Plan of Attack

15 December 2023 09:17

Limitations of Kmeans

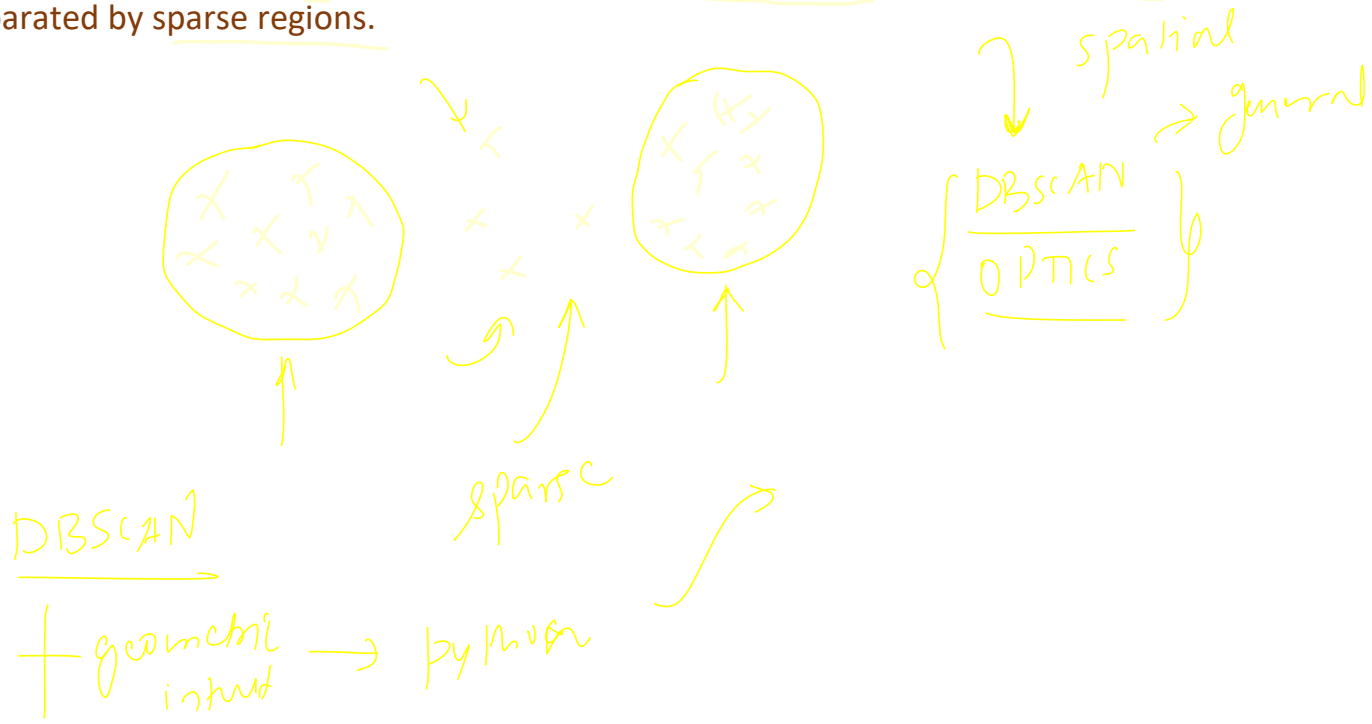
15 December 2023 09:22

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

Density based clustering

15 December 2023 09:17

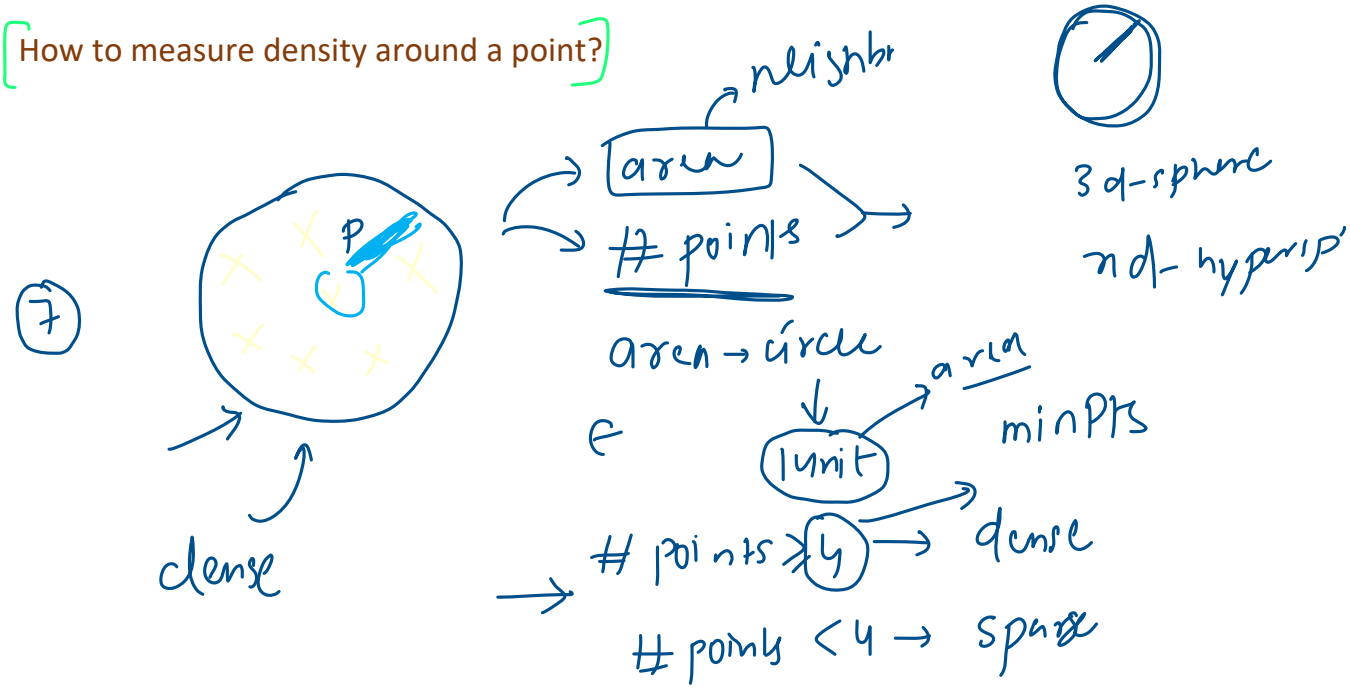
Density based clustering algorithms divides your entire dataset into dense regions separated by sparse regions.



MinPts & Epsilon → hyperparameters

15 December 2023 09:17

How to measure density around a point?



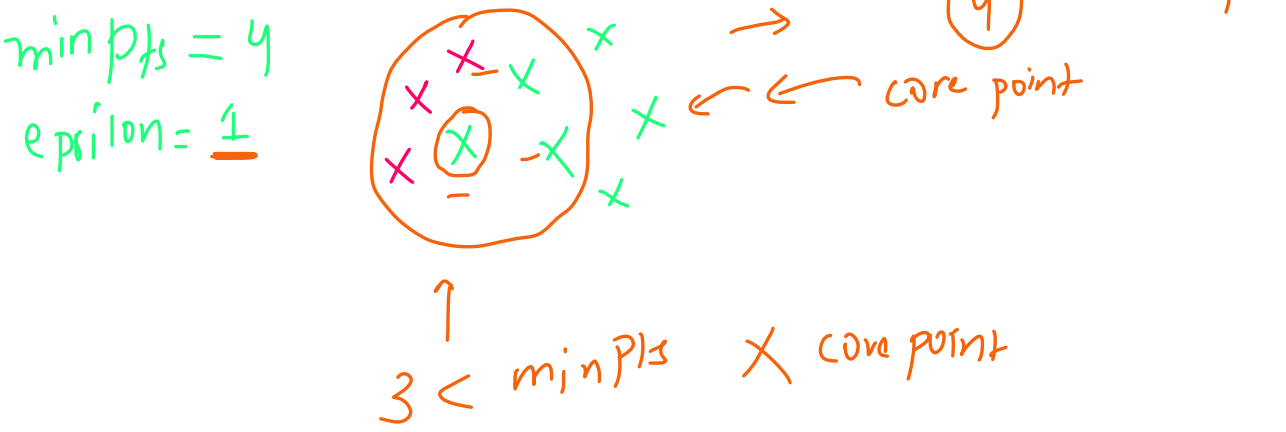
MinPts stands for "Minimum Points," is a parameter that specifies the minimum number of points required to form a dense region, which is considered a cluster.

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

15 December 2023 09:18

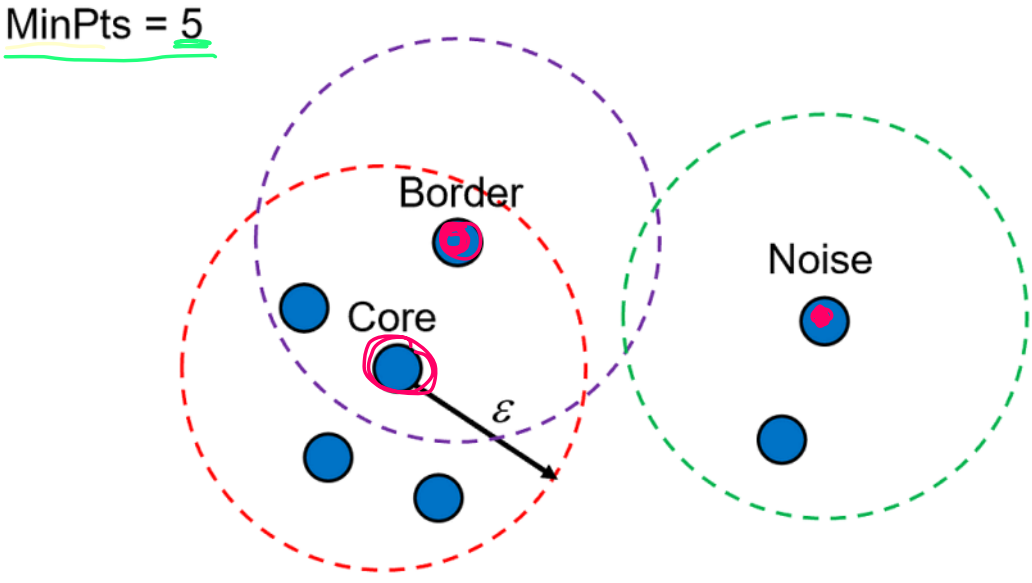
A point is considered a **core point** if it has a minimum number of other points (specified by MinPts) within a given radius ε of itself.

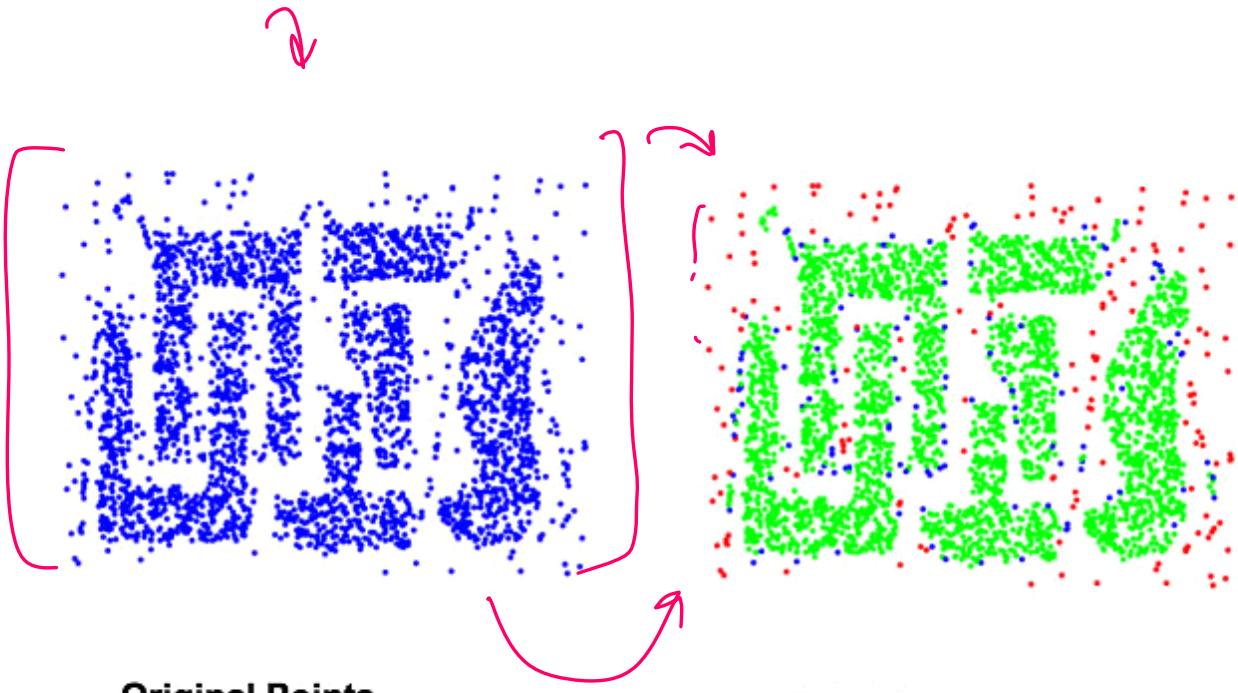


A **border point** is defined as follows:

- 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 lies on the edge of a cluster, within the radius ε of at least one core point.

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





Original Points

Point types: core, border and noise

Eps = 10, MinPts = 4

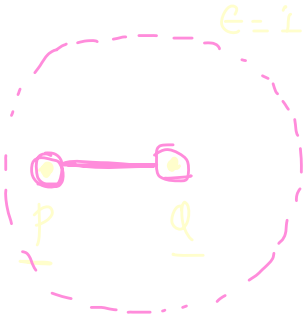
Density Connected Points

15 December 2023 09:18

Directly Density Reachable

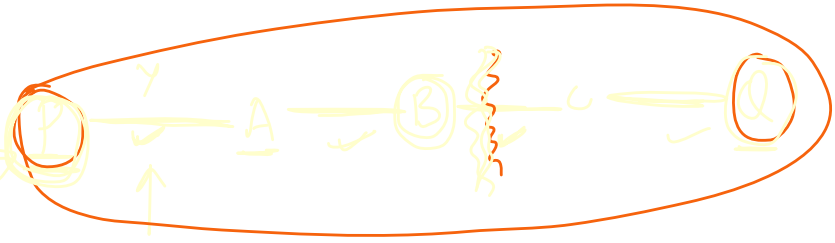
A point P is directly density-reachable from a point Q given Eps, MinPts if:

1. P is in the Eps-neighborhood 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 ..... P_n, P_1 = P and P_n = Q such that P_{i+1} is directly density reachable from P_i

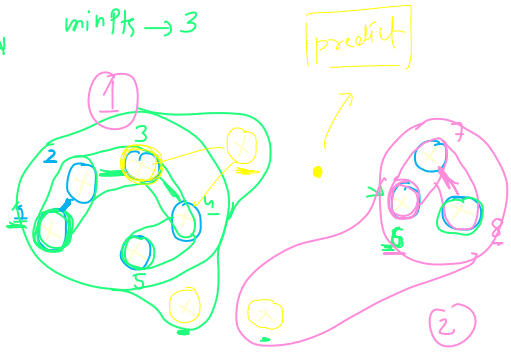


Simplified DBSCAN Algorithm

15 December 2023 09:18

Step 0 → minPts / epsilon

minPts → 3



Step 1 - Identify all points as either core point, border point or noise point

Step 2 - For all of the unclustered core points

Step 2a - Create a new cluster

Step 2b - add all the points that are unclustered and density connected to the current point into this cluster

Step 3 - For each unclustered border point assign it to the cluster of nearest core point

Step 4 - Leave all the noise points as it is.

Animations

15 December 2023 11:32

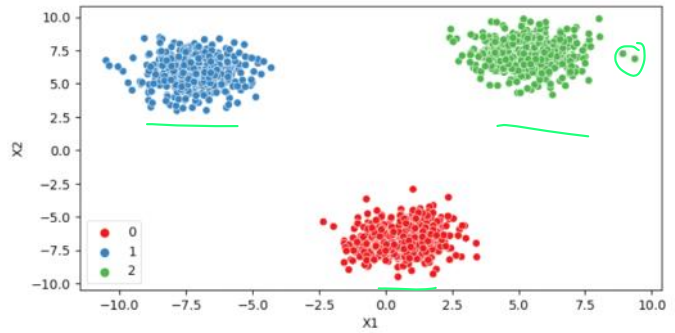
Code Example

15 December 2023 09:19

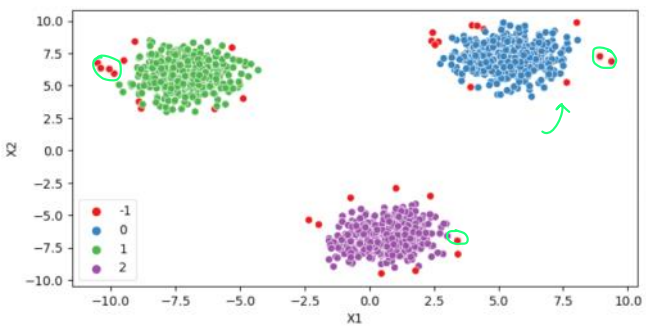
Kmeans Vs DBSCAN

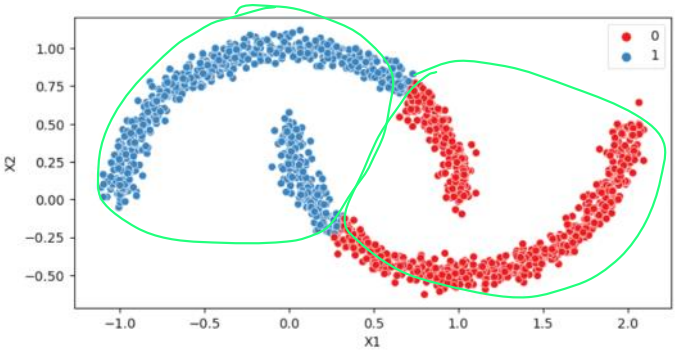
15 December 2023 09:29

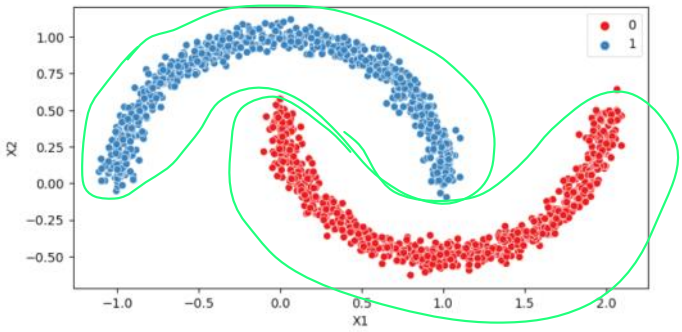
Kmeans

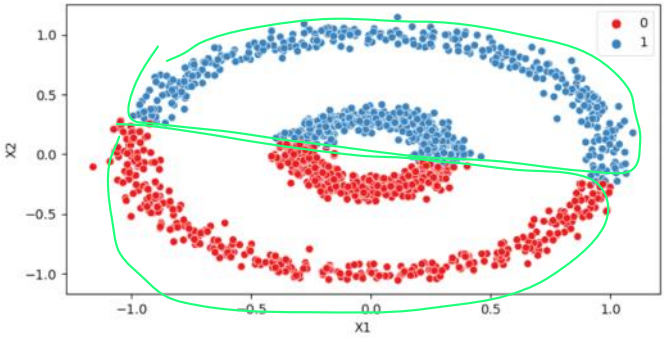


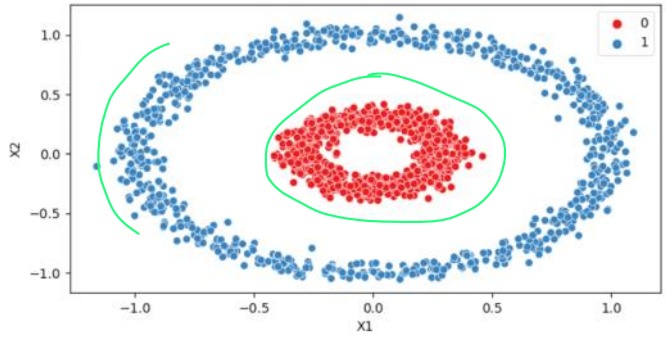
DBSCAN











Limitations & Advantages

15 December 2023 09:19

Advantages

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

minPts

epsilon

Disadvantages

1. Sensitivity to hyperparameters
2. Difficulty with varying density clusters
3. Does not predict

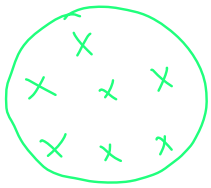
predict

new

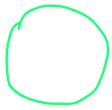
retrain

kmeans

epsilon minPts



OPTICS



Application Areas

15 December 2023 09:19

1. **Spatial Data Analysis**: DBSCAN is particularly well-suited for spatial data clustering due to its ability to find clusters of arbitrary shapes, which is common in geographic data. It's used in applications like identifying regions of similar land use in satellite images or grouping locations with similar activities in GIS (Geographic Information Systems).
2. **Anomaly Detection**: The algorithm's effectiveness in distinguishing noise or outliers from core clusters makes it useful in anomaly detection tasks, such as detecting fraudulent activities in banking transactions or identifying unusual 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 applied in bioinformatics for tasks such as gene expression data analysis, where it helps to identify groups of genes with similar expression patterns, which might indicate a functional relationship.
5. **Customer Segmentation**: In marketing and business analytics, DBSCAN can be used for customer segmentation by identifying clusters of customers with similar buying behaviours or preferences.
6. **Astronomy**: The algorithm 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 environmental characteristics.
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 algorithm can be used to detect communities or groups within social networks based on interaction patterns or shared interests.