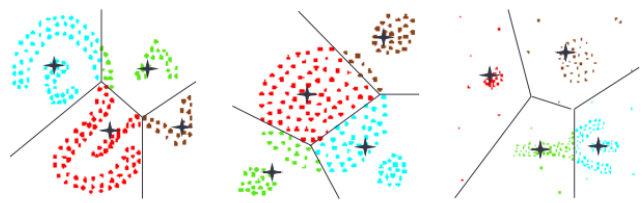
Cluster Analysis

- Density Based Clustering
  - DBSCAN
  - OPTICS

Density Based Clustering

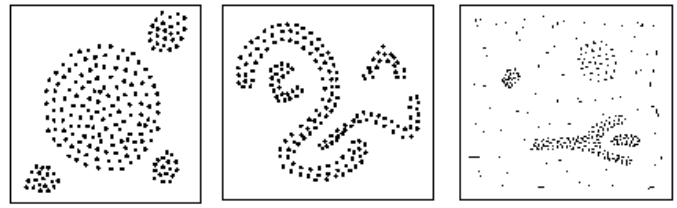
- Results of k-means algorithm for k = 4



- The result is not satisfiable!

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.



3

Density Based Clustering

- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - Need density parameters
  - One scan
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)

4

Density Concepts

- Two global parameters:
  - ***Eps***: Maximum radius of the neighbourhood
  - **MinPts**: Minimum number of points in an Eps-neighbourhood of that point
- 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.

5

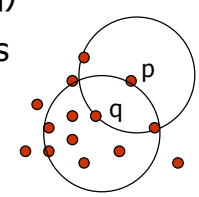
Density Concepts

- Eps-Neighborhood – Objects within a radius of *Eps* from an object.

  $N_{Eps}(p)$: {q belongs to D | dist(p,q) ≤ Eps}
- An object p is directly density-reachable from object q, if q is a core object and p is in q's Eps-neighborhood.

  1) p belongs to N_{Eps} (q)

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



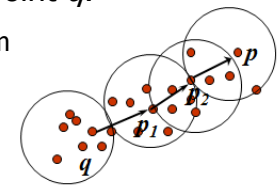
MinPts = 5

Eps = 1 cm

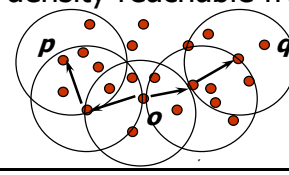
6

Density Concepts

- A point *p* is density-reachable from a point *q*.
  - A point p is directly density-reachable from
  - p2 is directly density-reachable from p1;
  - p1 is directly density-reachable from q;
  - p←p2←p1←q form a chain



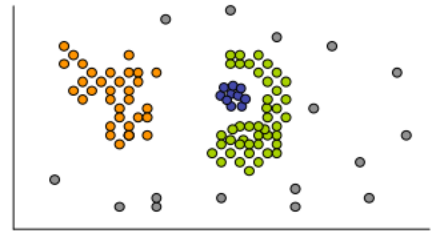
- A pair of points p and q are density-connected if they are commonly density-reachable from a point o.



7

Density Concepts

- A cluster is a set of **density-connected** objects which is maximal with respect to density reachability.
- Noise is the set of objects not contained in any cluster.



8

DBSCAN

- Density Based Spatial Clustering of Applications with Noise
- Proposed by Ester, Kriegel, Sander, and Xu (KDD96)
- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points.
- Discovers clusters of arbitrary shape in spatial databases with noise.

9

DBSCAN

- Given a data set D, parameter Eps and MinPts,
- A cluster C is a subset of D satisfying two criteria:
  - **Maximality:**

    ∀ p, q if p ∈ C and if q is density-reachable from p, then also q ∈ C
  - **Connectivity:**

    ∀ p, q ∈ C, p and q are density-connected

10

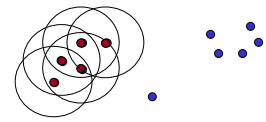
DBSCAN: The Algorithm

- Arbitrary select a point ***p***
- Retrieve all points density-reachable from ***p***. If ***p*** is a core point, a cluster is formed.
- If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

11

DBSCAN Algorithm: Example

- Parameter
  - *Eps*= 2 cm
  - *MinPts* = 3

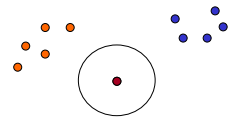


for each $o \in D$ **do**
  if o is not yet classified **then**
    if o is a core-object **then**
      collect all objects density-reachable from o
      and assign them to a new cluster.
    else
      assign o to NOISE

12

DBSCAN Algorithm: Example

- Parameter
  - *Eps* = 2 cm
  - *MinPts* = 3

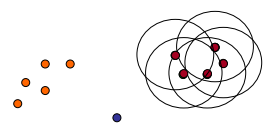


```
for each o ∈ D do
    if o is not yet classified then
        if o is a core-object then
            collect all objects density-reachable from o
            and assign them to a new cluster.
        else
            assign o to NOISE
```

13

DBSCAN Algorithm: Example

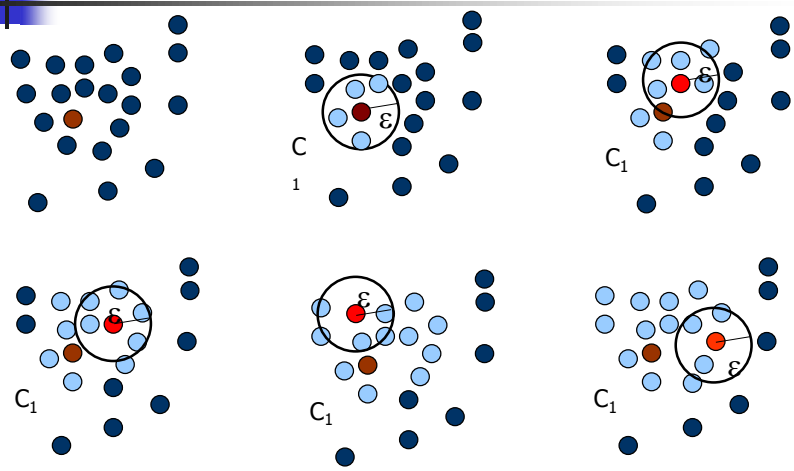
- Parameter
  - *Eps* = 2 cm
  - *MinPts* = 3



```
for each o ∈ 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
```

14

DBSCAN Algorithm: Example



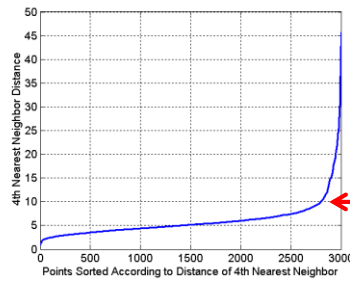
15

Determining Eps and MinPts

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbor at farther distance
- So, plot sorted distance of every point to its k^{th} nearest neighbor.
- Find the distance d where there is a "knee" in the curve
  - Eps = d, MinPts = k

16

Determining Eps and MinPts



Eps ~ 7-10
MinPts = 4

17

Advantages - Disadvantages

- Advantages
  - Clusters can have arbitrary shape and size
  - Number of clusters is determined automatically
  - Can separate clusters from surrounding noise
  - Can be supported by spatial index structures
- Disadvantages
  - Input parameters may be difficult to determine
  - In some situations very sensitive to input parameter setting

18

OPTICS

- DBSCAN
  - Input parameter – hard to determine.
  - Algorithm very sensitive to input parameters.
- OPTICS – Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
  - Based on DBSCAN.
  - Does not produce clusters explicitly.
  - Can be represented graphically or using visualization techniques

19

Core and Reachability Distance

- Parameters: "generating" distance , fixed value *MinPts*
- $core\text{-}distance_{Eps,MinPts}(o)$

  "smallest distance such that *o* is a core object"
- $reachability\text{-}distance_{Eps,MinPts}(p, o)$

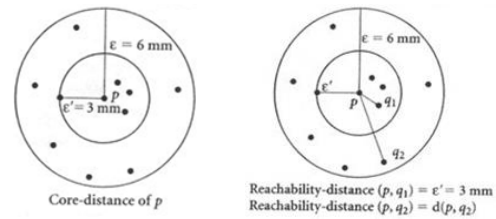
  "smallest distance such that *p* is *directly* density-reachable from *o*"

20

Core and Reachability Distance

- The core distance of p is the distance ε'.
- The reachability distance of q1 with respect to p is the core distance of p (ε'=3mm).
- The reachability distance of q2 with respect to p is the distance between p and q2.

(MinPts=5)



21

OPTICS

- The OPTICS algorithm finds clusters using the following steps:
  - Create an ordering of the objects in a database, storing the core-distance and a reachability distance for each objects.
  - Based on ordering information produced by OPTICS, use another algorithm to extract clusters.
  - Extract density-based clusters with respect to ant distance ε' that is smaller than the distance ε used in generating the order.

22

OPTICS: The Algorithm

- Begin with arbitrary object from the input database as the current object, p.
- It retrieves the ε-neighborhood of p, determines the core-distance and sets the reachability-distance to undefined.
- The current object, p, is then written to output.
- If p is not a core object,
  - OPTICS simply moves on to the next objext in the OrderSeeds list.

23

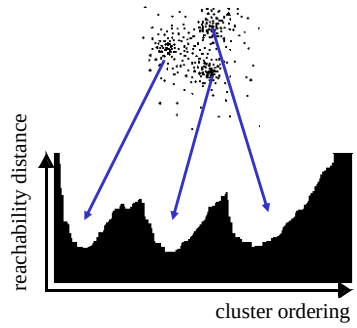
OPTICS: The Algorithm

- If p is a core object,
  - Then for each object, q, in the ε-neighborhood of p,
    - OPTICS updates its reachability-distance from p
    - And inserts q into OrderSeeds if q has not yet been processed.

24

OPTICS

- Represents the density-based clustering structure
- Easy to analyze
- Independent of the dimension of the data



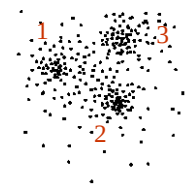
25

OPTICS

- Relatively insensitive to parameter settings

MinPts = 10, ε = 10

MinPts = 10, **ε = 5**



26