K-means Limitations

Friday 10 May 2024 10:04 PM

→ Centroid Based

→ # of clusters to be formed must be specified in advance

→ Not good with arbitrary clusters

→ Sensitive to outliers

Density Based Clustering

Friday 10 May 2024 10:12 PM

→ Divides your entire dataset into dense regions separated by sparse regions

→ DBSCAN

→ OPTICS

→ With the help of sparse regions we are able to separate dense regions

MinPts and Epsilon

Friday 10 May 2024 10:34 PM

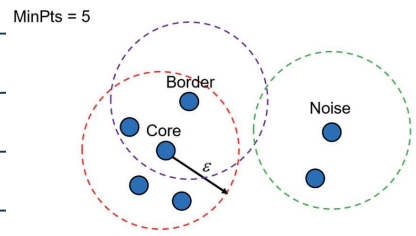
→ MinPts : Minimum Points is a parameter that specifies the minimum # of points required to form a dense region, which is considered a cluster.

→ Epsilon (ε) defines the radius of the neighbourhood around a given data point. ε is the maximum distance b/w 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.

Points

Saturday 11 May 2024 2:02 AM

→ A point is considered a core point if at least minPts points are within radius ε of it (including itself).



→ Noise point: Which can neither be a core nor a border point.

→ Border point
- A border point doesn't meet the criteria to be a core point.
- A border point is within the ε distance of one or more core points.

Density Connected Points

Saturday 11 May 2024 2:15 AM

Directly Density Reachable

A point P is directly density-reachable from a point Q given ε, minPts if:

→ P is in the ε neighborhood of Q

→ Both P & Q are core points

Density Connected points

Should we put them in the same cluster?

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

Simplified Algorithm

Saturday 11 May 2024 2:36 AM

[Read Research Paper later]

1) Identify all points as either — Core, Border, Noise

2) For all unclustered core points

   a) create a new cluster

   b) add all points that are unclustered & density connected to the current point into this cluster.

3) For each unclustered border point assign it to the cluster of nearest core point.

4) leave all the noise points as it is.

→ Not used for prediction

→ Used to cluster existing data

Pros
- Robust to outliers
- No need to specify clusters
- can find arbitrary shaped clusters (shape is not a concern)
- only 2 hyperparameters to tune

Cons
- sensitivity to hyperparameters
- Difficulty with varying density clusters (one value for ε & minPts) → one size doesn't fit all
- Does not predict

→ in general clustering algos don't perform well in higher dimensions. [Distances aren't reliable in 4 dimension]

Dimension Reduction techniques → Clustering

→ Clustering Metrics??