DBSCAN

Limitation of KMeans

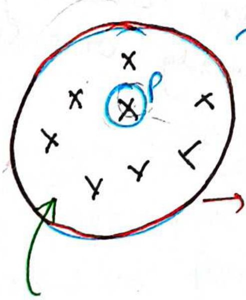
1. You have to tell the number of cluster to be formed
2. Not good with arbitrary cluster
3. Sensitive to outliers

Density Based Clustering

Density based Clustering algorithm divides your entire datasets into dense regions separated by sparse regions.

Min points & Epsilon

How to measure around a point?



→ $\boxed{\text{area}}$

→ # points

→ 2-d circle

→ 3-d circle

→ nd circle

area → circle

radius

1 unit → epsilon

dense

min pts

no. of points ≥ 4 → dense

no of points ≤ 4 → Sparse

Min Pts stands for "Minimum Points", is a parameter that specifies the minimum number of points required to form a dense region, which is considered a cluster.

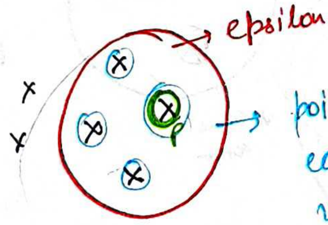
Epsilon (ε) is a key parameter that defines the radius of the neighbourhood around a given data point. Specifically, ε is the maximum distance between two points for them to be considered as part of the same neighbourhood. This parameter is crucial in determining whether points are close enough to be included in a cluster.

Core points, Border Points & Noise Points

A point is considered a core point if it has a minimum number of other points (specified by MinPts) within a given radius ε of itself.

min pts = 4

epsilon = 1



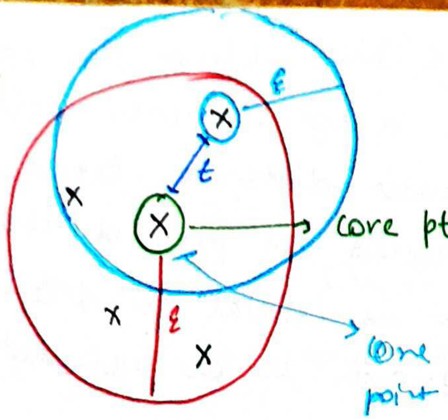
points inside epsilon circle is equal to or greater than min pts. So, p is core point

A border point is defined as follows:

→ Not a core point: A border point does not meet the criteria to be a core point. It has fewer than MinPts within its ε- neighbourhood.

→ Neighbour of a core Point: A border point is within the ε distance of one or more core points. In other words, it lie on the edge of a cluster, within the radius ε of atleast one core point.

minPts = 5

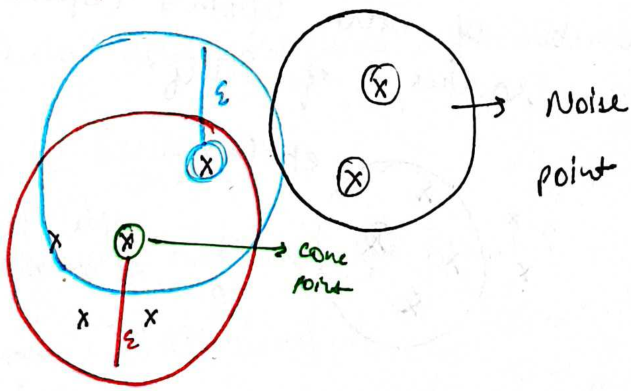


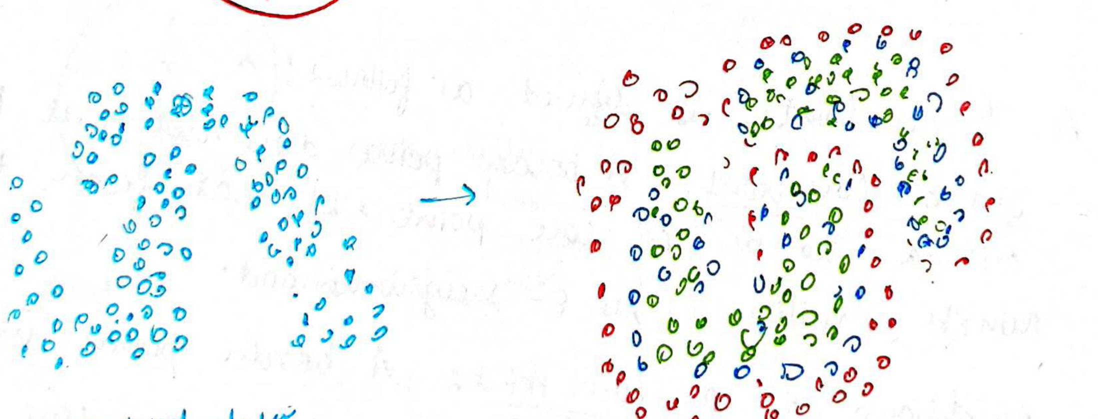
1) $t < \varepsilon$ and one core pt present in circle

2) In the circle pt is less the min pts and one core pt present in circle

border point

A noise point is a data point which can neither a core point nor a border point.





Original data

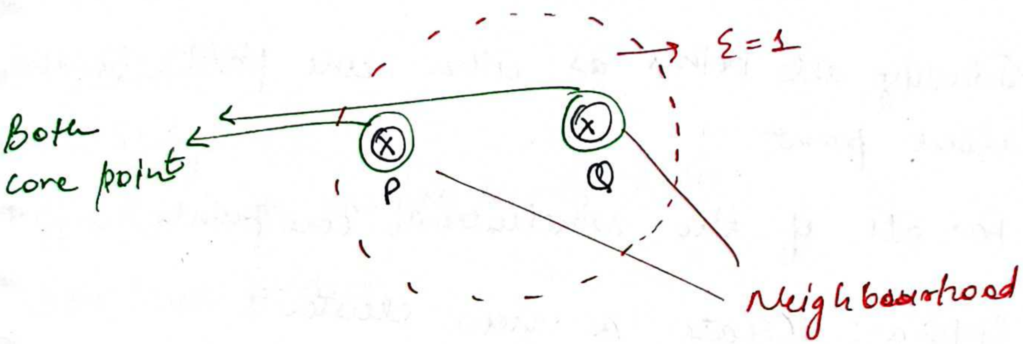
noise, border, core point

Density Connected Points

Directly Density Reachable

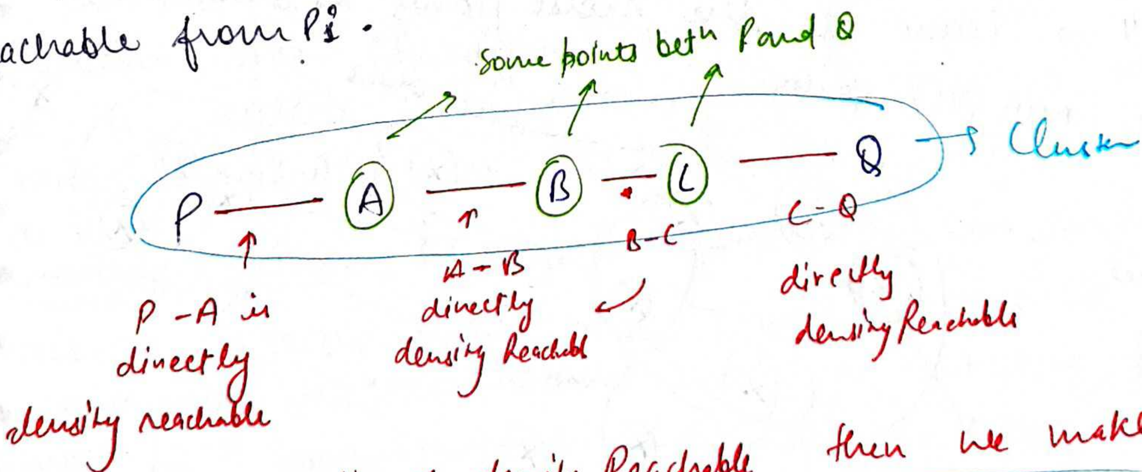
A point P is directly density reachable from a point Q given Eps, MinPts if:

1. P is the Eps - neighbourhood ~~of~~ Q
2. Both P and Q are core points



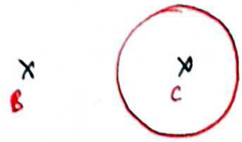
Density Connected Points

A point P is density connected to Q given Eps, MinPts if there is a chain of points $P_1, P_2, P_3, \ldots P_n$ $P_1 = P$ and $P_n = Q$ such that P_{i+1} is directly density reachable from P_i.



If all are directly density Reachable then we make cluster from P to Q

PABCQ
↓
core point



Not directly density Reachable

P — A — B | C — D

Not directly density Reachable

P– A–B in same cluster

Simplified DBSCAN Algo

Step1 → Identify all points as either core point, border point or noise point

Step2 → For all of the unclustered core points

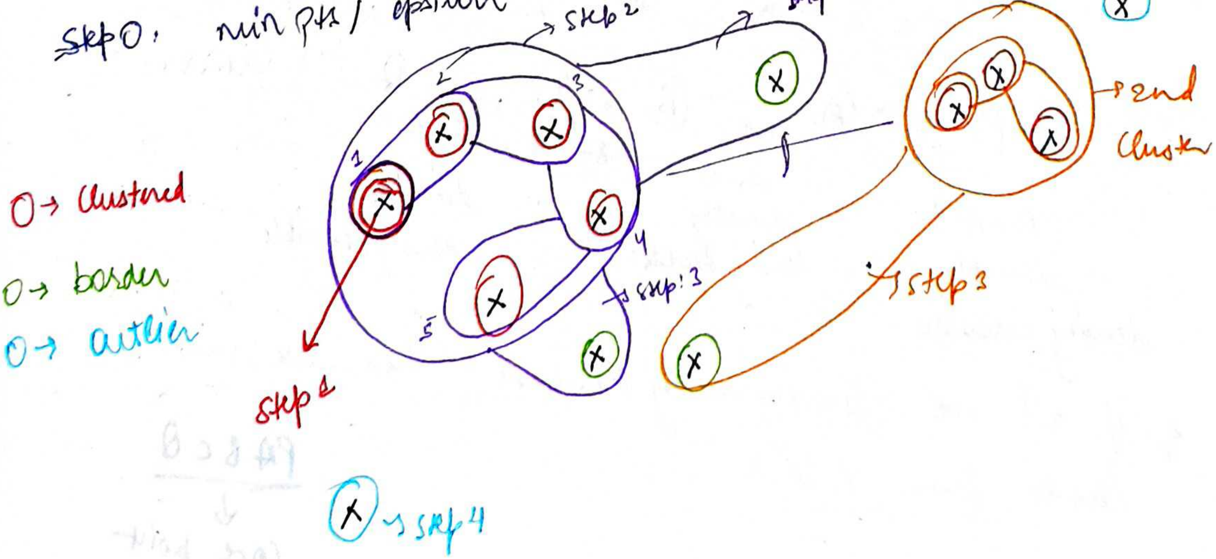
Step2a : Create a new cluster

Step2b: and all the points that are unclustered and density connected to the current point into this cluster.

Step3 → For each unclustered border point assign it to the cluster of nearest core point.

Step4 → Leave all the noise points as it is.

Step0: min pts / epsilon



Limitations & Advantages

Advantages

1. Robust to outliers
2. No need to specify cluster
3. Can find arbitrary shaped cluster
4. Only 2 hyperparameters to tune

Disadvantage

1. Sensitivity
2. Difficulty with varying density cluster
3. Does not predict

Application Areas

1. Spatial Data Analysis: DBSCAN is particularly well-suited for spatial data clustering due to its ability to find cluster of find clusters of arbitrary shapes, which is common is geographic data. It's used in application like identifying regions of similar land use in satellite images or grouping location with similar activities in GIS (Geographic Information Systems).

2. Anomaly Detection: The algo effectiveness in distinguishing noise or outliers from core cluster make it useful in anomaly detection tasks, such as detecting fraudulent activities in banking transaction or identifying usual patterns in network traffic.

3) Image Processing: In image analysis, DBSCAN can be used for tasks like object recognition and image segmentation, where the goal is to group pixels or features that form meaningful structures.

4) Bioinformatics: DBSCAN is ~~[illegible]~~ applied in bioinformatics for tasks such as gene expression data analysis, where it help to identify groups of genes with similar expression patterns which might indicate a functional relationship.

5) Customer Segmentation: In Marketing and bussiness analytics, DBSCAN can be used for customer segmentation by identifying cluster of customer with similar buying behaviours or preferences.

6) Astronomy: The algo is employed in astronomy for tasks like star cluster identification, where it groups stars based on their physical proximity or other attributes.

7) Environmental Studies: DBSCAN can be used in Environmental monitoring, for example, to cluster areas based on pollution levels or to identify regions with similar environmetal characteristic

8) Traffic Analysis: In traffic and transportation studies, DBSCAN is useful for identifying hotspots of traffic congestion or for clustering routes with similar traffic patterns

9) Machine Learning and Data Mining : More broadly, in the fields of Machine learning and data mining, DBSCAN is employed for exploratory data analysis, helping to uncover natural structures or patterns in data that might not be apparent otherwise.

10) Social Network Analysis : The algo can be used to detect communities or group within social networks based on interaction patterns or shared interests.