DBSCAN algorithm

DBSCAN stands for **d**ensity-**b**ased **s**patial **c**lustering of **a**pplications with **n**oise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

There are two key parameters of DBSCAN:

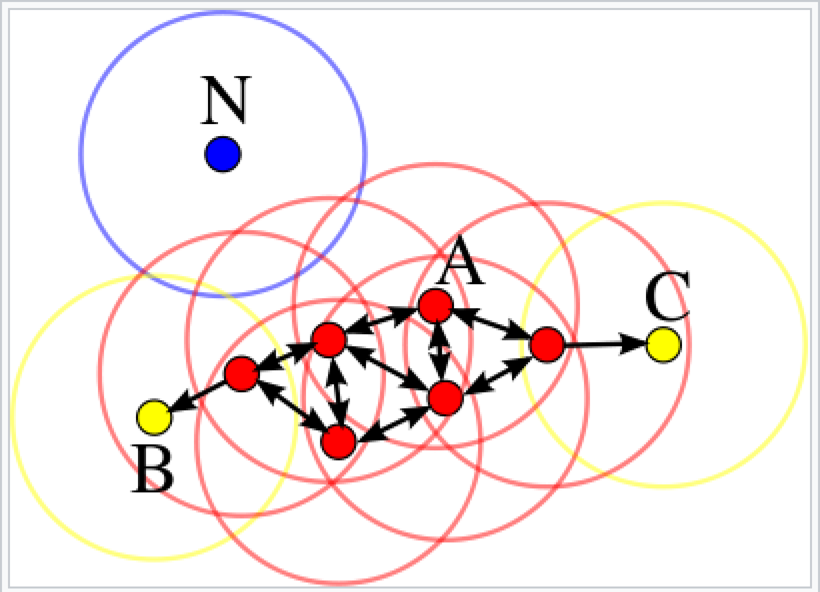
- **eps**: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to eps.
- **Min Pts:** Minimum number of data points to define a cluster.

Based on these two parameters, points are classified as core point, border point, or outlier:

- **Core point:** A point is a core point if there are at least minPts number of points (including the point itself) in its surrounding area with radius eps.
- **Border point:** A point is a border point if it is reachable from a core point and there are less than minPts number of points within its surrounding area.

- **Outlier:** A point is an outlier if it is not a core point and not reachable from any core points.

These points may be better explained with visualizations. The following figure is taken from Wikipedia:



In this case, minPts is 4. Red points are core points because there are **at least** 4 points within their surrounding area with radius eps. This area is shown with the circles in the figure. The yellow points are border points because they are reachable from a core point and have less than 4 points within their neighborhood. Reachable means being in the surrounding area of a core point. The points B and C have two points (including the point itself) within their neighborhood. Finally, N is an outlier because it is not a core point and cannot be reached from a core point.

We have learned the definitions of parameters and different type points. Now we can talk about how the algorithm works. It is actually quite simple:

- minPts and eps are determined.
- A starting point is selected at random at its neighborhood area is determined using radius eps. If there are at least minPts number of points in the neighborhood, the point is marked as core point and a cluster formation starts. If not, the point is marked as noise. Once a cluster formation starts (let's say cluster A), all the points within the neighborhood of initial point become a part of cluster A. If these new points are also core points, the points that are in the neighborhood of them are also added to cluster A.

Note: A point that is marked as noise may be revisited and be part of a cluster.

- Next step is to randomly choose another point among the points that have not been visited in the previous steps. Then same procedure applies.
- This process is finished when all points are visited.

The distance between points is determined using a distance measurement method as in k-means algorithm. The most commonly used method is Euclidean distance.

By applying these steps, DBSCAN algorithm is able to find high density regions and separate them from low density regions.

A cluster includes core points that are neighbors (i.e. reachable from one another) and all the border points of these core points. The required condition to form a cluster is to have at least one core point. Although very unlikely, we may have a cluster with only one core point and its border points.