CS685: Data Mining
Density-Based Clustering Methods

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs685/

1st semester, 2020-21

Mon 1030-1200 (online)

Density-Based Clustering Methods

- Partitioning-based and hierarchical clustering methods mostly work on distances
- Hence, they tend to find spherical or convex clusters

Density-Based Clustering Methods

- Partitioning-based and hierarchical clustering methods mostly work on distances
- Hence, they tend to find spherical or convex clusters
- Points in the same cluster tend to have similar neighborhood densities
- Also, outliers are generally in sparse regions
- Density-based clustering methods aim to exploit these properties

Density-Based Clustering Methods

- Partitioning-based and hierarchical clustering methods mostly work on distances
- Hence, they tend to find spherical or convex clusters
- Points in the same cluster tend to have similar neighborhood densities
- Also, outliers are generally in sparse regions
- Density-based clustering methods aim to exploit these properties
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Ordering Points to Identify the Clustering Structure (OPTICS)
- Density-based Clustering (DENCLUE)

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Density of a point is determined using an ϵ-neighborhood
- Since the radius is fixed, it is simply the number of points

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Density of a point is determined using an ϵ-neighborhood
- Since the radius is fixed, it is simply the number of points
- Based on the density, a point can be
    - Core point: If density exceeds a threshold τ
    - Border point: If density $< \tau$, but it is in the ϵ-neighborhood of a core point
    - Noise point: If it is neither a core point nor a border point

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Density of a point is determined using an ϵ-neighborhood
- Since the radius is fixed, it is simply the number of points
- Based on the density, a point can be
  - Core point: If density exceeds a threshold τ
  - Border point: If density $< \tau$, but it is in the ϵ-neighborhood of a core point
  - Noise point: If it is neither a core point nor a border point
- DBSCAN does not use border points directly

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Density of a point is determined using an ϵ-neighborhood
- Since the radius is fixed, it is simply the number of points
- Based on the density, a point can be
  - Core point: If density exceeds a threshold τ
  - Border point: If density $< \tau$, but it is in the ϵ-neighborhood of a core point
  - Noise point: If it is neither a core point nor a border point
- DBSCAN does not use border points directly
- It also removes all noise points as outliers

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Density of a point is determined using an ϵ-neighborhood
- Since the radius is fixed, it is simply the number of points
- Based on the density, a point can be
  - Core point: If density exceeds a threshold τ
  - Border point: If density $< \tau$, but it is in the ϵ-neighborhood of a core point
  - Noise point: If it is neither a core point nor a border point
- DBSCAN does not use border points directly
- It also removes all noise points as outliers
- Uses notions of density-reachability and density-connectivity

Density-Reachability and Density-Connectivity

- A point p is directly density-reachable from q if
  - q is a core point
  - p is within the ϵ-neighborhood of q

Density-Reachability and Density-Connectivity

- A point p is directly density-reachable from q if
  - q is a core point
  - p is within the ϵ-neighborhood of q
- A point p is density-reachable from q if
  - There is a chain of points $p_1, \ldots, p_n$ where $p_1 = q$ and $p_n = p$
  - Each p_{i+1} is directly density-reachable from p_i

Density-Reachability and Density-Connectivity

- A point p is directly density-reachable from q if
  - q is a core point
  - p is within the ϵ-neighborhood of q
- A point p is density-reachable from q if
  - There is a chain of points $p_1, \ldots, p_n$ where $p_1 = q$ and $p_n = p$
  - Each p_{i+1} is directly density-reachable from p_i
- Density-reachability is *not symmetric*
- Directly density-reachability is *not symmetric* either

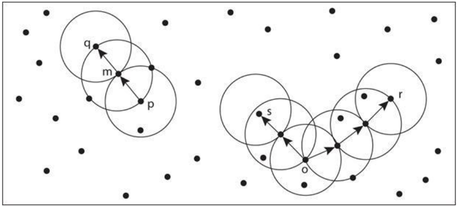
Density-Reachability and Density-Connectivity

- A point p is directly density-reachable from q if
  - q is a core point
  - p is within the ϵ-neighborhood of q
- A point p is density-reachable from q if
  - There is a chain of points $p_1, \ldots, p_n$ where $p_1 = q$ and $p_n = p$
  - Each p_{i+1} is directly density-reachable from p_i
- Density-reachability is *not symmetric*
- Directly density-reachability is *not symmetric* either
- Two points p_1 and p_2 are density-connected if
  - There is a point q from which both p_1 and p_2 are density-reachable

Density-Reachability and Density-Connectivity

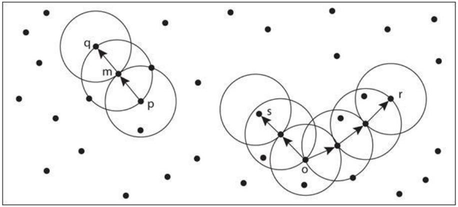
- A point p is directly density-reachable from q if
  - q is a core point
  - p is within the ϵ-neighborhood of q
- A point p is density-reachable from q if
  - There is a chain of points $p_1, \dots, p_n$ where $p_1 = q$ and $p_n = p$
  - Each p_{i+1} is directly density-reachable from p_i
- Density-reachability is *not symmetric*
- Directly density-reachability is *not symmetric* either
- Two points p_1 and p_2 are density-connected if
  - There is a point q from which both p_1 and p_2 are density-reachable
- Density-connectivity is *symmetric*

Example



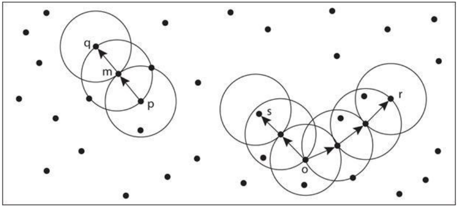
- Suppose $\tau = 3$
- Core points are

Example



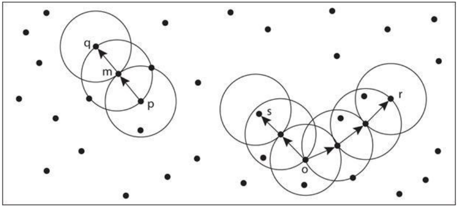
- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p?

Example



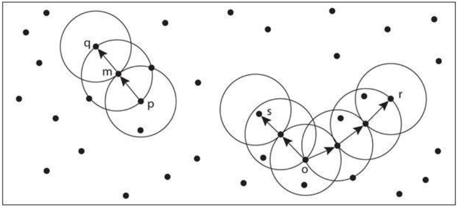
- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m?

Example



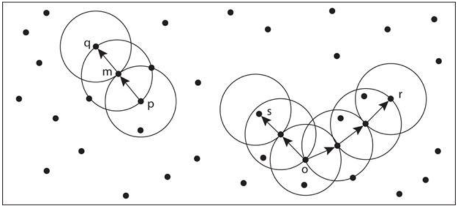
- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m? Yes
- Is q directly density-reachable from m?

Example



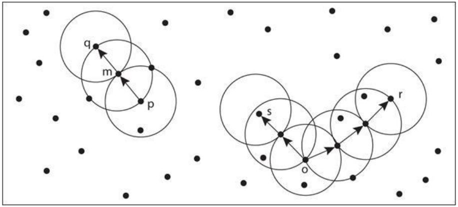
- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m? Yes
- Is q directly density-reachable from m? Yes
- Is m directly density-reachable from q?

Example



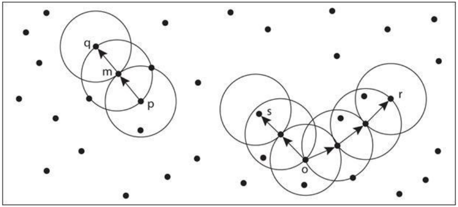
- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m? Yes
- Is q directly density-reachable from m? Yes
- Is m directly density-reachable from q? No
- Is q density-reachable from p?

Example



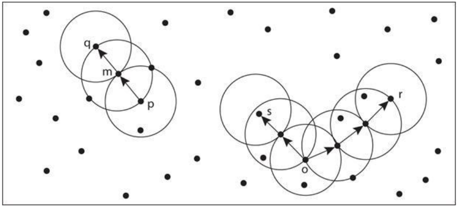
- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m? Yes
- Is q directly density-reachable from m? Yes
- Is m directly density-reachable from q? No
- Is q density-reachable from p? Yes
- Is p density-reachable from q?

Example



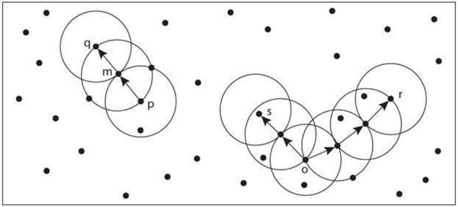
- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m? Yes
- Is q directly density-reachable from m? Yes
- Is m directly density-reachable from q? No
- Is q density-reachable from p? Yes
- Is p density-reachable from q? No
- Are s and r density-connected?

Example



- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m? Yes
- Is q directly density-reachable from m? Yes
- Is m directly density-reachable from q? No
- Is q density-reachable from p? Yes
- Is p density-reachable from q? No
- Are s and r density-connected? Yes
- Are p and q density-connected?

Example



- Suppose $\tau = 3$
- Core points are p, m, o, r but not q, s
- Is m directly density-reachable from p? Yes
- Is p directly density-reachable from m? Yes
- Is q directly density-reachable from m? Yes
- Is m directly density-reachable from q? No
- Is q density-reachable from p? Yes
- Is p density-reachable from q? No
- Are s and r density-connected? Yes
- Are p and q density-connected? Yes

Density-Based Cluster

- A density-based cluster is the closure of density-connected points
- A set of points C is a cluster if
  - For any two points $p, q \in C$, p and q are density-connected
  - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected

Density-Based Cluster

- A density-based cluster is the closure of density-connected points
- A set of points C is a cluster if
  - For any two points $p, q \in C$, p and q are density-connected
  - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected
- Algorithm
  - Arbitrarily select a point p

Density-Based Cluster

- A density-based cluster is the closure of density-connected points
- A set of points C is a cluster if
  - For any two points $p, q \in C$, p and q are density-connected
  - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected
- Algorithm
  - Arbitrarily select a point p
  - If p has $< \tau$ neighbors, it is marked as noise
  - Otherwise, p and all its ϵ-neighbors are added to the cluster
  - Then, neighbors of these points are checked and so on till the cluster cannot be expanded any more

Density-Based Cluster

- A density-based cluster is the closure of density-connected points
- A set of points C is a cluster if
  - For any two points $p, q \in C$, p and q are density-connected
  - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected
- Algorithm
  - Arbitrarily select a point p
  - If p has $< \tau$ neighbors, it is marked as noise
  - Otherwise, p and all its ϵ-neighbors are added to the cluster
  - Then, neighbors of these points are checked and so on till the cluster cannot be expanded any more
  - The next cluster is started by selecting another non-processed point arbitrarily

Density-Based Cluster

- A density-based cluster is the closure of density-connected points
- A set of points C is a cluster if
  - For any two points $p, q \in C$, p and q are density-connected
  - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected
- Algorithm
  - Arbitrarily select a point p
  - If p has $< \tau$ neighbors, it is marked as noise
  - Otherwise, p and all its ϵ-neighbors are added to the cluster
  - Then, neighbors of these points are checked and so on till the cluster cannot be expanded any more
  - The next cluster is started by selecting another non-processed point arbitrarily
- Time taken is $O(n^2)$
- Can be made $O(n \log n)$ by using efficient range search

Density-Based Cluster

- A density-based cluster is the closure of density-connected points
- A set of points C is a cluster if
  - For any two points $p, q \in C$, p and q are density-connected
  - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected
- Algorithm
  - Arbitrarily select a point p
  - If p has $< \tau$ neighbors, it is marked as noise
  - Otherwise, p and all its ϵ-neighbors are added to the cluster
  - Then, neighbors of these points are checked and so on till the cluster cannot be expanded any more
  - The next cluster is started by selecting another non-processed point arbitrarily
- Time taken is $O(n^2)$
- Can be made $O(n \log n)$ by using efficient range search
- Assumes clusters have similar densities

Density-Based Cluster

- A density-based cluster is the closure of density-connected points
- A set of points C is a cluster if
  - For any two points $p, q \in C$, p and q are density-connected
  - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected
- Algorithm
  - Arbitrarily select a point p
  - If p has $< \tau$ neighbors, it is marked as noise
  - Otherwise, p and all its ϵ-neighbors are added to the cluster
  - Then, neighbors of these points are checked and so on till the cluster cannot be expanded any more
  - The next cluster is started by selecting another non-processed point arbitrarily
- Time taken is $O(n^2)$
- Can be made $O(n \log n)$ by using efficient range search
- Assumes clusters have similar densities
- Depends heavily on the parameters ϵ and τ

OPTICS

- Ordering Points To Identify the Clustering Structure (OPTICS)
- More a data ordering algorithm than a clustering method
- Tries to overcome the difficulty of choosing ϵ in DBSCAN
- Produces a cluster ordering of the data
- A linear order that represents the density-based clustering structure of the data

OPTICS

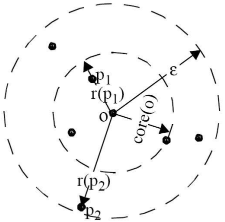
- Ordering Points To Identify the Clustering Structure (OPTICS)
- More a data ordering algorithm than a clustering method
- Tries to overcome the difficulty of choosing ϵ in DBSCAN
- Produces a cluster ordering of the data
- A linear order that represents the density-based clustering structure of the data
- Uses notions of core distance and reachability distance

Core Distance and Reachability Distance

- The core distance of a point p is the *smallest* radius ϵ' such that the ϵ'-neighborhood of p contains τ points, i.e., p becomes a core point
  - If $\epsilon' > \epsilon$, it is considered *undefined*

Core Distance and Reachability Distance

- The core distance of a point p is the *smallest* radius ϵ' such that the ϵ'-neighborhood of p contains τ points, i.e., p becomes a core point
  - If $\epsilon' > \epsilon$, it is considered *undefined*
- The reachability distance of point p to point q is the *minimum* radius that makes q *directly density-reachable* from p
- p must become a core point
- It is, therefore, the *maximum* of *core distance* of p and distance of p to q
  - If p is not a core point with respect to ϵ, it is considered *undefined*

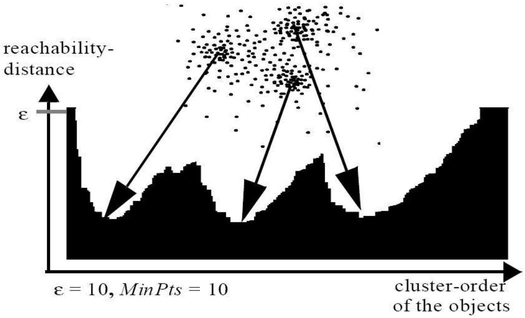


Reachability Plot

- OPTICS outputs a reachability plot
- For each point in the database, it plots its reachability distance to the *nearest* core point
- Bumps mark the boundaries of clusters
- Valleys denote the clusters
  - Deeper the valley, denser the cluster

Reachability Plot

- OPTICS outputs a reachability plot
- For each point in the database, it plots its reachability distance to the *nearest* core point
- Bumps mark the boundaries of clusters
- Valleys denote the clusters
  - Deeper the valley, denser the cluster



Algorithm

- Arbitrarily select a point p
- Find its core distance and set its reachability distance to undefined
- Insert p to a priority list L
- If p is not a core point, select next point from L (or dataset)
- Otherwise, pick q from p's neighborhood
- Update reachability distance of q and insert q in L
- Continue till all points are processed

Algorithm

- Arbitrarily select a point p
- Find its core distance and set its reachability distance to undefined
- Insert p to a priority list L
- If p is not a core point, select next point from L (or dataset)
- Otherwise, pick q from p's neighborhood
- Update reachability distance of q and insert q in L
- Continue till all points are processed
- Can show nested clusters as well

Algorithm

- Arbitrarily select a point p
- Find its core distance and set its reachability distance to undefined
- Insert p to a priority list L
- If p is not a core point, select next point from L (or dataset)
- Otherwise, pick q from p's neighborhood
- Update reachability distance of q and insert q in L
- Continue till all points are processed
- Can show nested clusters as well
- Time complexity is $O(n \log n)$

Algorithm

- Arbitrarily select a point p
- Find its core distance and set its reachability distance to undefined
- Insert p to a priority list L
- If p is not a core point, select next point from L (or dataset)
- Otherwise, pick q from p's neighborhood
- Update reachability distance of q and insert q in L
- Continue till all points are processed
- Can show nested clusters as well
- Time complexity is $O(n \log n)$
- Still uses parameters: τ, ϵ

DENCLUE

- DENsity-based CLUstEring (DENCLUE)
- General parameter-free clustering
- Tries to capture *natural* clusters in the data

DENCLUE

- DENsity-based CLUstEring (DENCLUE)
- General parameter-free clustering
- Tries to capture *natural* clusters in the data
- Density of a point is modeled as sum of influence functions associated with each data point
- Influence of a data point y on an arbitrary point x in the space is some function $f_y(x)$
- Results in an overall density function at every point in the space
- Density at a point x is $\sum_{y \in D} f_y(x)$

DENCLUE

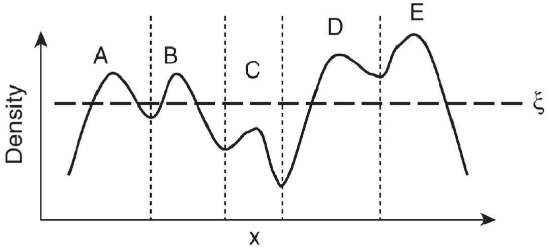
- DENsity-based CLUstEring (DENCLUE)
- General parameter-free clustering
- Tries to capture *natural* clusters in the data
- Density of a point is modeled as sum of influence functions associated with each data point
- Influence of a data point y on an arbitrary point x in the space is some function $f_y(x)$
- Results in an overall density function at every point in the space
- Density at a point x is $\sum_{y \in D} f_y(x)$
- Local peaks denote cluster centres
- Points are attracted towards the nearest peak

Thresholding

- Uses a minimum density threshold ξ
- If density at a peak is too low, it is noise
- If two peaks are connected with high density points, the corresponding clusters are merged

Thresholding

- Uses a minimum density threshold ξ
- If density at a peak is too low, it is noise
- If two peaks are connected with high density points, the corresponding clusters are merged
- Cluster corresponding to C is discarded
- Clusters A and B remain separated
- Clusters D and E get merged



Density Estimation

- Uses kernel density estimation
- Probability density of a point depends on its distance to other points
- Contribution of each point to overall density is measured by a kernel function that acts as the influence function

Density Estimation

- Uses kernel density estimation
- Probability density of a point depends on its distance to other points
- Contribution of each point to overall density is measured by a kernel function that acts as the influence function
- For iid points $x_1, \ldots, x_n$, kernel density approximation is

$$\hat{f}_h(x) = \frac{1}{n.\sigma} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\sigma}\right)$$

- σ is the smoothing factor and K is the kernel function

Density Estimation

- Uses kernel density estimation
- Probability density of a point depends on its distance to other points
- Contribution of each point to overall density is measured by a kernel function that acts as the influence function
- For iid points $x_1, \dots, x_n$, kernel density approximation is

$$\hat{f}_h(x) = \frac{1}{n.\sigma} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\sigma}\right)$$

- σ is the smoothing factor and K is the kernel function
- DENCLUE uses a Gaussian kernel

$$K\left(\frac{x - x_i}{\sigma}\right) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{(x - x_i)^2}{2\sigma^2}}$$

Density Estimation

- Uses kernel density estimation
- Probability density of a point depends on its distance to other points
- Contribution of each point to overall density is measured by a kernel function that acts as the influence function
- For iid points $x_1, \ldots, x_n$, kernel density approximation is

$$\hat{f}_h(x) = \frac{1}{n.\sigma} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\sigma}\right)$$

- σ is the smoothing factor and K is the kernel function
- DENCLUE uses a Gaussian kernel

$$K\left(\frac{x - x_i}{\sigma}\right) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

- A point x^* is a local density attractor if it is a local maximum of the estimated density function

Density Estimation

- Uses kernel density estimation
- Probability density of a point depends on its distance to other points
- Contribution of each point to overall density is measured by a kernel function that acts as the influence function
- For iid points $x_1, \ldots, x_n$, kernel density approximation is

$$\hat{f}_h(x) = \frac{1}{n.\sigma} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\sigma}\right)$$

- σ is the smoothing factor and K is the kernel function
- DENCLUE uses a Gaussian kernel

$$K\left(\frac{x - x_i}{\sigma}\right) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

- A point x^* is a local density attractor if it is a local maximum of the estimated density function
- Points are attracted towards local density attractors using *hill climbing*
- Uses the gradient of the Gaussian kernel

Discussion

- Very robust to noise

Discussion

- Very robust to noise
- Strong mathematical foundation

Discussion

- Very robust to noise
- Strong mathematical foundation
- Generalization of several clustering methods

Discussion

- Very robust to noise
- Strong mathematical foundation
- Generalization of several clustering methods
- Can be very slow

Discussion

- Very robust to noise
- Strong mathematical foundation
- Generalization of several clustering methods
- Can be very slow
- Density estimated only at actual data points

Discussion

- Very robust to noise
- Strong mathematical foundation
- Generalization of several clustering methods
- Can be very slow
- Density estimated only at actual data points
- Influence function constrained to a range
- May use grids where each point influences its own cell and the neighboring cells only