



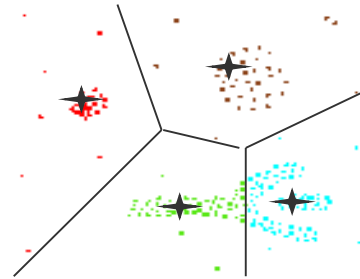
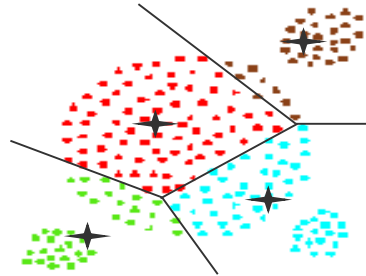
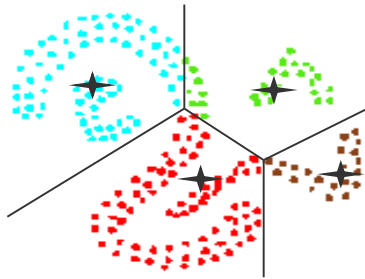
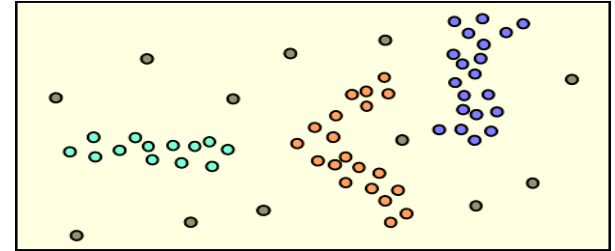
Big Data Analytics

– Week 3: Density-Based Clustering –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

Density-Based Clustering

- Basic Idea:
 - Clusters are dense regions in the data space, separated by regions of lower object density
- Why Density-Based Clustering?



Results of a k -medoid algorithm for $k=4$

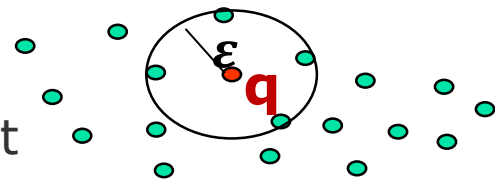
- Different density-based approaches exist (see Textbooks & Papers)
Here we discuss the ideas underlying the DBSCAN algorithm

Density-Based Clustering: Basic Concept

- Intuition for the formalization of the basic idea
 - For any point in a cluster, the local point density around that point has to exceed some threshold
 - The set of points from one cluster is spatially connected
- Local point density at a point q defined by two parameters
 - **ε -radius** for the neighborhood of point q :
 $N_\varepsilon(q) := \{p \in D \mid \text{dist}(p, q) \leq \varepsilon\}$ *(contains q itself !)*
 - **MinPts** – minimum number of points in the given neighborhood $N_\varepsilon(q)$
- q is called a **core object** (or core point) w.r.t. ε , MinPts if

$$| N_\varepsilon(q) | \geq \text{MinPts}$$

$\text{MinPts} = 5 \rightarrow \mathbf{q}$ is a core object



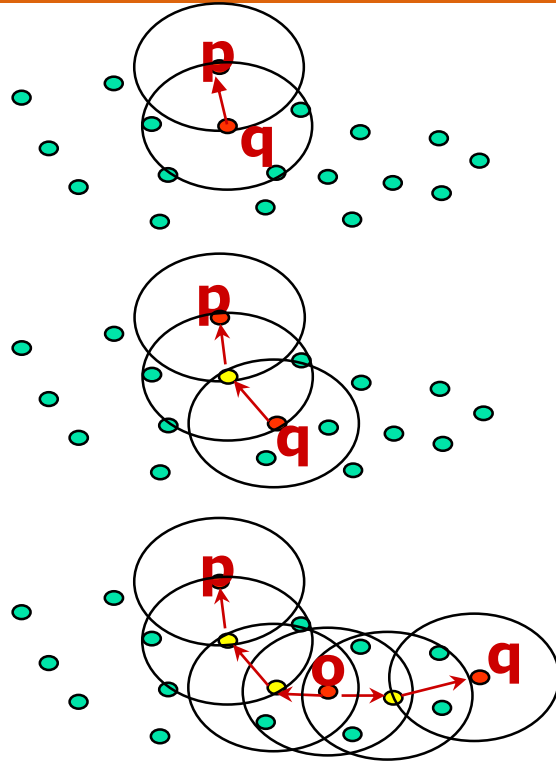
Density-Based Clustering: Basic Definitions

p **directly density-reachable** from q w.r.t. ε , $MinPts$ if

- 1) $p \in N_\varepsilon(q)$ and
- 2) q is a core object w.r.t. ε , $MinPts$

density-reachable: transitive closure of *directly* density-reachable

p is **density-connected** to a point q w.r.t. ε , $MinPts$ if there is a point o such that both, p and q are density-reachable from o w.r.t. ε , $MinPts$.



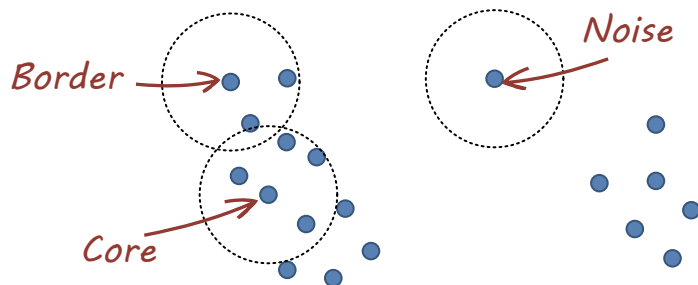
Density-Based Clustering: Basic Definitions

Density-Based Cluster: non-empty subset S of database D satisfying:

- 1) **Maximality:** if p is in S and q is density-reachable from p then q is in S
- 2) **Connectivity:** each object in S is density-connected to all other objects in S

Density-Based Clustering of a database $D : \{S_1, \dots, S_n; N\}$ where

- $S_1, \dots, S_n$: all density-based clusters in the database D
- $N = D \setminus \{S_1 \cup \dots \cup S_n\}$ is called the **noise** (objects not in any cluster)



$$\varepsilon = 1.0$$

$$MinPts = 5$$

Density-Based Clustering: DBSCAN Algorithm

Ester M., Kriegel H.-P., Sander J., Xu X.: „A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise“, *In KDD 1996*, pp. 226–231.

Density Based Spatial Clustering of Applications with Noise

Basic Theorem:

- Each object in a density-based cluster C is density-reachable from any of its core-objects
- Nothing else is density-reachable from core objects.

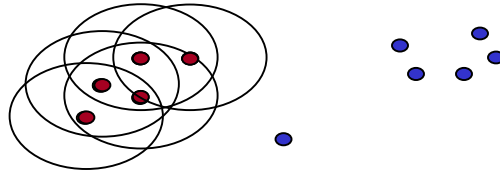
```
for each  $o \in D$  do
    if  $o$  is not yet classified then
        if  $o$  is a core-object then
            collect all objects density-reachable from  $o$ 
            and assign them to a new cluster.
        else
            assign  $o$  to NOISE
```

- density-reachable objects are collected by performing successive ε -neighborhood queries.

DBSCAN Algorithm: Example

Parameter

- $\varepsilon = 2.0$
- $MinPts = 3$



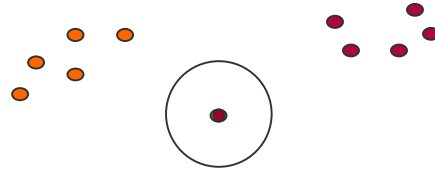
```

for each  $o \in D$  do
    if  $o$  is not yet classified then
        if  $o$  is a core-object then
            collect all objects density-reachable from  $o$ 
            and assign them to a new cluster.
        else
            assign  $o$  to NOISE
    
```

DBSCAN Algorithm: Example

Parameter

- $\varepsilon = 2.0$
- $MinPts = 3$



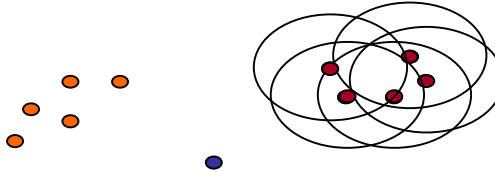
```

for each  $o \in D$  do
  if  $o$  is not yet classified then
    if  $o$  is a core-object then
      collect all objects density-reachable from  $o$ 
      and assign them to a new cluster.
    else
      assign  $o$  to NOISE
  
```

DBSCAN Algorithm: Example

Parameter

- $\varepsilon = 2.0$
- $MinPts = 3$

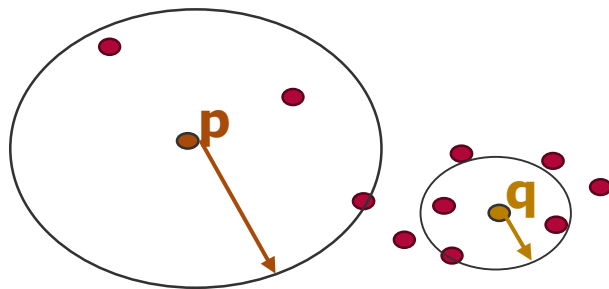


```

for each  $o \in D$  do
    if  $o$  is not yet classified then
        if  $o$  is a core-object then
            collect all objects density-reachable from  $o$ 
            and assign them to a new cluster.
        else
            assign  $o$  to NOISE
    
```

Determining the Parameters ε and MinPts

- Cluster: Point density higher than specified by ε and *MinPts*
- Idea: use the point density of the least dense cluster in the data set as parameters – but how to determine this?
- Heuristic: look at the distances to the k -nearest neighbors



4-distance(p) : 

4-distance(q) : 

- Function $k\text{-distance}(p)$: distance from p to its k -nearest neighbor
- $k\text{-distance plot}$: k -distances of all objects, sorted in decreasing order

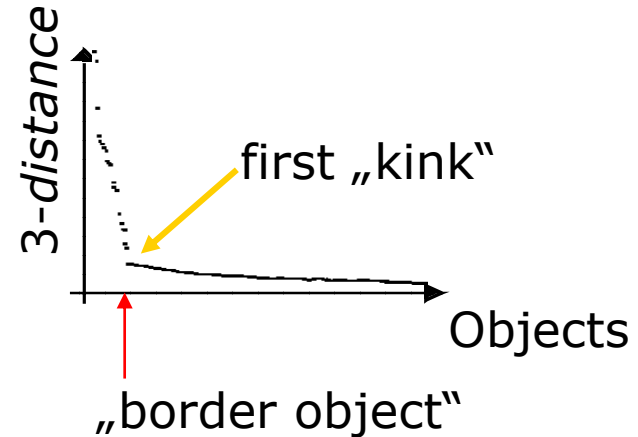
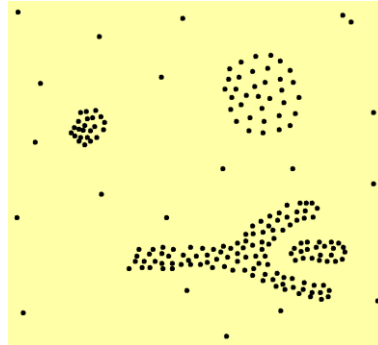
Determining the Parameters ε and MinPts

Heuristic method:

- Fix a value for *MinPts*
- (default: $2 \times d - 1$, d = dimension of data space)
- User selects "border object" o from the *MinPts-distance* plot;
 ε is set to *MinPts-distance*(o)

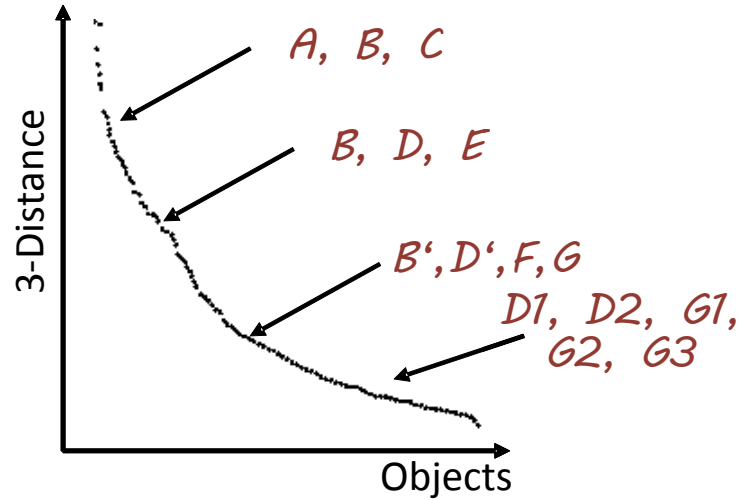
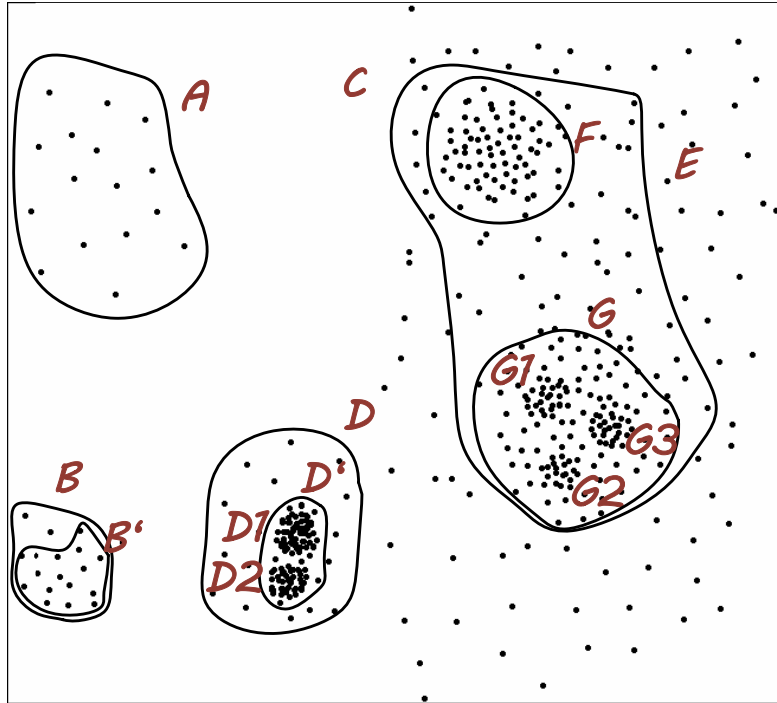
Example *k*-distance plot

- 1 dim = 2 $\rightarrow$ MinPts = 3
- 2 Identify border object
- 3 Set ε



Determining the Parameters ε and MinPts

Problematic example



Density-Based Clustering: Discussion

Advantages

- Clusters can have arbitrary shape and size, i.e. clusters are not restricted to have convex shapes
- Number of clusters is determined automatically
- Can separate clusters from surrounding noise
- Can be supported by spatial index structures
- Complexity: N_ϵ -query: $O(n)$ DBSCAN: $O(n^2)$

Disadvantages

- Input parameters may be difficult to determine
- In some situations very sensitive to input parameter setting