boldsymbol{x}$ is *density reachable* from $\boldsymbol{y}$ if there exists a chain of points, $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$, such that $\boldsymbol{x} = \boldsymbol{x}_0$ and $\boldsymbol{y} = \boldsymbol{x}_l$, and $\boldsymbol{x}_i$ is directly density reachable from $\boldsymbol{x}_{i-1}$ for all $i = 1, \ldots, l$. In other words, there is set of core points leading from $\boldsymbol{y}$ to $\boldsymbol{x}$.

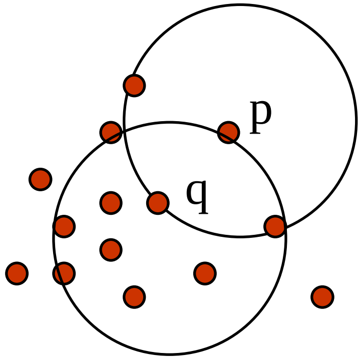
Two points $\boldsymbol{x}$ and $\boldsymbol{y}$ are *density connected* if there exists a core point $\boldsymbol{z}$, such that both $\boldsymbol{x}$ and $\boldsymbol{y}$ are density reachable from $\boldsymbol{z}$.

A *density-based cluster* is defined as a maximal set of density connected points.

DBSCAN

- Directly density-reachable**:** A point ***p*** is directly density-reachable from a point ***q*** wrt. ***Eps***, ***MinPts*** if
  - 1) ***p*** belongs to $\boldsymbol{N_{Eps}(q)}$
  - 2) core point condition:

$$|\boldsymbol{N_{Eps}}\ \boldsymbol{(q)}| >= \boldsymbol{MinPts}$$

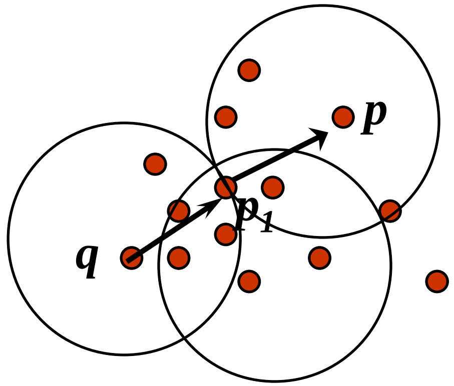


MinPts = 5

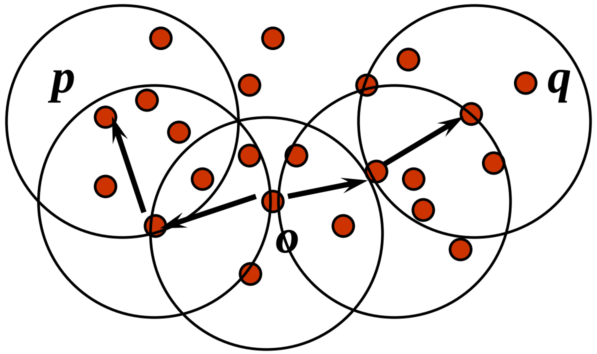
Eps = 1 cm

DBSCAN

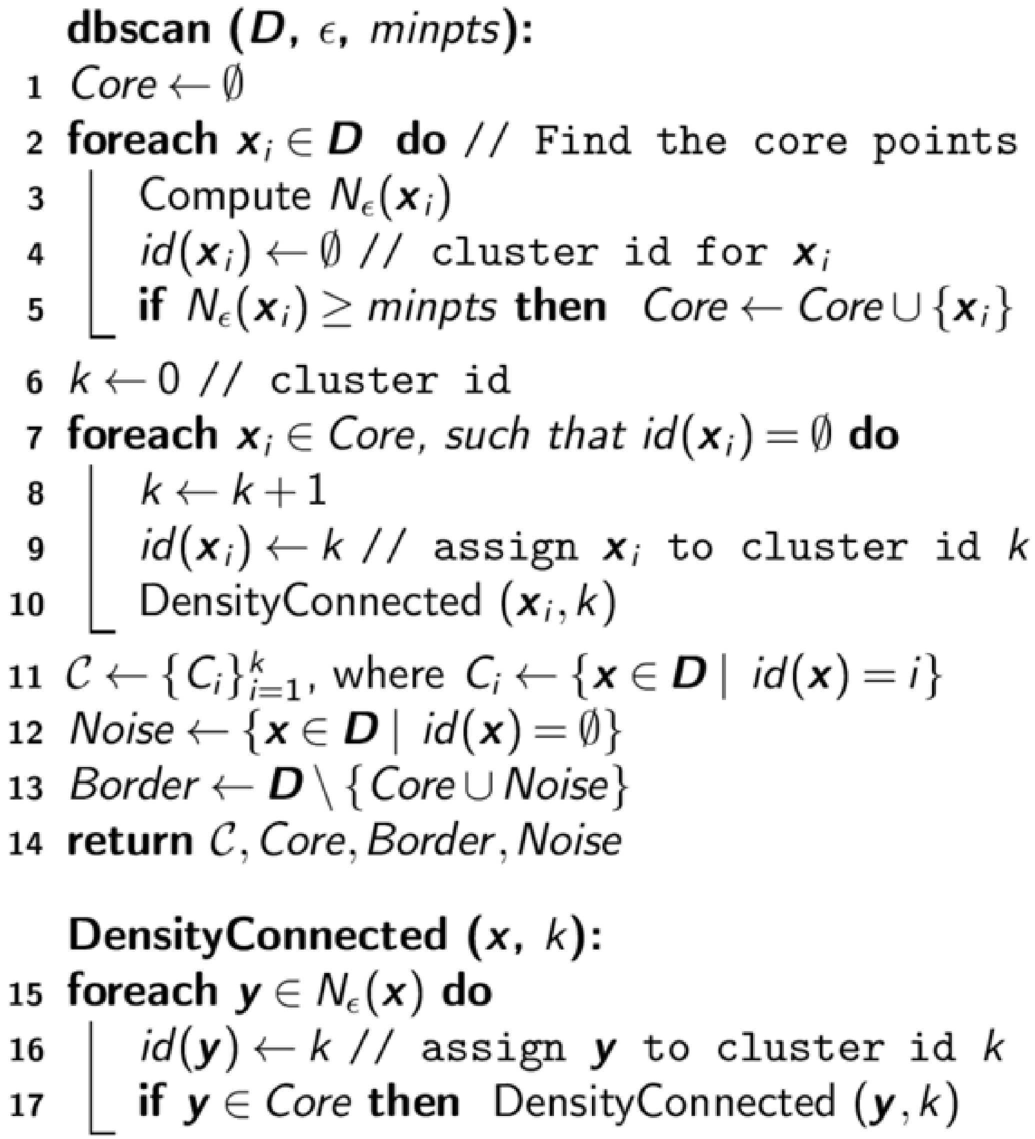
- Density-reachable:
  - A point *p* is density-reachable from a point *q* wrt. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



- Density-connected
  - A point *p* is density-connected to a point *q* wrt. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* wrt. *Eps* and *MinPts*.



DBSCAN Algorithm



```
dbscan ($D$, $\epsilon$, $minpts$):
Core ← $\emptyset$
foreach $x_i \in D$ do // Find the core points
    Compute $N_\epsilon(x_i)$
    $id(x_i) \leftarrow \emptyset$ // cluster id for $x_i$
    if $N_\epsilon(x_i) \geq minpts$ then $Core \leftarrow Core \cup \{x_i\}$
$k \leftarrow 0$ // cluster id
foreach $x_i \in Core$, such that $id(x_i) = \emptyset$ do
    $k \leftarrow k + 1$
    $id(x_i) \leftarrow k$ // assign $x_i$ to cluster id $k$
    DensityConnected $(x_i, k)$
$\mathcal{C} \leftarrow \{C_i\}_{i=1}^{k}$, where $C_i \leftarrow \{x \in D \mid id(x) = i\}$
$Noise \leftarrow \{x \in D \mid id(x) = \emptyset\}$
$Border \leftarrow D \setminus \{Core \cup Noise\}$
return $\mathcal{C}, Core, Border, Noise$

DensityConnected ($x$, $k$):
foreach $y \in N_\epsilon(x)$ do
    $id(y) \leftarrow k$ // assign $y$ to cluster id $k$
    if $y \in Core$ then DensityConnected $(y, k)$
```

DBSCAN Algorithm

DBSCAN computes the ϵ-neighborhood $N_\epsilon(\boldsymbol{x}_i)$ for each point $\boldsymbol{x}_i$ in the dataset $\boldsymbol{D}$, and checks if it is a core point. It also sets the cluster id $id(\boldsymbol{x}_i) = \emptyset$ for all points, indicating that they are not assigned to any cluster.

Starting from each unassigned core point, the method recursively finds all its density connected points, which are assigned to the same cluster.

Some border point may be reachable from core points in more than one cluster; they may either be arbitrarily assigned to one of the clusters or to all of them (if overlapping clusters are allowed).

Those points that do not belong to any cluster are treated as outliers or noise.

Each DBSCAN cluster is a maximal connected component over the core point graph.

DBSCAN is sensitive to the choice of ϵ, in particular if clusters have different densities. The overall complexity of DBSCAN is $O(n^2)$.

Complexity DBSCAN

- Time Complexity: O(n^2)—for each point it has to be determined if it is a core point, can be reduced to O(n*log(n)) in lower dimensional spaces by using efficient data structures (n is the number of objects to be clustered);
- Space Complexity: O(n).

DBSCAN Good?

- Good:
  - can detect arbitrary shapes
  - not very sensitive to noise, supports outlier detection
  - complexity ?
- Bad:
  - does not work well in high-dimensional datasets
  - parameter selection
  - Does not create a real density function, only a graph of density-connected points