(III) Density Based Methods.

Need :- Partitioning & hierarchical methods are designed to find spherical-shaped clusters. They have difficulty finding clusters of arbitrary shape such as the "S" shape (for eg.)

Sol :- To find clusters of arbitrary shape, we can model clusters as dense regions in the data space, separated by sparse regions. This is the main idea behind density-based clustering method, which can discover clusters of non spherical shape.

Techniques of Density based clustering :-

(1) DBSCAN : Density based clustering based on connected regions with high Density.
(Density Based Spatial Clustering of Applications with Noise)

(2) OPTICS : (Ordering Points to Identify the Clustering Structure)

(3) DENCLUE : (DENsity-based CLUstering)

(1) DBSCAN: Density Based clustering Based on connected Regions with High Density.

- Density of an object 'o' can be measured by number of objects close to 'o'.
- DBSCAN finds core objects, that is, objects that have dense neighborhoods. It connects core objects & their neighborhoods to form dense regions as clusters.
- A user-specified parameter $\epsilon > 0$ is used to specify the radius of a neighborhood we consider for every object.
- The ϵ-neighborhood of an object 'o' is the space within a radius ϵ centered at 'o'.
- Due to fixed neighborhood size parametrized by ϵ, the density of a neighborhood can be measured simply by the number of objects in the neighborhood.
- DBSAN uses another user-specified parameter, MinPts, which specifies the density threshold of dense regions.
- An object is a core object if the ϵ neighborhood of the object contains at least MinPts objects
- Core objects are pillars of dense regions.
- Given a set, D, of objects, we can identify all core objects w.r.t the given parameters ϵ & MinPts.

- The clustering task is thus reduced to using core objects & their neighborhoods to form dense regions, where the dense regions are clusters

- For a core object 'q' & an object 'p', we say that 'p' is directly density reachable from q if 'p' is within the ϵ-neighborhood of q.

- using directly-density reachable, a core object can bring all objects from its ϵ-neighborhood into a dense region.

How to assemble a large dense region using small dense regions centered by core objects?

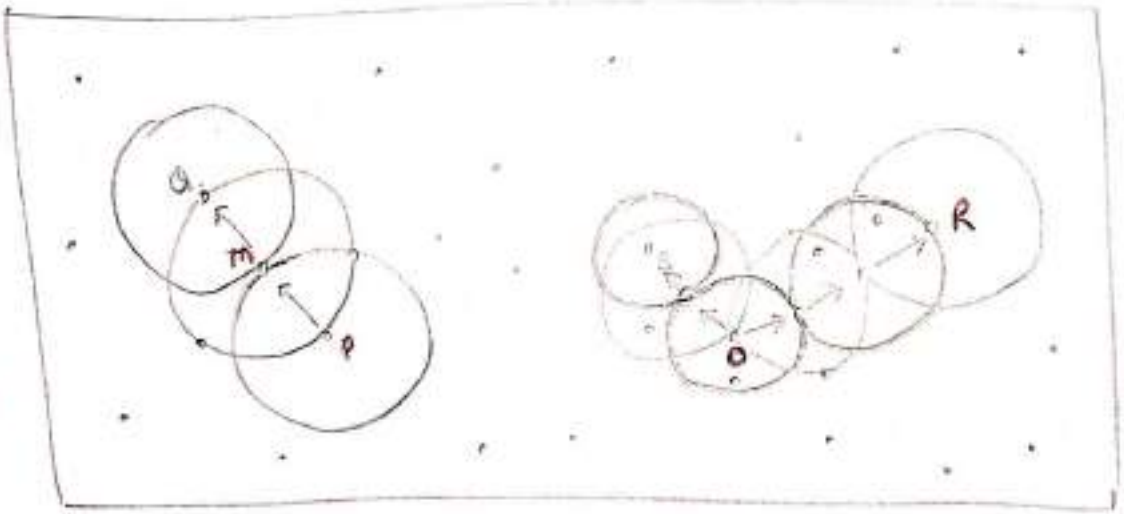
- In DBSCAN, p is density-reachable from q if there is a chain of objects $p_1, \ldots, p_n$ such that $p_1 = q$, $p_n = p$ & and p_{i+1} is directly density reachable from p_i w.r.t. ϵ & MinPts for $1 \le i \le n$, $p_i \in D$

- Density-reachability is not an equivalence relation because it is not symmetric. If both o_1 & o_2 are core objects & o_1 is density-reachable from o_2, then o_2 is density-reachable from o_1. If o_2 is a core object but o_1 is not, then o_1 may be density-reachable from o_2, but not vice-versa.

Density connectedness :- Two objects $p_1, p_2 \in D$ are density connected w.r.t. ϵ & MinPts if there is an object $q \in D$ such that both p_1 & p_2 are density reachable from 'q' w.r.t ϵ & MinPts.

- density connectedness is an equivalence relation.
- Objects O_1, O_2 & O_3 ~~are said~~ can have the following relation. If O_1 & O_2 are density connected & O_2 & O_3 are density-connected, then so are O_1 and O_3.

Eg:- consider figure for a given ϵ rep. by the radius of the circles & let MinPts = 3.



mark in Red - core objects

- of the labeled points M, P, O, R are core objects because each is in an ϵ-neighborhood containing at least three points.
- object Q is directly density-reachable from M.
- object M is directly density-reachable from P & vice-versa.

- Object Q is (indirectly) density-reachable from P because Q is directly density-reachable from M & M is directly density-reachable from P.
- However, P is not density-reachable from Q because Q is not a core object.
- Similarly, R & S are density-reachable from O & O is density-reachable from R. Thus O, R & S are all density-connected.

- We can use the closure of density-connectedness to find connected dense regions as clusters.
- Each closed set is a <u>density-based cluster</u>.

- A subset $C \subseteq D$ is a cluster if

(1). for any two objects $o_1, o_2 \in C$, o_1 & o_2 are density-connected.

(2). there does not exist an object $o \in C$ & another object $o' \in (D - C)$ such that o & o' are density-connected.

How does DBSCAN find clusters?

- Initially all objects in a given data set D are marked as "unvisited".
- DBSCAN randomly selects an unvisited object p, marked p as visited & check whether the ϵ-neighborhood of p contains at least MinPts objects.
- If not, p is marked as noise point.
- Otherwise, a new cluster C is created for p, and all the objects in the ϵ-neighborhood of p are added to a candidate set, N.
- DBSCAN iteratively adds to C those objects in N that do not belong to any cluster.
- In this process, for an object p' in N that carries the label "unvisited", DBSCAN marks it as "visited" & checks its ϵ-neighborhood.
- If the ϵ-neighborhood of p' has atleast MinPts objects those objects in the ϵ-neighborhood of p' are added to N.
- DBSCAN continues adding objects to C until C can no longer be expanded, i.e. N is empty.
- At this time, cluster C is completed & thus is the output.

- To find next cluster, DBSCAN randomly select an unvisited object from the remaining ones.
- The clustering process continues until all objects are visited.
- If spatial index is used, the computational complexity of DBSCAN is $O(n \log n)$, where n is no. of database objects.
- Otherwise, the complexity is $O(n^2)$.

Algorithm: DBSCAN: a density based clustering algorithm.

Input:

D: a data set containing n objects.
ϵ: The radius parameter, and
MinPts: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

(1) Mark all objects as unvisited.
(2) do
(3) randomly select an unvisited object p.
(4) mark p as visited
(5) If the ϵ-neighborhood of p has atleast MinPts objects
(6) Create a new cluster C. Add p to C.
(7) Let N be the set of objects in the ϵ-neighborhood of p.
(8) for each point p' in N
(9) If p' is unvisited
(10) mark p' as visited
(11) If the ϵ-neighborhood of p' has atleast MinPts, add these pts. to N
(12) If p' is not yet a member of any cluster, add p' to C.
(13) end for
(14) O/P C;
(15) else mark p as noise;
(16) until no object is unvisited;

Density based clustering

- Uses the local density of points to determine the clusters, rather than using only the distance b/w points.
- We define a ball of radius ϵ around a point $x \in R^d$, called ϵ-neighborhood of x as follows:

$$N_\epsilon(x) = B_d(x, \epsilon) = \{ y \mid \delta(x,y) \leq \epsilon \}$$

$\{ \delta(x,y) = \|x-y\|_2$ i.e. ass. to be Euc. dista

- Density based methods are able to mine non-convex clusters, where distance based methods may have difficulty.

- **Core point** :- $x \in D$ is a core point if there are atleast 'minpts' points in its ϵ-neighborhood. i.e

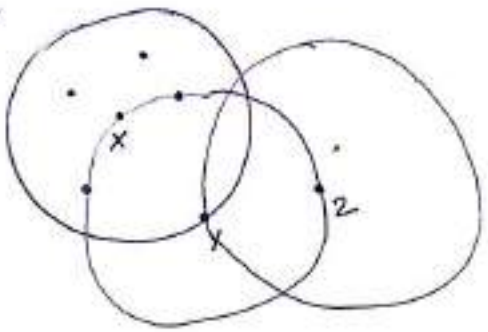
$$|N_\epsilon(x)| \geq \text{minpts}$$

'minpts' is a user-defined local density or frequency threshold.

- **Border point** :- $x \in D$, is a border point that does not meet the minpts threshold i.e. $|N_\epsilon(x)| < \text{minpts}$ but it belongs to the ϵ-neighborhood of some core point z ($z \in D$) i.e

$$x \in N_\epsilon(z).$$

- **Noise point** :- If a point is neither a core nor a border point, then it is called a noise point or an outlier.



minpts = 6
X - core point :- $|N_\epsilon(x)| = 6$
Y - border point :- $|N_\epsilon(y)| <$ minpts but $y \in N_\epsilon(x)$.
Z - noise point.

Directly density reachable :- A point x is directly density reachable from another point y if

- y is a core point and
- $x \in N_\epsilon(y)$.

If y is core point then all other pts in cluster are D.D.R.

Density reachable :- A point x is density reachable from y if there exists a chain of points $x_0, x_1, \ldots x_l$ such that $x = x_0$ & $y = x_l$ and x_i is directly density reachable from x_{i-1} $\forall$ $i = 1$ to l.

OR

There is a set of core points leading from y to x.

Density connected :- Two points x & y are density connected if there exist a core point z, such that both x & y are density reachable from z.

Density based cluster :- is defined as maximal set of density connected points.

DBSCAN(D, ϵ, minpts):

1. core $\leftarrow \emptyset$
2. foreach $x_i \in D$ do // find the core points
3. | compute $N_\epsilon(x_i)$
4. | id(x_i) $\leftarrow \emptyset$ // cluster id for x_i // Till now, not assigned to any cluster i.e $\emptyset$
5. | If $N_\epsilon(x_i) \geq$ minpts then core $\leftarrow$ core $\cup \{x_i\}$
6. $k \leftarrow 0$ // cluster id.
7. for each $x_i \in$ Core, such that id(x_i) = $\emptyset$ do
8. | $k \leftarrow k+1$
9. | id(x_i) $\leftarrow k$ // assign x_i to cluster id k
10. | DensityConnected(x_i, k)
11. $\mathcal{C} \leftarrow \{C_i\}_{i=1}^{k}$ where $C_i \leftarrow \{x \in D \mid id(x) = i\}$
12. Noise $\leftarrow \{x \in D \mid id(x) = \emptyset\}$
13. Border $\leftarrow D \setminus \{core \cup Noise\}$
14. Return $\mathcal{C}$, Core, Border, Noise.

DensityConnected(x, k):

15. for each $y \in N_\epsilon(x)$ do
16. | id(y) $\leftarrow k$ // assign y to cluster id k.
17. | if $y \in$ Core then DensityConnected(y, k).

Complexity.

- main cost is computing ϵ-neighborhood for each point.
- If dimensionality is not too high → done efficiently using a spatial index structure in $O(n \lg n)$ time.
- If dimensionality is high, it takes $O(n^2)$ to compute $N_\epsilon(x)$, $x \in D$ (for each point)
- Once $N_\epsilon(x)$ computed, only a single pass is needed to find density connected clusters. $\therefore O(n^2)$ in worst case.

Drawback

- DBSCAN is sensitive to the choice of ϵ, if clusters have different densities.
- If ϵ is too small, sparses cluster will be categorized as noise.
- If ϵ is too large, denser clusters may be merged together.

ie If there are clusters with different local densities, then a single ϵ-value may not be sufficient.