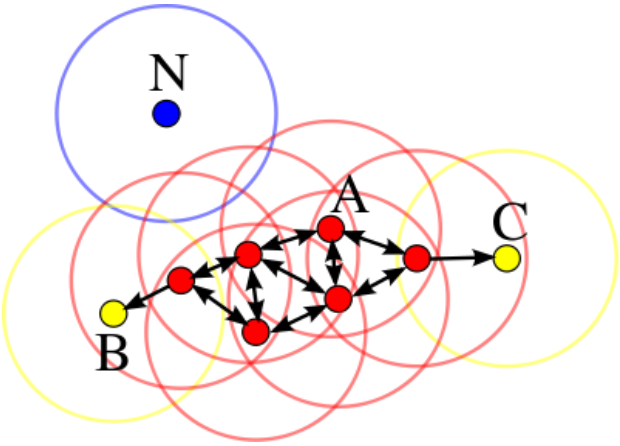
DBSCAN

Definition and purpose

Density-based spatial clustering of applications with noise (DBSCAN) is a **data clustering algorithm**. It is a density-based clustering **non-parametric algorithm**: given a set of points in some space, **it groups together points that are closely packed together** (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

Let *ε* be a parameter specifying the radius of a neighborhood with respect to some point. For the purpose of DBSCAN clustering**, the points are classified as core points, (density-) reachable points and outliers**, as follows:

- **Core Points**: A point p is a **core point** if **at least m points are within distance ε** of it (including p).
- **Reachable Points**: A point q is directly reachable from p if point q is within distance ε from core point p. **Points are only said to be directly reachable from core points.**
- **Outliers**: All points not reachable from any other point are outliers or noise points.



In this diagram, m = 4. **Point A** and the other red points **are core points**, because the area surrounding these points in an ε radius contain **at least 4 points** (including the point itself). Because they are all reachable from one another, they form a single cluster. **Points B and C are** not core points but are **reachable** from A (via other core points) and thus belong to the cluster as well. **Point N is a noise point** that is neither a core point nor directly reachable.

Algorithm

The DBSCAN algorithm can be abstracted into the following steps:

1. Find the points in the ε neighborhood of every point, **and identify the core points with more than m neighbors**.
2. Find the connected components of core points on the neighbor graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbor, otherwise assign it to noise.

Properties

Advantage

1. **DBSCAN does not require one to specify the number of clusters** in the data a priori, as opposed to k-means.
2. DBSCAN **can find arbitrarily-shaped clusters.** It can even find a cluster completely surrounded by (but not connected to) a different cluster.
3. DBSCAN has a notion of noise, **and is robust to outliers.**
4. DBSCAN requires just two parameters and is **mostly insensitive to the ordering of the points** in the database. (However, **points sitting on the edge of two different clusters might swap cluster membership** if the ordering of the points is changed, and the cluster assignment is unique only up to isomorphism.)
5. DBSCAN is designed for use with databases that can **accelerate region queries**, e.g. using an R* tree.
6. The parameters **m and ε can be set by a domain expert**, if the data is well understood.

Disadvantage

1. **DBSCAN is not entirely deterministic**: **border points** that are reachable from more than one cluster **can be part of either cluster**, **depending on the order the data are processed**. **But both on core points and noise points, DBSCAN is deterministic.**
2. DBSCAN cannot cluster data sets well with large differences in densities, **since the m-ε combination cannot then be chosen appropriately for all clusters**. (Same parameters for all clusters)
3. If the data and scale are not well understood, **choosing a meaningful distance threshold ε can be difficult**.