Optics

Avishek

2023-05-29

Preliminary

Loading & Cleaning Data

Introduction

OPTICS stands for Ordering Points To Identify Clustering Structure.
This algorithm can be seen as a generalization of DBSCAN. A major issue with DBSCAN is that it fails to find clusters of varying density due to fixed eps.
This is solved in OPTICS by using an approach of finding reachability of each point from the core points and then deciding the clusters based on reachability plot.

Relevant terminologies for OPTICS

- ϵ, epsilon (eps): is known as the Maximum allowed distance from one point to the other point for both of them to be considered in one group/cluster
- MinPts: is the minimum number of points which should be present close to each other at a distance of epsilon (ϵ) so that they all can be form a group/cluster Core Point: That point which has at least MinPts number of points near to it, within the distance of ϵ(eps)
- Border Point/Non-Core Point: That point in data which has less than the minimum number of points(MinPts) within its reach (a distance of eps)
- Noise: That point which has no point near to it within a distance of eps
- Core Distance: The minimum distance required by a point to become a core point. It means it is possible to find the MinPts number of points within this distance. Core distance can be less than the pre-decided value of ϵ, epsilon (eps), which is the maximum allowed distance to find MinPts.
- Reachability distance: This is the distance to reach a point from the cluster. Now if the point lies within the Core Distance, then Reachability Distance=Core Distance. And, if the point lies outside the Core Distance, then Reachability Distance is the minimum distance from the extreme point of the cluster.

Alogrithm Steps

1. For the given values of MinPts and eps(ϵ). Find out if a point is close to MinPts number of points within a distance less than or equal to eps. Tag it as a Core Point. Update the reachability distance = core distance for all the points within the cluster.
2. If it is not a core point then find out its density connected distance from the nearest cluster. Update the reachability distance.

3. Arrange the data in increasing order of reachability distance for each cluster. The smallest distances come first and represent the dense sections of data and the largest distances come next representing the noise section. This is a special type of dendrogram
4. Find out the places where a sharp decline is happening in the reachability distance plot.
5. “Cut” the plot in the y-axis by a suitable distance to get the clusters.