Lecture 10. Density-based clustering

Xin Chen

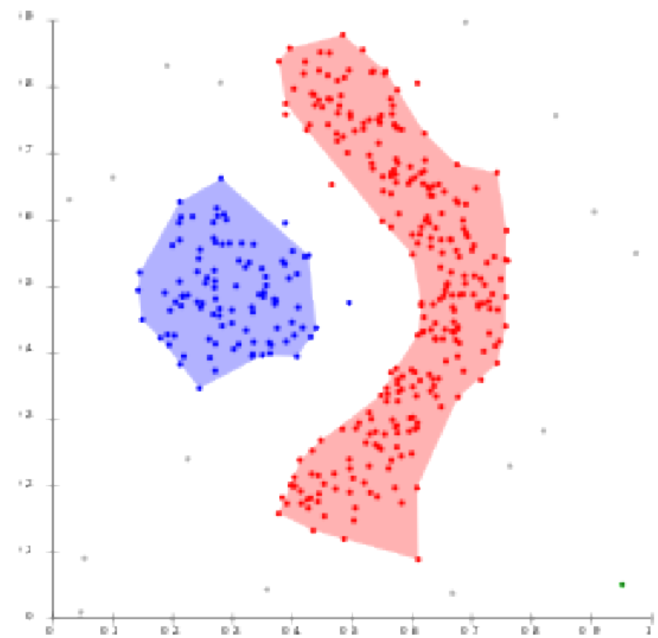
These slides are based on slides from Mahdi Roozbahani

Outline

- Overview
- Basic concepts
- The DBSCAN algorithm
- Analysis of DBSCAN algorithm

Density-Based Clustering

- Basic ideas
  - Clusters are dense regions in the data space, separated by regions of lower density
  - A cluster is defined as a maximal set of density-connected points
  - Detect arbitrarily shaped clusters



- Method
  - DBSCAN (Density-Based Spatial Clustering of Application with Noise)

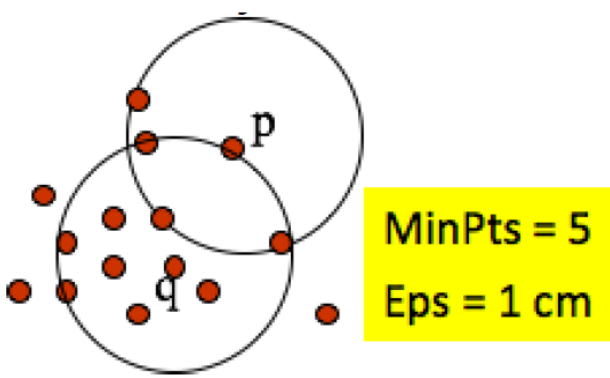
Outline

- Overview
- Basic concepts
- The DBSCAN algorithm
- Analysis of DBSCAN algorithm

High Density v.s. Low Density

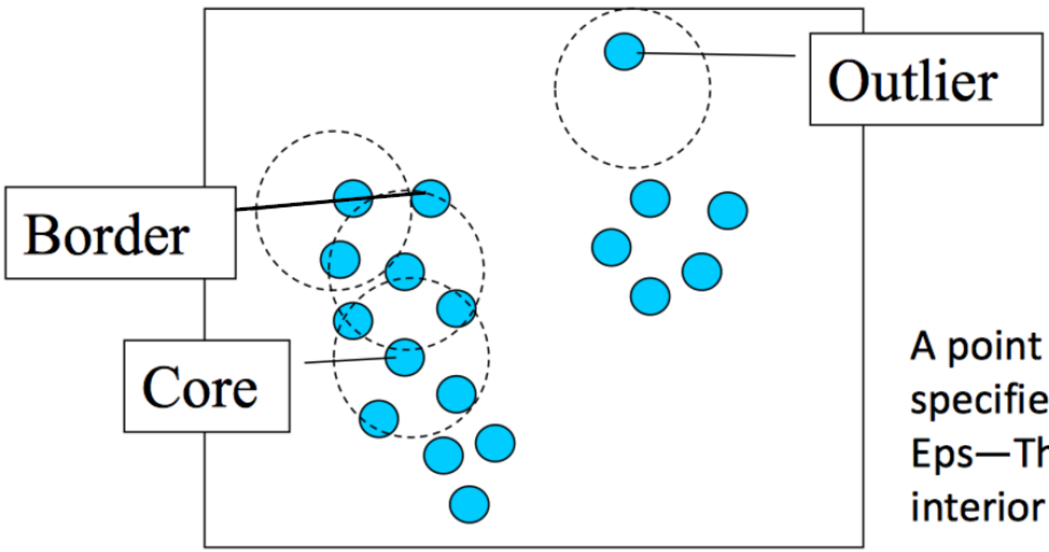
- Two parameters
  - Eps(ε): maximum radius of the neighborhood
  - MinPts: minimum number of points in the Eps-neighborhood of a point

- High density: ε-neighborhood of an object contains at least MinPts of objects



Density of ***p*** is low
Density of ***q*** is high

Core Points, Border Points, and Outliers



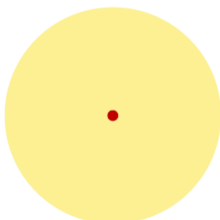
ε = 1unit, MinPts = 5

Given ε and *MinPts*, categorize the objects into three exclusive groups.

A point is a core point if it has more than a specified number of points (MinPts) within Eps—These are points that are at the interior of a cluster.

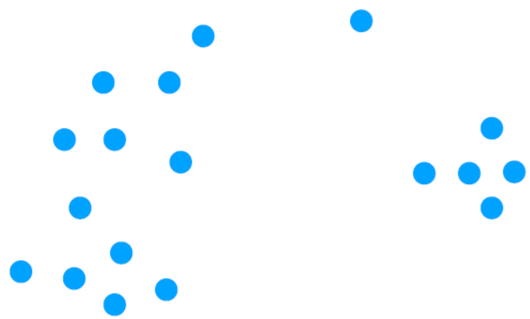
A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.

A noise point is any point that is not a core point nor a border point.



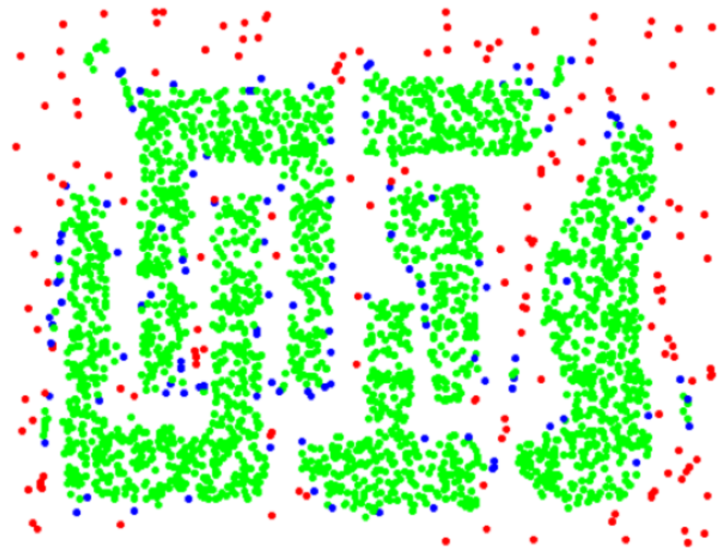
ε= 1 unit

MinPts = 5



Examples

Original Points

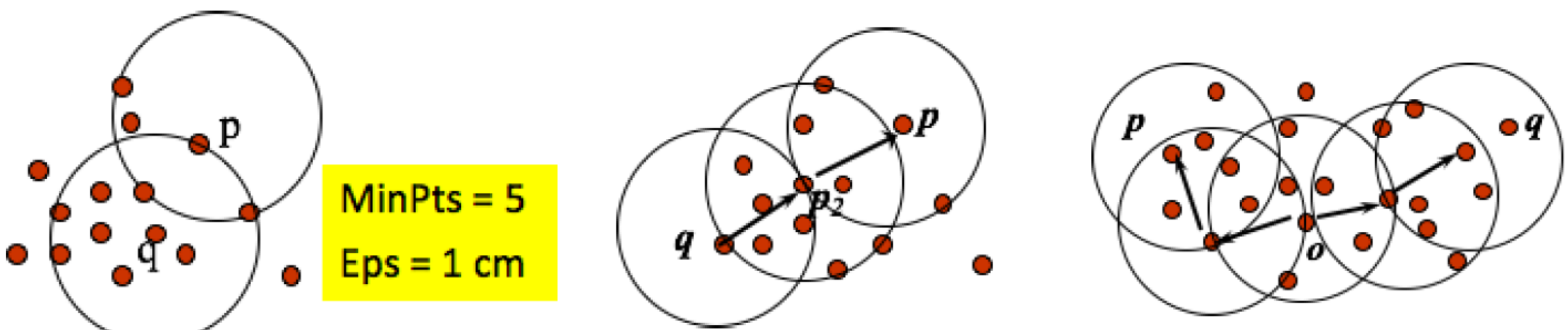


Point types: core, border and outliers

ε = 10, MinPts = 4

Density-based related points

- Direct density reachability
  - An object p is directly density-reachable from object q if
    - q is a core object
    - p is in q’s ε-neighborhood



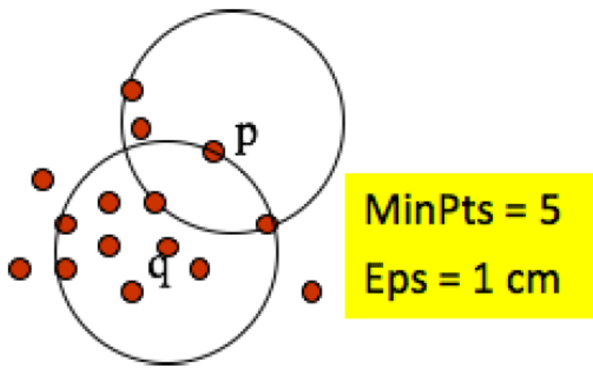
Directly Density-Reachable

Density-Reachable

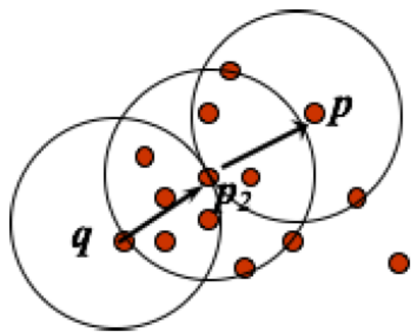
Density-Connected

Density-based related points

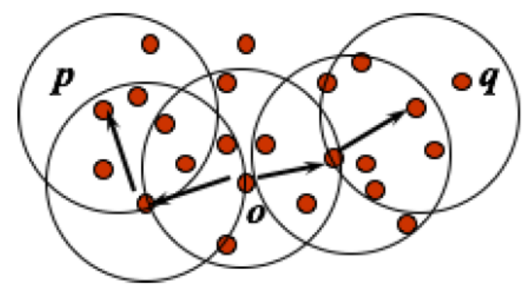
- Density reachability
  - A point p is density-reachable from a point q if there is a chain of point $p_1, p_2, \ldots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i
  - $p_1 = q \rightarrow p_2 \rightarrow \cdots \rightarrow p_n = q$



Directly Density-Reachable



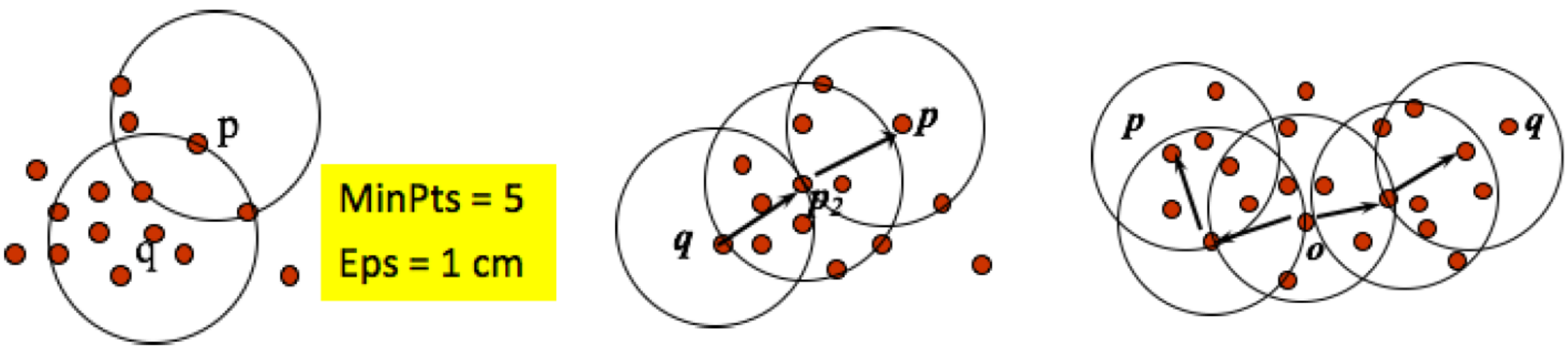
Density-Reachable



Density-Connected

Density-based related points

- Density connectivity
  - A point p is density-connected to a point q if there is a point o such that both p and q are density-reachable from o.



Directly Density-Reachable

Density-Reachable

Density-Connected

Outline

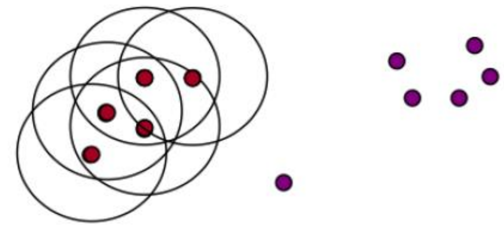
- Overview
- Basic concepts
- The DBSCAN algorithm
- Analysis of DBSCAN algorithm

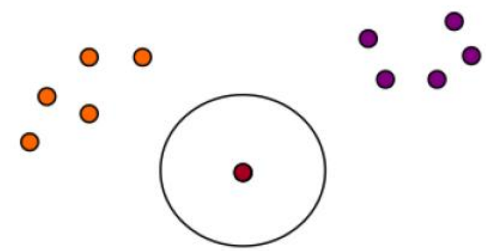
The DBSCAN algorithm

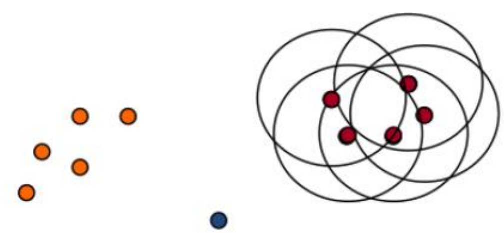
```
DBSCAN(D, eps, MinPts)
C = 0
for each unvisited point P in dataset D
    mark P as visited
    NeighborPts = regionQuery(P, eps)
    if sizeof(NeighborPts) < MinPts
        mark P as NOISE
    else
        C = next cluster
        expandCluster(P, NeighborPts, C, eps, MinPts)

expandCluster(P, NeighborPts, C, eps, MinPts)
    add P to cluster C
    for each point P' in NeighborPts
        if P' is not visited
            mark P' as visited
            NeighborPts' = regionQuery(P', eps)
            if sizeof(NeighborPts') >= MinPts
                NeighborPts = NeighborPts joined with NeighborPts‘
        if P' is not yet member of any cluster
            add P' to cluster C

regionQuery(P, eps) return all points within P's eps-neighborhood (including P)
```







https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Outline

- Overview
- Basic concepts
- The DBSCAN algorithm
- Analysis of DBSCAN algorithm

DBSCAN is sensitive to parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

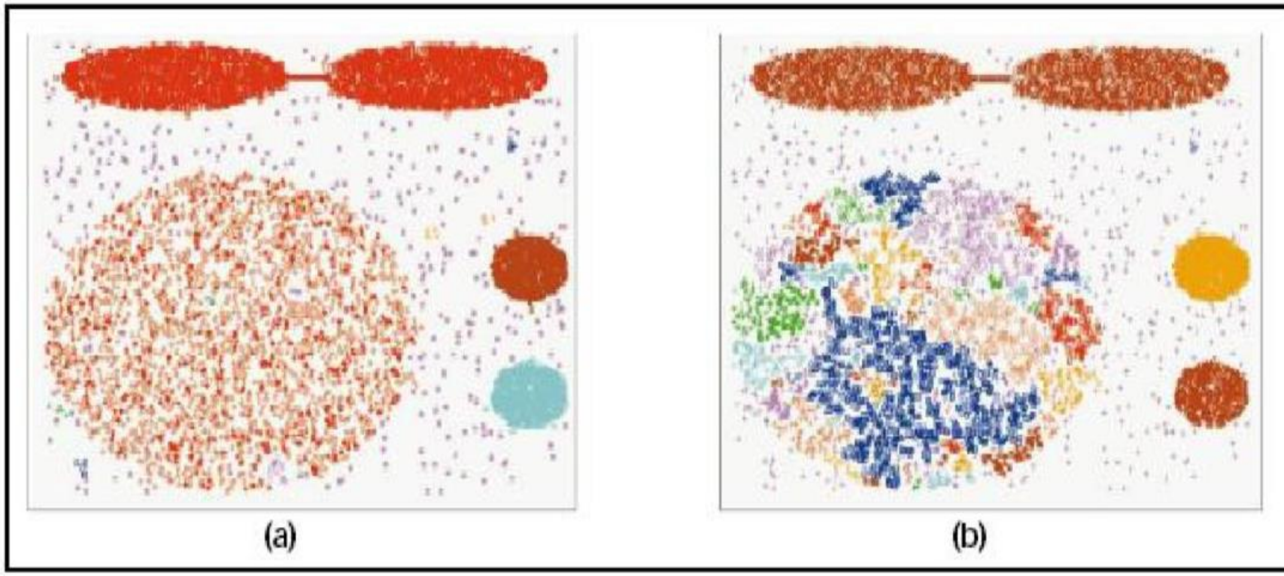


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.

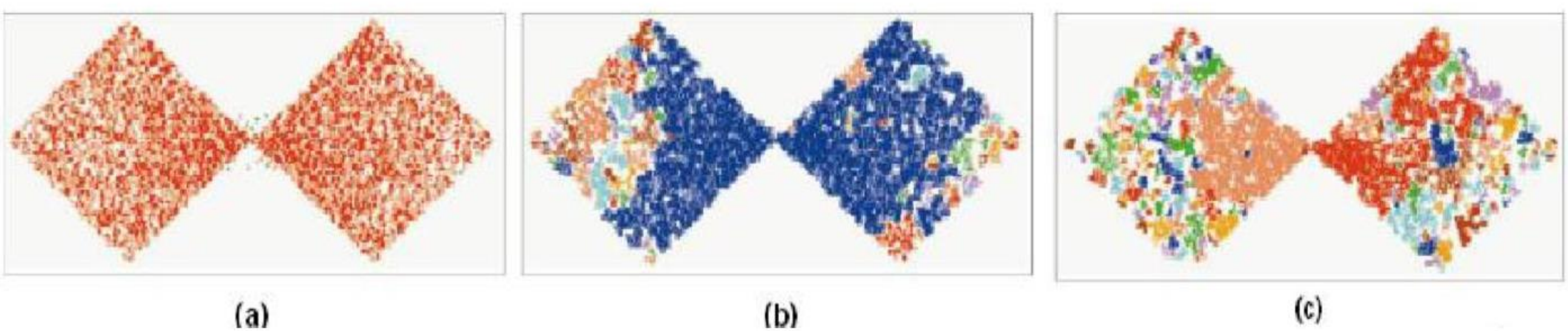


Image Credit: George Karypis.

One parameter

ε → High value (what will happen?) → Clusters will merge and the majority of data points will be in the same cluster

ε → Low value (what will happen?) → A large part of data won't be clustered and considered as outliers.

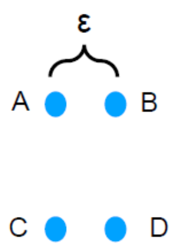
Do we need to define the number of clusters in DBSCAN? → No

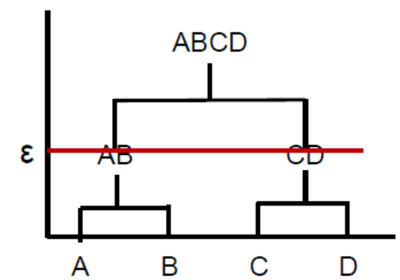
Minimum number of points (MinPts)

MinPts=1?

Every point will be a cluster on its own.

MinPts=2?





Dendogram cut at height ε

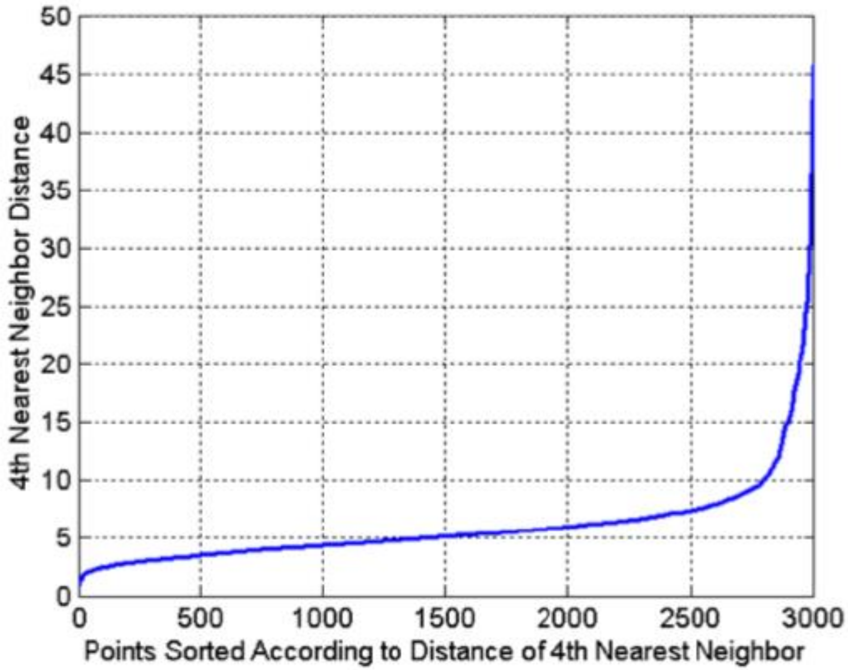
Hierarchical clustering with single link

MinPts should be at least 3

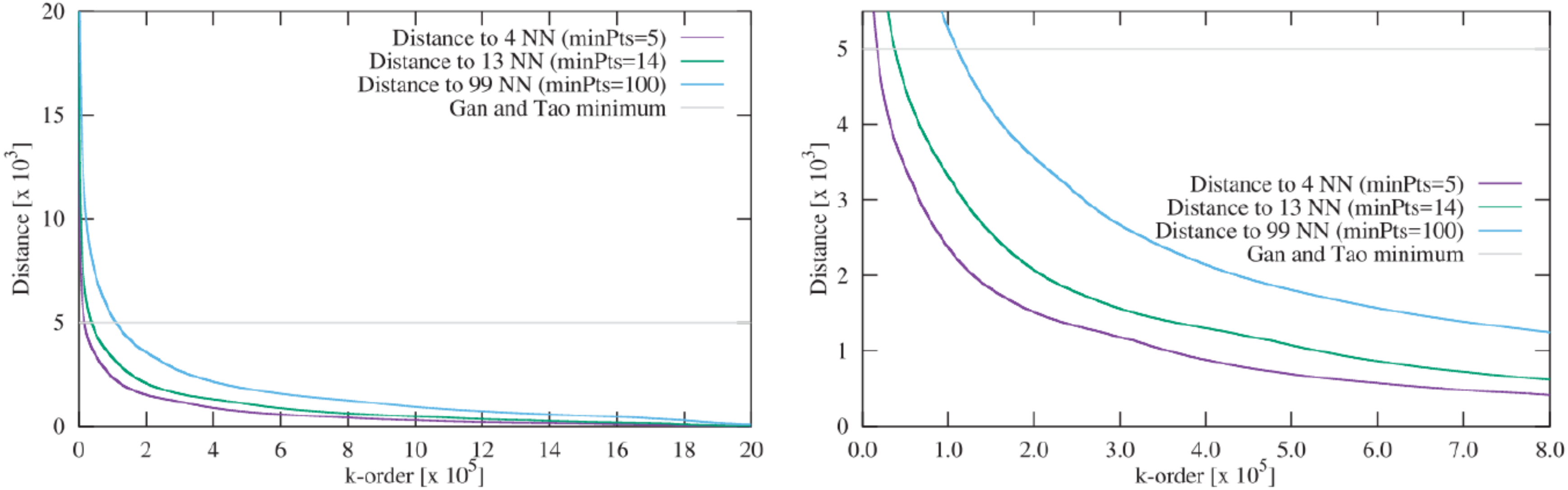
As a rule of the thumb, minPts=2*dim can be used, but it can choose large values for large data and for noise data.

How about Eps? (Elbow effects)

- Idea is that for points in a cluster, their kth nearest neighbors are at roughly the same distance.
- Noise points have the kth nearest neighbor at farther distance
- So, plot sorted distance of every point to its kth nearest neighbor



Elbow effect another example



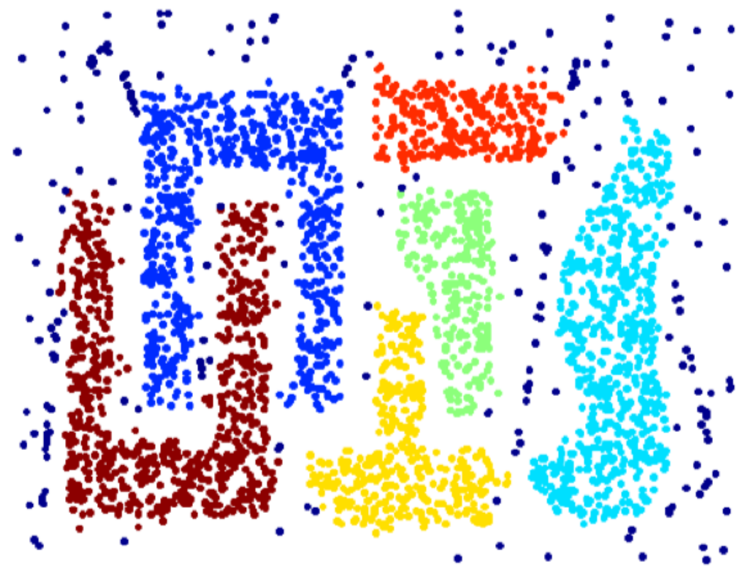
(a) k-distance plots

(b) k-distance plots (magnified region)

minPts often does not have a significant impact on the clustering results.

When DBSCAN works well

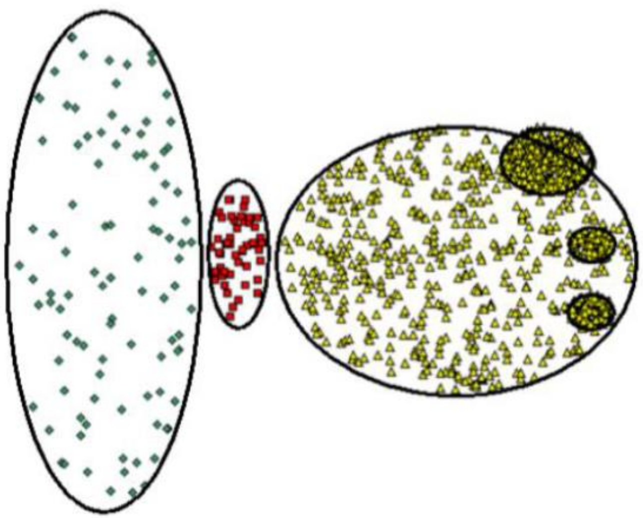
Original points

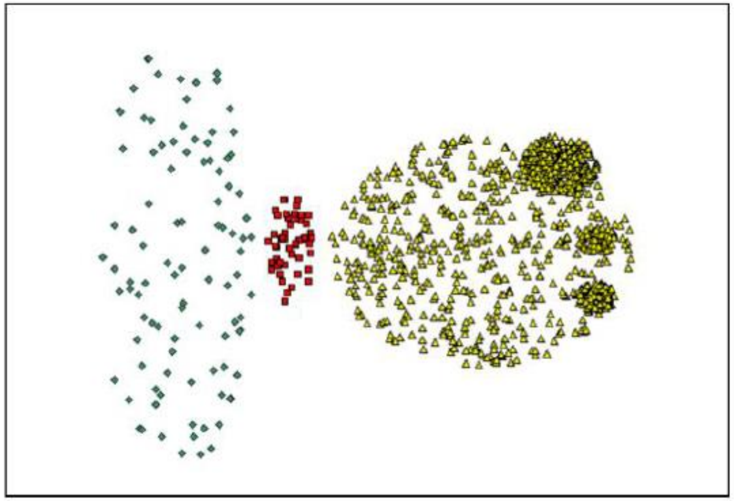


Clustered results

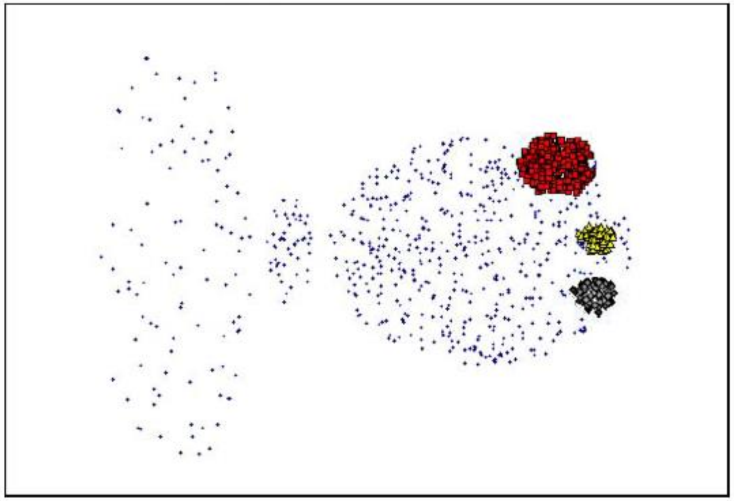
- Robust to noise
- Can detect arbitrarily-shaped clusters

When DBSCAN does not work well





(MinPts=4, Eps=9.92).



(MinPts=4, Eps=9.75)

- Cannot handle varying density
- Can detect arbitrarily-shaped clusters

Take-home Messages

- The basic idea of density-based clustering
- The two important parameters and the definitions of neighborhood and density in DBSCAN
- Core, border and outlier points
- DBSCAN algorithm
- DBSCAN's pros and cons