Clustering

DBSCAN: Density-Based Clustering Based on Connected Regions with High Density

- The density of an object o can be measured by the number of objects close to o. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) finds core objects, that is, objects that have dense neighborhoods.
- It connects core objects and their neighborhoods to form dense regions as clusters.
- A user-specified parameter $\varepsilon > 0$ is used to specify the radius of a neighborhood we consider for every object.
- The ε-neighborhood of an object o is the space within a radius centered at o.

DBSCAN: Density-Based Clustering Based on Connected Regions with High Density

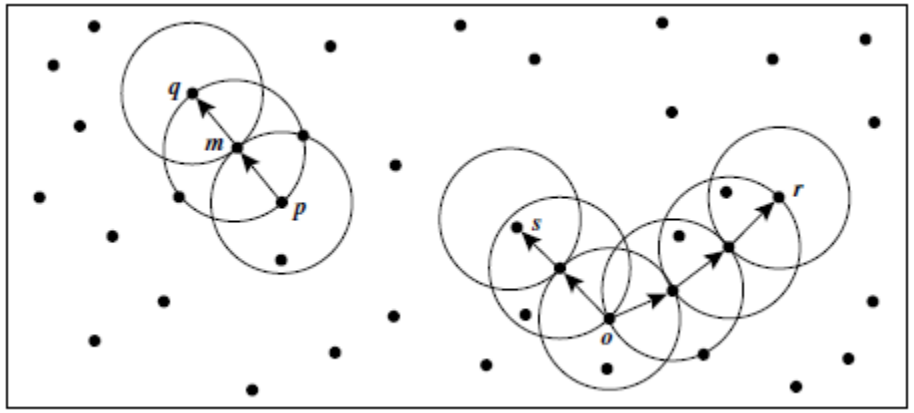
- Due to the fixed neighborhood size parameterized by ε, the density of a neighborhood can be measured simply by the number of objects in the neighborhood.
- To determine whether a neighborhood is dense or not, DBSCAN uses another user-specified parameter, **MinPts**, which specifies the density threshold of dense regions.
- An object is a core object if the ε -neighborhood of the object contains at least MinPts objects.
- Core objects are the pillars of dense regions.

DBSCAN: Density-Based Clustering Based on Connected Regions with High Density

- Given a set, D, of objects, we can identify all core objects with respect to the given parameters, ε and MinPts.
- The clustering task is therein reduced to using core objects and their neighborhoods to form dense regions, where the dense regions are clusters.
- For a core object **q** and an object **p**, we say that p is directly density-reachable from q (with respect to ε and MinPts) if p is within the ε neighborhood of q.
- Clearly, an object p is directly density-reachable from another object q if and only if q is a core object and p is in the ε -neighborhood of q.
- Using the directly density-reachable relation, a core object can “bring” all objects from its ε -neighborhood into a dense region.

DBSCAN: Density-Based Clustering Based on Connected Regions with High Density

- In DBSCAN, p is density-reachable from q (with respect to ε and MinPts in D) if there is a chain of objects p1, … ,pn, such that p1 = q, pn = p, and pi+1 is directly density-reachable from pi with respect to ε and MinPts, for 1 <= i <= n, pi ∈ D.



DBSCAN Algorithm

Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ε-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ε-neighborhood of p;
(8)          for each point p' in N
(9)              if p' is unvisited
(10)                 mark p' as visited;
(11)                 if the ε-neighborhood of p' has at least MinPts points,
                         add those points to N;
(12)             if p' is not yet a member of any cluster, add p' to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

DBSCAN algorithm.

OPTICS: Ordering Points to Identify the Clustering Structure

- OPTICS stands for Ordering Points To Identify the Clustering Structure. It gives a significant order of database with respect to its density-based clustering structure.
- OPTICS needs two important pieces of information per object:
- The **core-distance** of an object p is the smallest value ε’ such that the ε’ -neighborhood of p has at least MinPts objects. That is, ε’ is the minimum distance threshold that makes p a core object. If p is not a core object with respect to ε and MinPts, the core-distance of p is undefined.

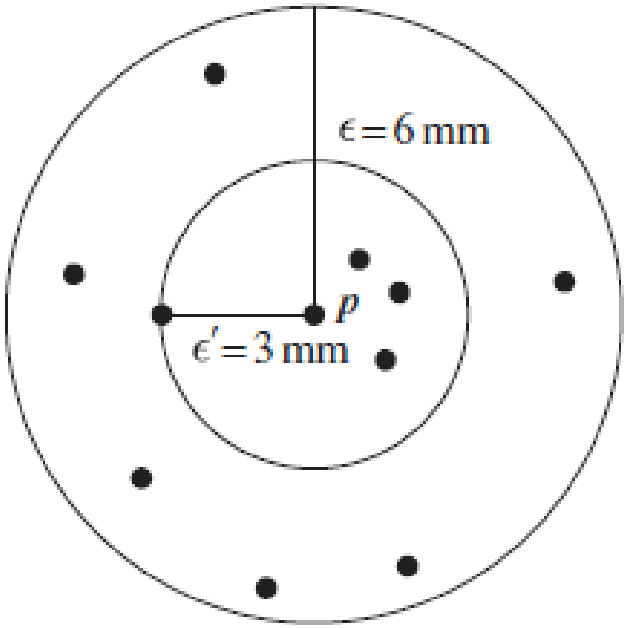
OPTICS: Ordering Points to Identify the Clustering Structure

- The **reachability-distance** to object p from q is the minimum radius value that makes p density-reachable from q. According to the definition of density-reachability, q has to be a core object and p must be in the neighborhood of q. Therefore, the reachability-distance from q to p is max{core-distance(q), dist(p, q)}. If q is not a core object with respect to ε and MinPts, the reachability-distance to p from q is undefined.

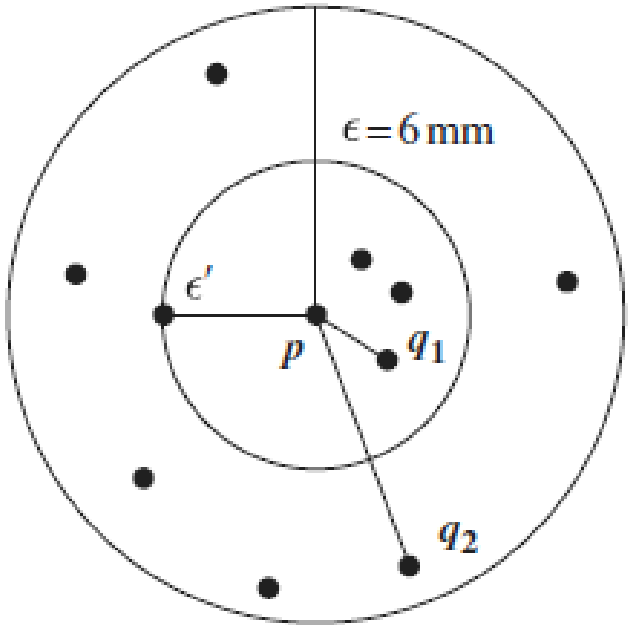
OPTICS: Ordering Points to Identify the Clustering Structure

- Figure illustrates the concepts of core distance and reachability-distance. Suppose that ε = 6 mm and MinPts = 5.
- The core distance of p is the distance, ε', between p and the fourth closest data object from p.
- The reachability-distance of q1 from p is the core-distance of p (i.e., ε' = 3mm) because this is greater than the Euclidean distance from p to q1.
- The reachability-distance of q2 with respect to p is the Euclidean distance from p to q2 because this is greater than the core-distance of p.

OPTICS: Ordering Points to Identify the Clustering Structure



Core-distance of p



Reachability-distance $(p, q_1) = \epsilon' = 3$ mm
Reachability-distance $(p, q_2) = dist\,(p, q_2)$

Grid-Based Methods

- The clustering methods discussed so far are data-driven—they partition the set of objects and adapt to the distribution of the objects in the embedding space.
- A grid-based clustering method takes a space-driven approach by partitioning the embedding space into cells independent of the distribution of the input objects.

- The grid-based clustering approach uses a multiresolution grid data structure.
- It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed.
- The main advantage of the approach is its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space.

- STING: STatistical INformation Grid
  - STING is a grid-based multiresolution clustering technique in which the embedding spatial area of the input objects is divided into rectangular cells.
- CLIQUE: An Apriori-like Subspace Clustering Method
  - CLIQUE partitions each dimension into nonoverlapping intervals, thereby partitioning the entire embedding space of the data objects into cells.

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Evaluation of Clustering

- Clustering evaluation assesses the feasibility of clustering analysis on a data set and the quality of the results generated by a clustering method.
- The tasks include assessing clustering tendency, determining the number of clusters, and measuring clustering quality.

- Assessing clustering tendency
  - For a given data set, we assess whether a nonrandom structure exists in the data.
  - Blindly applying a clustering method on a data set will return clusters; however, the clusters mined may be misleading.
  - Clustering analysis on a data set is meaningful only when there is a nonrandom structure in the data.

- Determining the number of clusters in a data set.
  - A few algorithms, such as k-means, require the number of clusters in a data set as the parameter. Moreover, the number of clusters can be regarded as an interesting and important summary statistic of a data set.
  - Therefore, it is desirable to estimate this number even before a clustering algorithm is used to derive detailed clusters.

- Measuring clustering quality.
  - After applying a clustering method on a data set, we want to assess how good the resulting clusters are. A number of measures can be used.
  - Some methods measure how well the clusters fit the data set, while others measure how well the clusters match the ground truth, if such truth is available.
  - There are also measures that score clusterings and thus can compare two sets of clustering results on the same data set.

Outlier Detection

- Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation.
- Such objects are called outliers or anomalies.
- Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance, and intrusion detection.
- Outlier detection and clustering analysis are two highly related tasks.
- Clustering finds the majority patterns in a data set and organizes the data accordingly, whereas outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns.
- Outlier detection and clustering analysis serve different purposes.

- An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism.
- We may refer to data objects that are not outliers as “normal” or expected data. Similarly, we may refer to outliers as “abnormal” data.
- Outliers are different from noisy data.

- Noise is a random error or variance in a measured variable.
- In general, noise is not interesting in data analysis, including outlier detection.
- For example, in credit card fraud detection, a customer's purchase behavior can be modeled as a random variable.
- A customer may generate some "noise transactions" that may seem like "random errors" or "variance," such as by buying a bigger lunch one day, or having one more cup of coffee than usual.
- Such transactions should not be treated as outliers; otherwise, the credit card company would incur heavy costs from verifying that many transactions.