DBSCAN Report

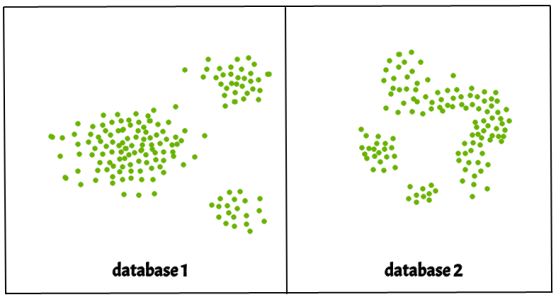
Donal Loitam

August 3, 2022

Contents

1 Introduction

- **Clusters** are dense regions in the data space, separated by regions of the lower density of points
- The **DBSCAN(Density-based spatial clustering of applications with noise) algorithm** is based on this intuitive notion of **"clusters"** and **"noise"**.
- The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.



2 The Problem with k-Means Clustering :

- Partitioning methods (K-means) work for finding spherical-shaped clusters or convex clusters i.e they are suitable only for compact and well-separated clusters
- Moreover, they are also severely affected by the presence of noise and outliers in the data.
- And as we know, **Real life data often contain irregularities and noise**

3 Key Points of the Algorithm

- **Parameters of DBSCAN algorithm:**

  1. **eps :** It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered neighbors. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the k-distance graph.
  2. **MinPts:** Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts ¿= D+1. The minimum value of MinPts must be chosen at least 3.

- **3 types of Datapoints in the algorithm**:

  1. **Core Point:** A point is a core point if it has **more than MinPts points** within eps radius
  2. **Border Point:** A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.
  3. **Noise or outlier:** A point which is not a core point or border point.
  4. Below is a figure depicting the three types of datapoints and the parameters of the algorithm

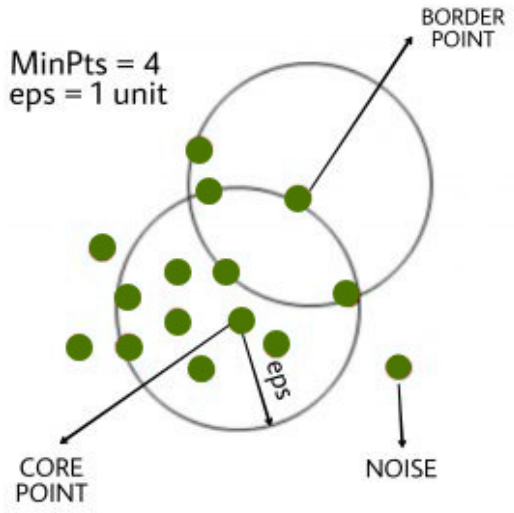


Figure 1: figure depicting types of datapoints and parameters involved

- **The Algorithm at a Glance:**

  1. Find all the neighbour points within eps and **identify the core points** or visited with more than MinPts neighbors.

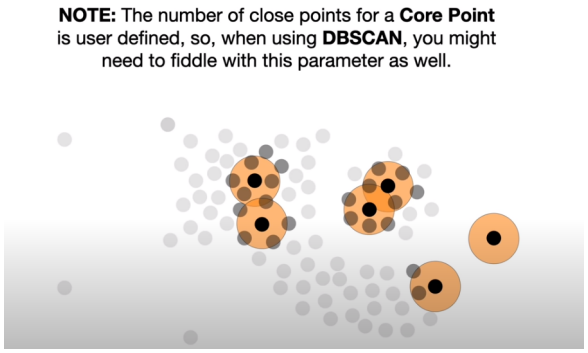


Figure 2: **MinPts is userdefined, here MinPts=4**

2. **For each core point**, if it is not already assigned to a cluster, **create a new cluster.**

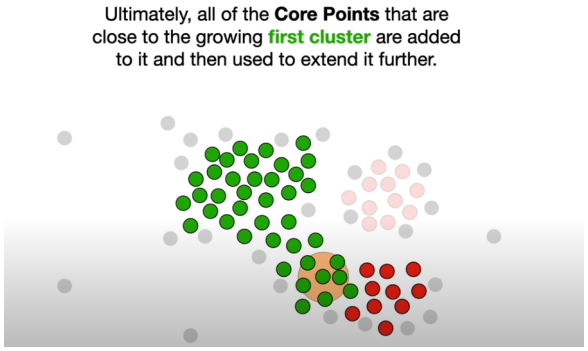


Figure 3: **Creating a cluster**

3. Find recursively all its density connected points and assign them to the same cluster as the core point.

   A point a and b are said to be density connected if there exist a point c which has a sufficient number of points in its neighbors and both the points a and b are within the eps distance. This is a chaining process. So, if b is neighbor of c, c is neighbor of d, d is neighbor of e, which in turn is neighbor of a implies that b is neighbor of a.

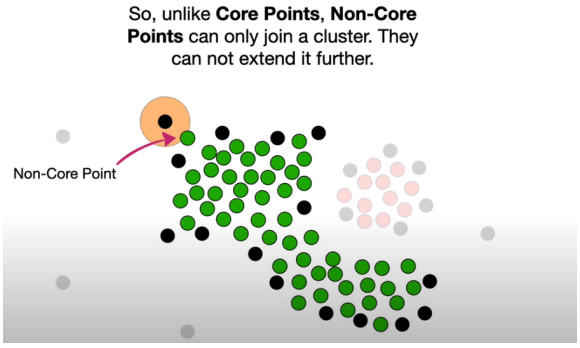


Figure 4: **Assigning Clusters to Non-Core points**

4. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are **noise.**

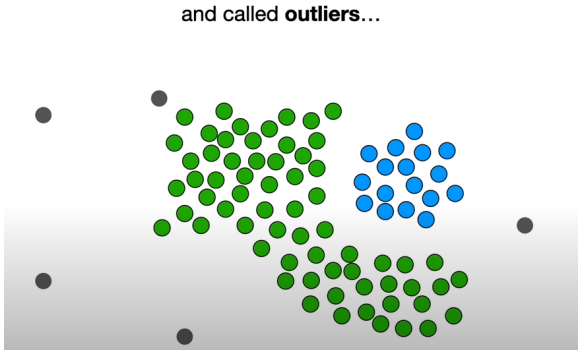


Figure 5: **Black points which are not near are called noise/outliers**

- **DBSCAN clustering algorithm in pseudocode:**
  DBSCAN(dataset, eps, MinPts){
  #cluster index
  C = 1
  for each unvisited point p in dataset {
  mark p as visited
  # find neighbors
  Neighbors N = find the neighboring points of p

  if $|N|$ >=MinPts:
  $N = NUN'$
  if p' is not a member of any cluster:
  add p' to cluster C
  }

4 Some Questions

1. How can we interpret the parameters "eps" and "Minpts" in high dimensions for the DBSCAN Algorithm?
Ans.
2. What are density reachability and density connectivity?
Ans.
3. What is the time complexity of the DBSCAN Clustering Algorithm?
Ans :
4. How is the parameter "eps" estimated in the DBSCAN Algorithm?
Ans.
5. Why does there arise a need for DBSCAN when we already have other clustering Algorithms?
Ans.