Cluster Analysis

Density Based Clustering

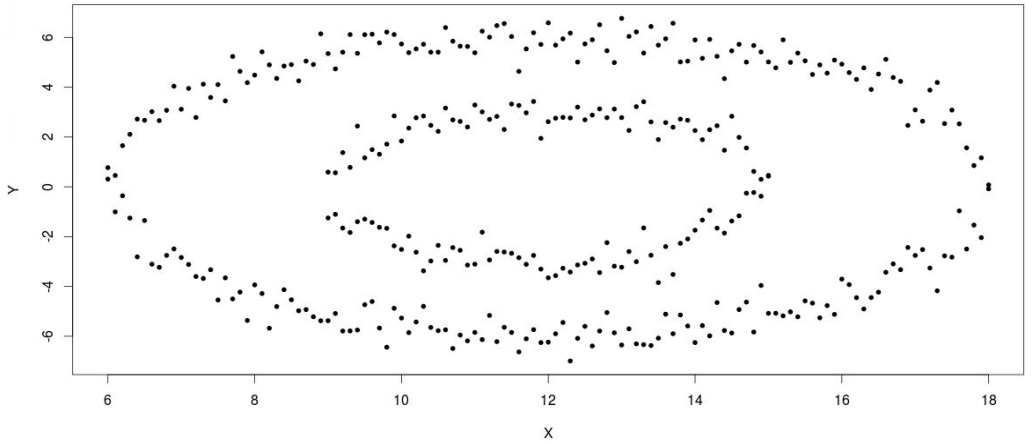
By
Subhasis Dasgupta
Asst Professor
Praxis Business School, Kolkata

Relook at partition based clustering

Centroid based clustering techniques (kmeans and kmedoid) are good when patterns are not irregularly shaped

Lets have a look, how kmeans clustering, clusters the dataset given below

There are two clusters having shapes of concentric circles

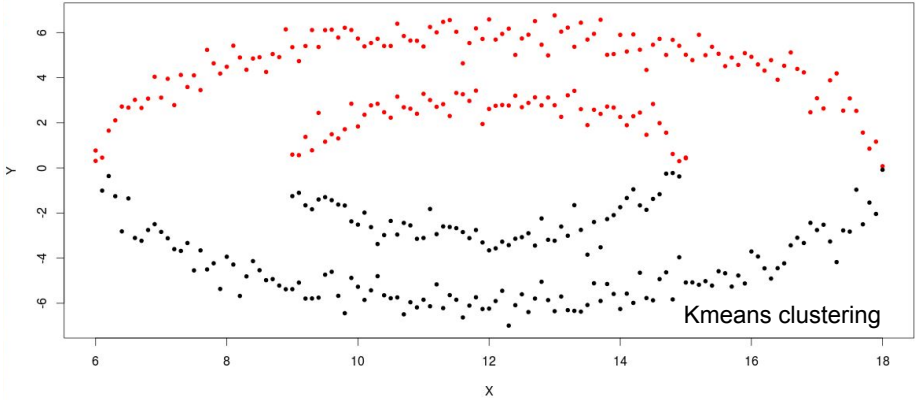


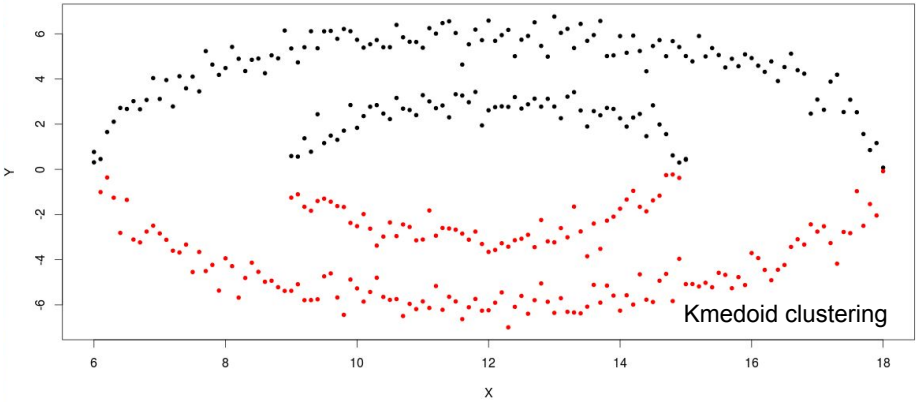
Relook at partition based clustering

This is how the dataset has been clustered using kmeans and kmedoid clustering

Essentially both have done similar clustering

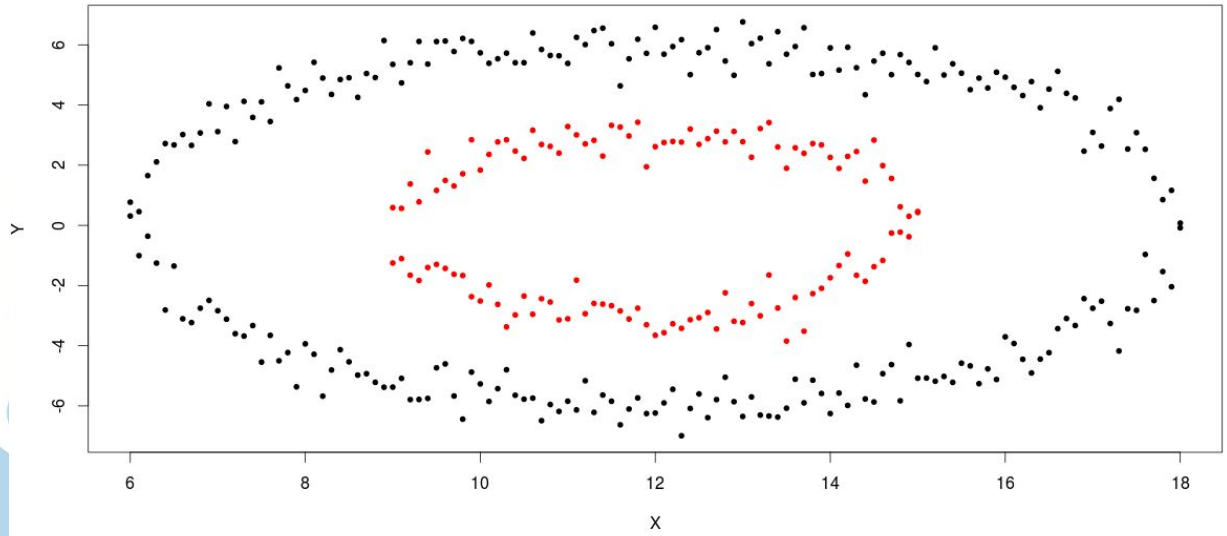
And both have failed to cluster the data points as we expected them to do





Density Based Clustering

However, density based clustering could detect two clusters as we expected



Density Based Clustering (DBSCAN)

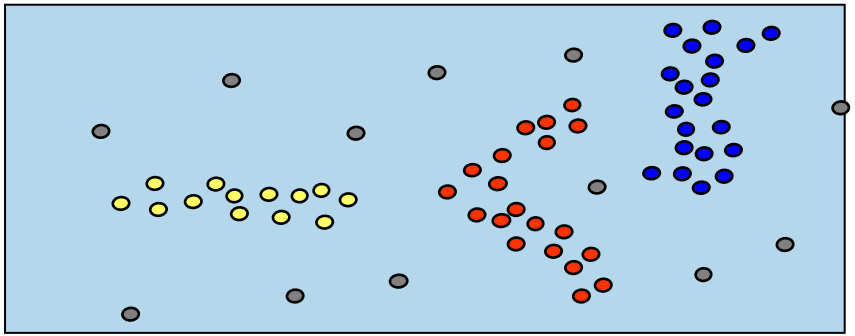
It works on the basis of density connections

Clusters are identified as a set of maximal density-connected points

Unlike K-Means clustering, density based clustering can detect clusters with irregular pattern

The basic idea of such clustering is:

- "Clusters are dense regions in data space separated by regions of low object density"



DBSCAN: Basic Concept

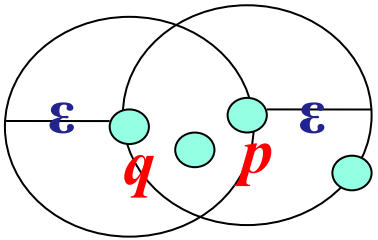
Local point density at a point *p* defined by two parameters

- ε-radius for the neighborhood of point p:
  $N_\varepsilon(p) := \{q \text{ in data set } D \mid dist(p, q) \le \varepsilon\}$
- *MinPts* – minimum number of points in the given neighbourhood *N*(*p*)

ε-Neighborhood – Objects within a radius of *ε* from an object

$$N_\varepsilon(p) : \{q \mid d(p,q) \le \varepsilon\}$$

"High density" - ε-Neighborhood of an object contains at least *MinPts* of objects

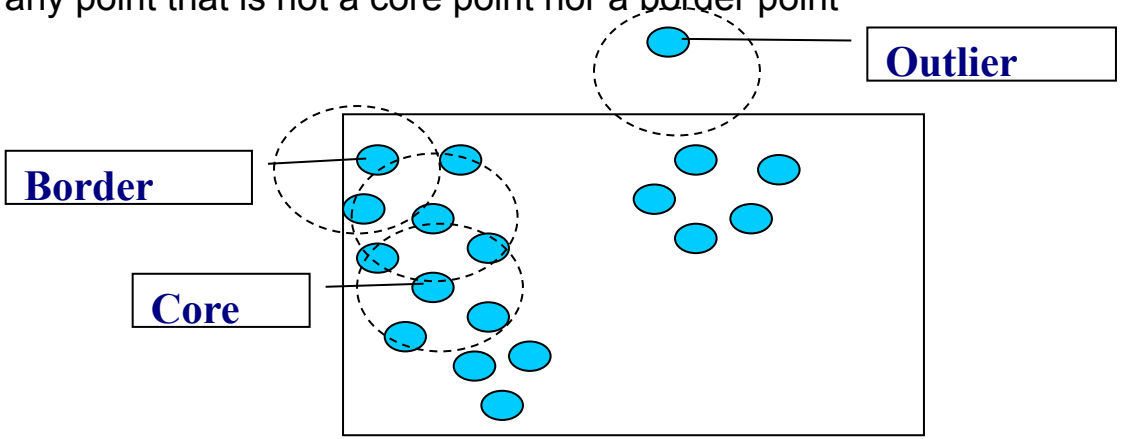


DBSCAN: Basic Concept (Core, Border and Outlier)

A point is a core point if it has more than a specified number of points (MinPts) within Eps These are points that are at the interior of a cluster

A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point

A noise point is any point that is not a core point nor a border point

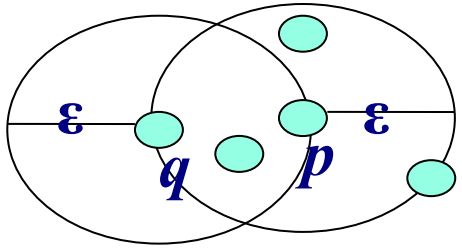


ε = 1unit, MinPts = 5

DBSCAN: Basic Concept (Density-Reachability)

Directly density-reachable

- An object q is directly density-reachable from object p if p is a core object and q is in p's ε-neighborhood



MinPts = 4

q is directly density-reachable from p

p is not directly density- reachable from q

Density-reachability is asymmetric

Two parameters are needed to be decided carefully

- Minimum points
- ε-neighborhood