# Density-based methods

Clustering based on density (local cluster criterion), such as density-connected points. Continue growing a given cluster as long as the density (number of objects) in the neighborhood exceeds some threshold. Major features:

* Discover clusters of arbitrary shape – model clusters as dense regions separated by sparse regions
* Handle noise – can filter out noise or outliers
* One scan
* Needs density parameters (Eps & MinPts) as termination condition

DBSCAN and OPTICS

**Basic concepts**

Two parameters:

* Eps: maximum radius of the neighborhood. Neighborhood is the area around a point with radius Eps
* MinPts: Minimum number of points in an Eps-neighborhood of that point, i.e. minimum number of points in the area for the area to be dense

where

is a subset of data D contained in the Eps-neighborhood of q.

Core point:

q is a core point if contains at least MinPts number of points.

Directly density reachable:

A point p is directly density-reachable from a point q with respect to Eps & MinPts if:

1. p belongs to Eps-neighborhood of q, i.e:  
   p belongs to ,
2. (core-point condition)

An object p is directly density-reachable from q if and only if q is a core point and p belongs to Eps-neighborhood of q.

This is not a symmetric definition because of the core point condition. Only from core points can other points be directly density-reachable

Density-reachable:

A point p is density-reachable from a point q w.r.t Eps and MinPts if there is a chain pf points where and such that is directly density-reachable from

This is not a symmetric definition. p doesn’t have to be a core point.

Only core points can be mutually density-reachable

Density-connected:

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

This is a symmetric definition!

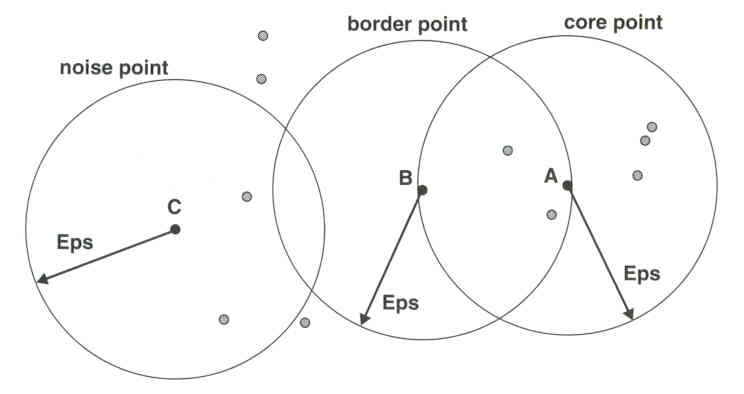
Density-based notion of a cluster:

A cluster is defined as a maximal set of density-connected points. A set of density-connected points which is maximal w.r.t density-reachability and where the noise is a set of points not contained in any cluster

**DBSCAN**

Relies on the density-based notion of a cluster. DBSCAN can find clusters of arbitrary shapes and filter out noise, since outliers are ignored. If we can’t reach a point with extreme values, it just leaces that point out.

But DBSCAN is very sensitive to parameters. For example, if Eps is bigger, more points will satisfy the core point condition.



Algorithm:

* Arbitrary select a point p
* Retrieve all points that are density-reachable from p w.r.t. Eps 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 data
* Process is continued until all points have been processed

**OPTICS**

Ordering Points To Identify the Clustering Structure – A cluster ordering method

OPTICS produced a special order of the data w.r.t its density-based clustering structure. This clustering-ordering contains information equivalent to the density-based clusterings corresponding to a broad range of parameter settings

Core distance:

p is a point in data D. Eps is a distance value. is the Eps-neighborhood of p.

Core distance of p w.r.t MinPts: smallest distance Eps’ between p and an object in its Eps-neighborhood such that p would be a core point for Eps’ &MinPts. Otherwise, undefined

MinPts-distance(p) – distance from p to its MinPts neighbor

Core distance of p: OR UNDEFINED if p is not a core point.

Reachability distance:

* p and o are points from data D.
* is the Eps-neighborhood of point o
* The reachability distance of p w.r.t o is if o is a core point. Otherwise undefined
* It’s the smallest distance such that p is directly density-reachable from o if o is a core point

Algorithm:

1. Select non-processed object o
2. Find neighbors (in the Eps-neighborhood of o)
   1. Compute core distance for o
   2. Write object o to ordered file and mark o as processed
   3. If o is not a core object, restart at step 1
   4. If o is a core object: Put neighors of o in Seedlist and order
      1. If neighbor n is not yet in Seedlist then add (n, reachability from o)
      2. Else if reachability from o<current reachability, then update reachability + order Seedlist w.r.t reachability
   5. Take new point from Seedlist with smallest reachability and restart at step 2

After running OPTICS, we run DBSCAN with the chosen parameters Eps. MinPts has to be the same in OPTICS and DBSCAN

**What is the relationship between DBSCAN & OPTICS?**

OPTICS makes use of the basic principles of DBSCAN by trying to give a graphical representation of reachability and Eps, while not being so sensitive as DBSCAN to the setting of Eps.

Often, OPTICS is used as a pre-processing step to DBSCAN to select an Eps for the actual performance of DBSCAN. MinPts has to be the same in both.

The data file is ordered in OPTICS to discover areas of high density. The following statistics are considered in OPTICS:

* Core-distance: Minimum value of Eps such that a point o is a core point.   
  -- which is the distance from point o to its MinPts neighbor. Only defined when o is a core point.
* Reachability of a point p from a core point o for the value of Eps so that p is density-reachable from o: - Only defined when o is a core point

OPTICS doesn’t really produce a clustering but it does a cluster analysis that identifies cluster structure. Then it is used as a pre-processing step for DBSCAN.