CS37300: Data Mining and Machine Learning

Nov 15, 2023

Density-based Clustering

Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

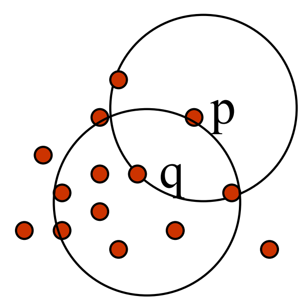
Density Concepts

- Core object (CO): an object with at least ‘M’ objects within a radius ‘ε-neighborhood’
- Directly density reachable (DDR): x is CO, y is in x’s ‘ε-neighborhood’
- Density reachable–there exists a chain of DDR objects from x to y
- Density based cluster: density connected objects w.r.t. reachability

Density-Based Clustering: Background

- Two parameters***:***
  - ***ε***: Maximum radius of the neighborhood
  - ***MinPts***: Minimum number of points in an ε-neighborhood of that point
- ***$N_\varepsilon(p)$***: ***{q belongs to D | dist(p,q) <= ε}***
- Directly density-reachable**:** A point ***p*** is directly density-reachable from a point ***q*** wrt. ***ε***, ***MinPts*** if
  - 1) ***p*** belongs to ***$N_\varepsilon(q)$***
  - 2) core point condition:

    $|N_\varepsilon(q)|$ >= ***MinPts***

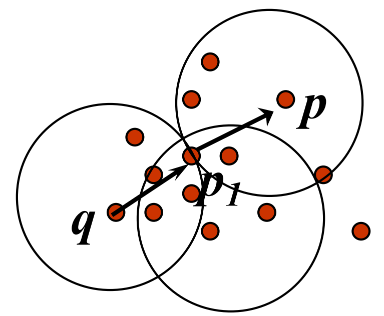


MinPts = 5

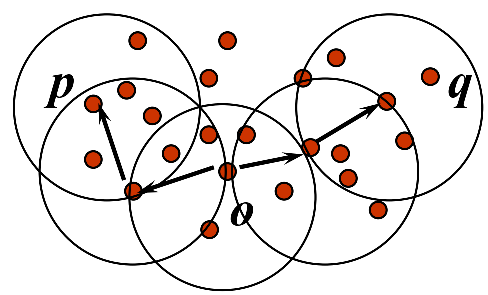
ε = 1 cm

Density-Based Clustering: Background

- Density-reachable:
  - A point *p* is density-reachable from a point *q* wrt. ε, *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

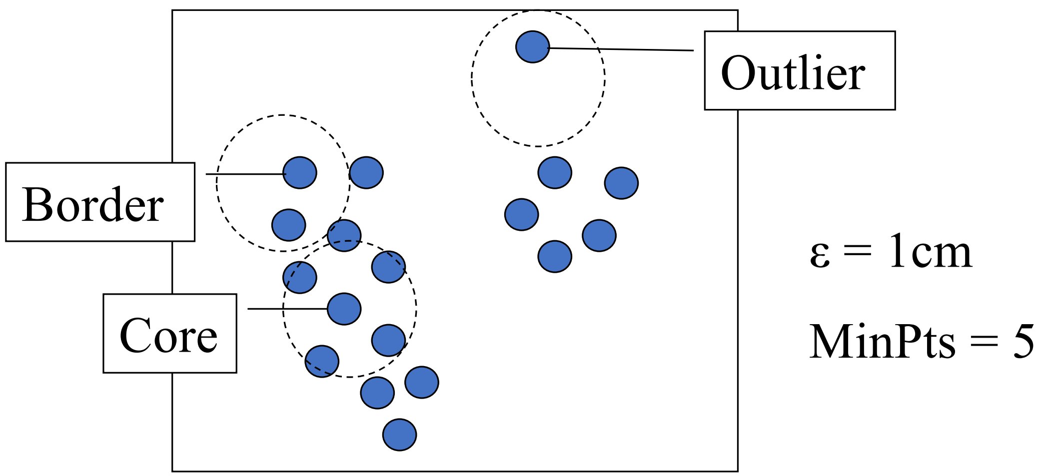


- Density-connected
  - A point *p* is density-connected to a point *q* wrt. ε, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o.*



DBSCAN: Density Based Spatial Clustering of Applications with Noise

- 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



DBSCAN: The Algorithm

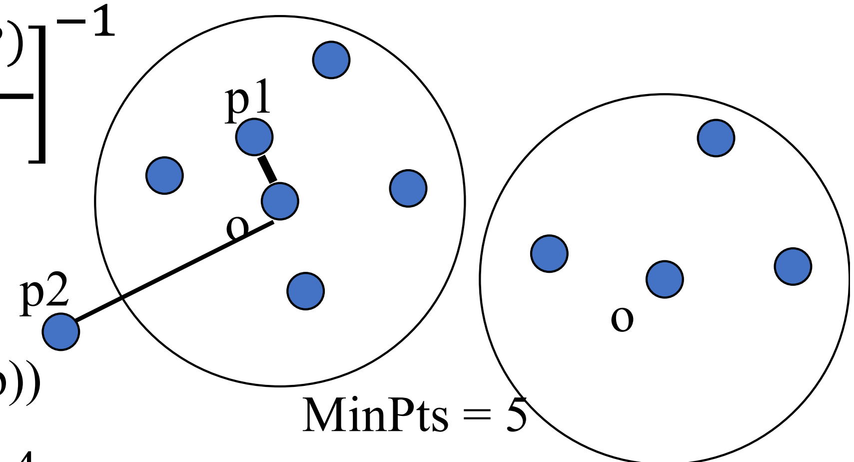
- Arbitrary select a point ***p***
- Retrieve all points directly density-reachable from ***p*** wrt **ε** and **MinPts**.
- If ***p*** is a core point, a cluster is formed.
- If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.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

Reachability distance (RD) and density(D)

- K-distance(A) = distance from A to its k-th closest neighbor
- RD of A from B:

$$RD_k\ (A, B) = \max\{kdistance(A), distance(A, B)\}$$

- Local Density of A = $\left[\frac{\sum_{B\in N_k(A)} RD_k(A,B)}{|N_k(B)|}\right]^{-1}$



Max (k-distance (o), d (o, p))

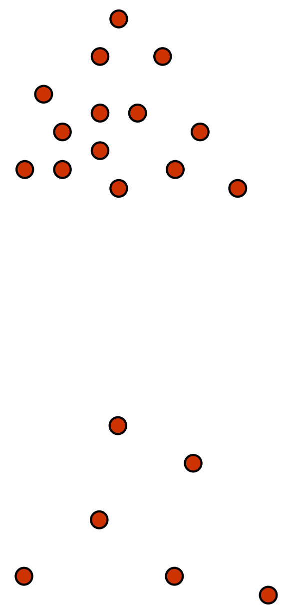
r(p1, o) = 2.8cm. r(p2,o) = 4cm

MinPts = 5

ε = 3 cm

Local Outlier Factor (LOF)

- Challenge in DBScan: Setting ε
  - What is the right neighborhood size?
  - Is it even constant across the data?
- LOF compares the ratio of the average of the reachability densities of its neighbors to its own reachability density
- Typically used for anomaly detection, but same idea can be used for clustering

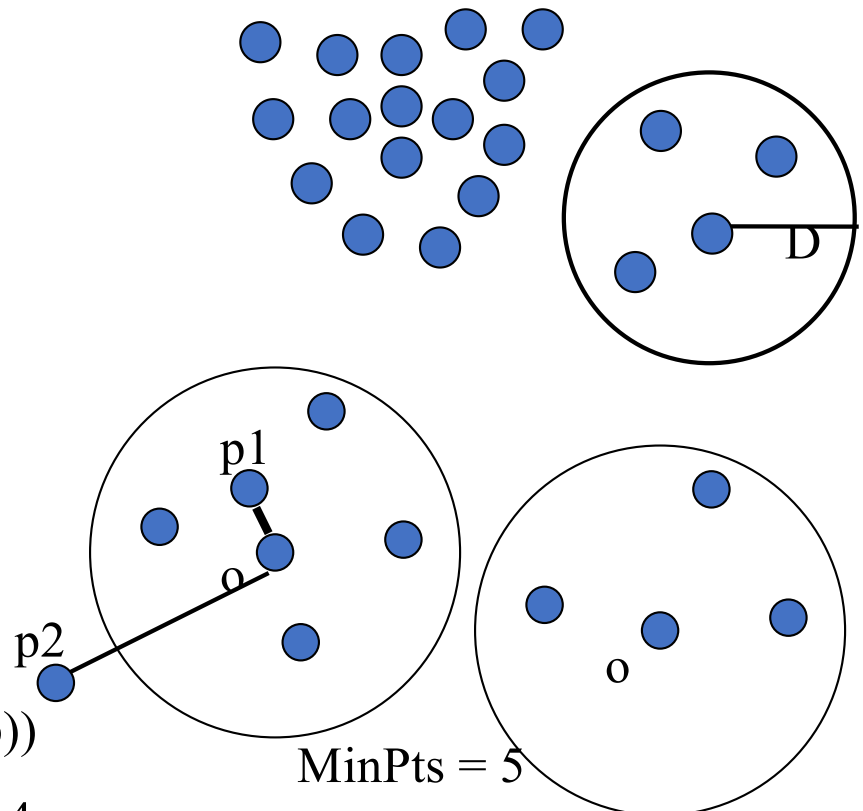


OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
  - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
  - Does not produce an explicit single clustering of data
  - Produces a special order of the database wrt its density-based clustering structure
  - This cluster-ordering contains info equivalent to the density-based clusterings corresponding to a broad range of parameter settings
  - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
  - Can be represented graphically or using visualization techniques

OPTICS: Extension from DBSCAN

- Index-based:
  - d = number of dimensions
  - N = number of data points
- Complexity: O(dN^2)
- Core or k-Distance
- Reachability Distance

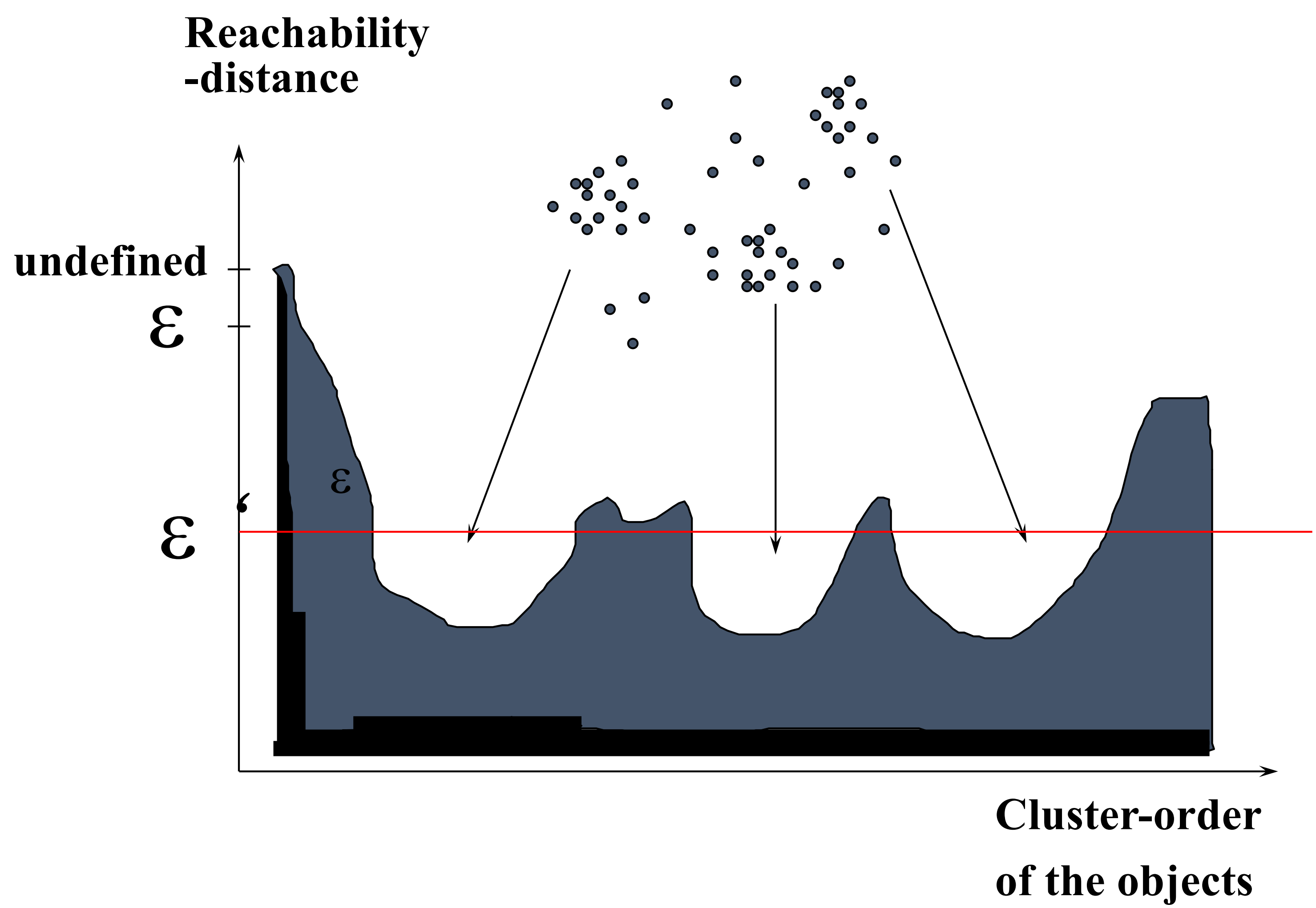


Max (k-distance (o), d (o, p))

r(p1, o) = 2.8cm. r(p2,o) = 4cm

MinPts = 5

ε = 3 cm


Reachability
-distance
undefined
ε
ε'
ε
Cluster-order
of the objects

DENCLUE: Using density functions

- DENsity-based CLUstEring by Hinneburg & Keim (KDD'98)
- Major features
  - Solid mathematical foundation
  - Good for data sets with large amounts of noise
  - Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
  - Significant faster than existing algorithm (faster than DBSCAN by a factor of up to 45)
  - But needs a large number of parameters

Denclue: Technical Essence

- Uses grid cells but only keeps information about grid cells that do actually contain data points and manages these cells in a tree-based access structure.
- Influence function: describes the impact of a data point within its neighborhood.
- Overall density of the data space can be calculated as the sum of the influence function of all data points.
- Clusters can be determined mathematically by identifying density attractors.
- Density attractors are local maximal of the overall density function.

Gradient: The steepness of a slope

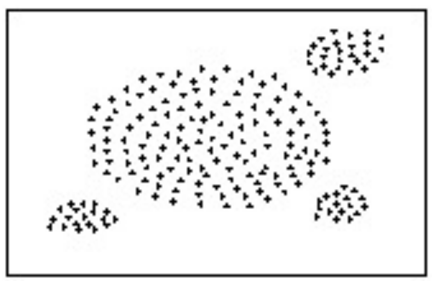
- Example

$$f_{Gaussian}(x,y) = e^{-\frac{d(x,y)^2}{2\sigma^2}}$$

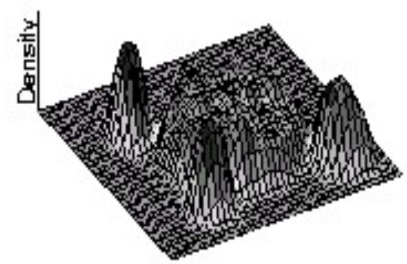
$$f_{Gaussian}^{D}(x) = \sum_{i=1}^{N} e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

$$\nabla f_{Gaussian}^{D}(x,x_i) = \sum_{i=1}^{N} (x_i - x) \cdot e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

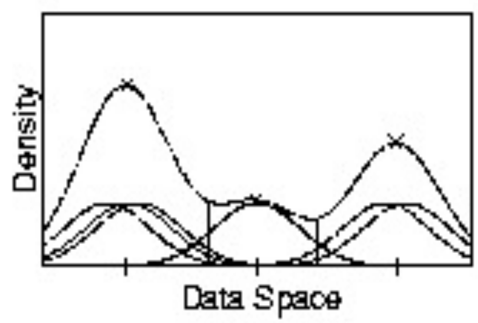
Density Attractor



(a) Data Set



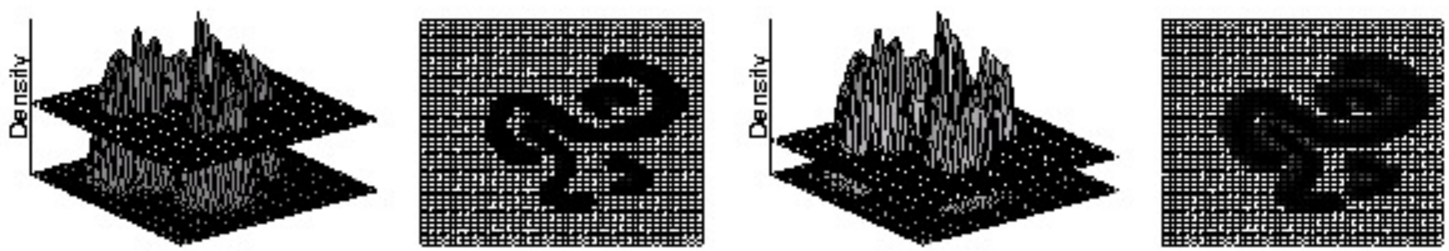
(c) Gaussian



Center-Defined and Arbitrary

(a) $\sigma = 0.2$ (b) $\sigma = 0.6$ (d) $\sigma = 1.5$

Figure 3: Example of Center-Defined Clusters for different σ



(a) $\xi = 2$ (b) $\xi = 2$ (c) $\xi = 1$ (d) $\xi = 1$

Figure 4: Example of Arbitray-Shape Clusters for different ξ