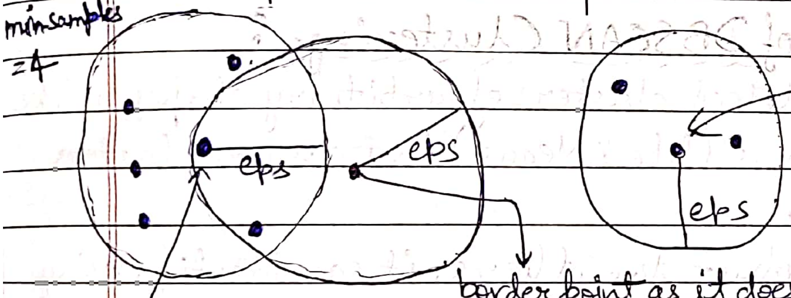
DBSCAN Clustering

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a powerful clustering algorithm used in ML. It stands out for its ability to discover clusters of arbitrary shape and identify noise (outliers) in datasets. Here we need to know few concepts to fully understand the working of DBSCAN clustering.

* **Core Point:** A point is said to be a core point if there are atleast "min-samples" points (including the point itself) within a radius distance "eps" (epsilon) from it.

* **Border Point:** A point is said to be a border point when it has less points than "min-samples" within "eps" radius but one of the points inside must be a core point.

* **Noise Point:** Points that are neither core points and nor border points are classified as noise or outliers.



Noise point, as it does not have min-samples and neither it has any core point in its eps distance.

border point as it does not have min-samples but has one node point in eps.

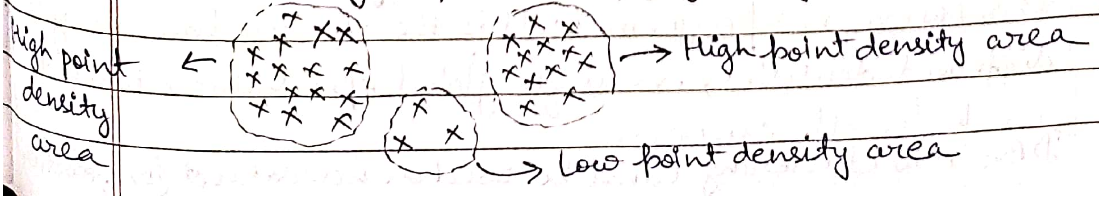
core point since it has more than 4 samples within eps radius.

- DBSCAN heavily relies on two ~~features~~/critical parameter

* **eps (epsilon):** The maximum distance between two points for one to be considered part of the neighbourhood of other.

* **min-samples:** The minimum number of points required to form a dense region (i.e., a core point). This includes the core point itself.

DBSCAN identifies clusters based on the density of points in a dataset. It assumes that clusters are regions of high point density separated by regions of low point density.



Process of DBSCAN clustering:

DBSCAN clusters data through the following steps:

(1) Start with an arbitrary point. If this point is a core point, a cluster is formed. If it it is a border point or noise, it is labeled accordingly.

(2) Find the neighbors of the core point. All points within a distance **eps** (epsilon) from the core point are considered its neighbors.

(3) Expand the cluster. If a core point is found, DBSCAN continues expanding the cluster by recursively finding all reachable points with in the **eps** neighbourhood.

(4) Assign clusters. If a point is not reachable from any core point, (i.e, no core point within eps), it is labelled as noise.

Advantages of DBSCAN Clustering:

(1) DBSCAN can detect clusters of arbitrary shapes, unlike other algorithms (like KMeans) that assumes clusters are spherical.

(2) DBSCAN is robust to outliers, it can identify noisy points and do not make them part of cluster.

(3) Unlike KMeans, DBSCAN does not require us to predefine the number of clusters.

Disadvantages:

(1) Choosing **eps** and **min-samples**. The algorithm performance depends heavily on selecting appropriate values for eps and min-samples. Too large or too small values can lead to poor clustering results.

(2) DBSCAN may struggle with high-dimensional data since it relies on distance measures, and the concept of "distance" becomes less meaningful in high dimensions (a phenomenon known as the curse of dimensionality).

(3) DBSCAN may not perform well if the clusters have very different densities, as a single global **eps** value might not work for all clusters.

(4) DBSCAN clustering can't be used on new dataset for prediction'

(5) DBSCAN clustering can't be used on distributed computing setup as all the points needs to be loaded at once in memory to get Eps and min-pts.

Ques.1 How do you define density in context of DBSCAN?

Ans→ In context of DBSCAN, density refers to the number of points in a given neighbourhood around a point. The density is determined by two key parameters **eps** and **min-samples**. Density at a point is considered high if there are atleast **min-samples** (including the point itself) within its **eps**-radius neighbourhood.

Ques.2. What is the meaning of directly-density reachable and density-connected points?

Ans→ A point **p** is said to be **directly-density reachable** from another point q if:

→ p is within eps-radius neighbourhood of q (i.e.; the distance between p&q is less than or equal to eps) and

→ q is a core point, meaning there are atleast min-sample points (including q) within its eps-radius neighbourhood.

In simple words, a point p is directly density reachable from q if q is a core point, and p lies within the neighbourhood defined by eps around q.

Two points **p** and **q** are said to be **density-connected** if there exists a chain of points $p_1, p_2, ---, p_n$ such that: $p_1 = p$ and $p_n = q$, and each point in the chain is directly density reachable from the previous one.

In other words, two points are density connected if there are a series of directly density-reachable connections between them through intermediate points. This means even if p&q are not density-reachable from each other, they can still belong to the same cluster if they are connected through other core points.

Ques. 3 How can we select the appropriate values for the **eps** and **min-samples** parameter in DBSCAN?

Ans → Selecting the appropriate values for Eps and min-samples in DBSCAN is crucial for achieving good clustering results. These parameters control how DBSCAN defines the "density" of points in the dataset, and their values can significantly impact the resulting clusters.

(1) Choosing the min-samples parameter:
A good heuristic for min-samples is to set it to atleast the dimensionality of data plus one (Min-samples ≥ D+1), where D is the number of dimensions. For example, for 2D data, we can start with minsamples = 3.
In case of large-dimensional data, we can use thumb rule of min-samples ≥ 2D. i.e, for 100-dimension data, min-samples ≥ 200.

(2) Choosing the eps parameter:
To select appropriate value of eps (the radius defining the neighbourhood), we can use different techniques like-

(a) k-nearest neighbors (k-NN) distance plot:
Steps: → Compute the distance from its k-th nearest neighbor, where k = min-samples.
→ Sort the distances in ascending order
→ Plot the distances. The "elbow" point (the point of maximum curvature) is often a good candidate for **eps**.
The elbow represents the transition from points that are within dense clusters to those that are more isolated (potential noise points).

(b) Cross-Validation: We can even test with different values of eps and evaluate the resulting clusters using internal cluster validation metrics like silhoutte score, davies-bouldin index, Dunn's index, Calinski-Harabasz index, etc.

(c) In some cases, domain knowledge can help set eps. For eg. in geographical datasets.

(d) Using OPTICS clusteringalgorithm.

Ques.4 How DBSCAN gets affected when it faces dataset having varying density?

Ans→ DBSCAN may struggle with datasets where clusters have varying densities. If the eps and min-sample parameters are set for detecting a dense cluster, they might fail to identify a larger less dense cluster and vice-versa. In such cases, parameter tuning or other variations of DBSCAN, such as HDBSCAN (Hierarichal DBSCAN), might be necessary to handle clusters with significantly different densities.

Ques.5 What is the effect of increasing or decreasing eps?

Ans→ Increasing "eps" leads to fewer, larger clusters and fewer noise points, but may result in over-aggregation of distinct clusters into one.

Decreasing "eps" leads to more, smaller clusters and more noise points, but it risks under-clustering if 'eps' is too small.