In this week you will learn about 2 techniques DBSCAN and LOF

DBSCAN: Density-Based Clustering Essentials

Density-based clustering algorithm, which can be used to identify clusters of any shape in a data set containing noise and outliers

Why DBSCAN??

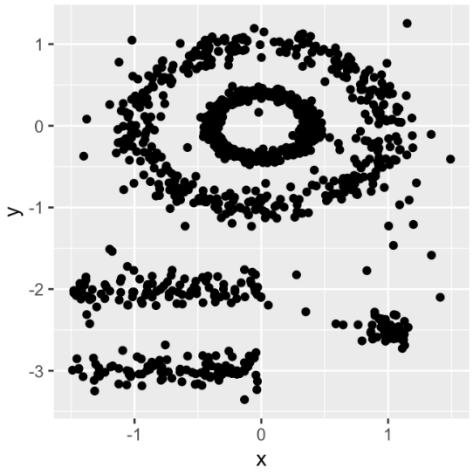
Partitioning methods (K-means, PAM clustering) and hierarchical clustering are suitable for finding spherical-shaped clusters or convex clusters. In other words, they work well only for compact and well separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data.

Unfortunately, real life data can contain:

i) clusters of arbitrary shape such as those shown in the figure below (oval, linear and “S” shape clusters)

ii) many outliers and noise.

The figure below shows a data set containing nonconvex clusters and outliers/noises



The plot above contains 5 clusters and outliers, including:

- 2 ovales clusters
- 2 linear clusters
- 1 compact cluster

Given such data, k-means algorithm has difficulties for identifying theses clusters with arbitrary shapes

In this week we're going to learn about a clustering approach that throws out the K-means assumption that clusters fall in convex globular clusters and does something different: spectral clustering

Pros

1. Unlike K-means, DBSCAN does not require the user to specify the number of clusters to be generated
2. DBSCAN can find any shape of clusters. The cluster doesn't have to be circular.
3. DBSCAN can identify outliers

Cons

1. Cannot handle varying densities
2. Sensitive to parameters

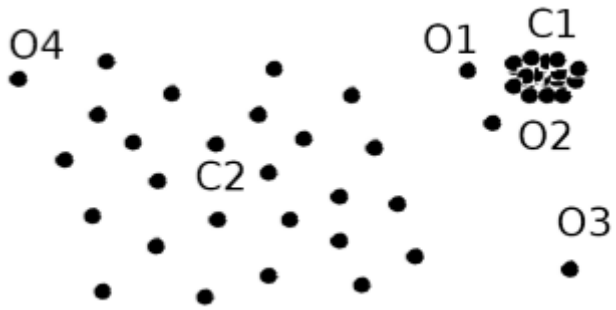
LOF : Density-Based Outlier Detection

Density based outlier detection algorithm, used for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbours

Intuition (density-based outlier detection): The density around an outlier object is significantly different from the density around its neighbours

Why LOF??

Due to the local approach, LOF can identify outliers in a data set that would not be outliers in another area of the data set. For example, a point at a "small" distance to a very dense cluster is an outlier, while a point within a sparse cluster might exhibit similar distances to its neighbours.



In the above Fig., o1 and o2 are local outliers to C1, o3 is a global outlier, but o4 is not an outlier. However, proximity-based clustering cannot find o1 and o2 are outlier (e.g., comparing with O4).

Cons

1. The resulting values are quotient-values and hard to interpret. A value of 1 or even less indicates a clear inlier, but there is no clear rule for when a point is an outlier. In one data set, a value of 1.1 may already be an outlier, in another dataset and parameterization (with strong local fluctuations) a value of 2 could still be an inlier. These differences can also occur within a dataset due to the locality of the method
2. Sensitive to parameters