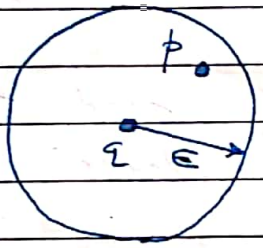
* Density Based Clustering
- Partitioning & hierarchical methods are designed to find spherical clusters. They have difficulty finding clusters of arbitrary shape.
- To find clusters of arbitrary shape, we can model clusters as dense regions in the object space, separated by sparse regions. This is the main idea behind density based clustering.

→ DBSCAN (Density Based Spatial Clustering of Applications with Noise)

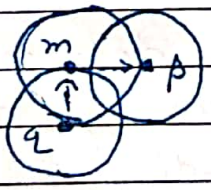
- The density of an object o can be measured by the no. of objects close to o.
- The algo find core objects, the objects whose neighborhoods are dense. The core objects & their neighborhoods are connected to form dense regions as clusters.
- The algo. has two user defined parameters:
(i) $\epsilon$-neighborhood to specify neighborhood of an object o with radius $\epsilon$ centered at o.
(ii) Minpts to specify the density threshold of a dense region.
- An object is "core object" if its $\epsilon$-neighborhood has at least Minpts no. of objects.

- Given a set D, of objects, the algo. first identifies core objects using ∈ & Minpts.
- These core objects & their neighborhoods are used to form dense regions.

- For a core object q & an object p, it is said that p is "directly density reachable" from q if p is within ∈-neighborhood of q.

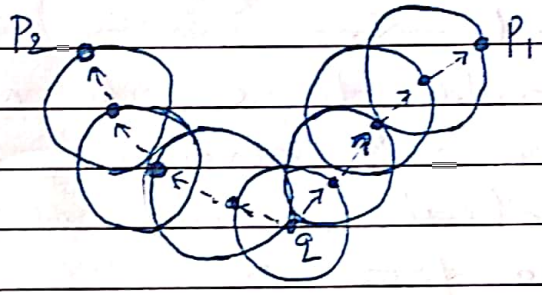



- An object p is "density reachable" from a core object q, if there is a chain of objects $p_1, p_2, \ldots p_n$ such that $p_1 = q$, $p_n = p$ and $p_{i+1}$ is directly density reachable from $p_i$, $1 \le i \le n$, w.r.t. ∈ & Minpts.




- m is directly density reachable from q & p is directly density reachable from m, so p is density reachable from q.
- If q & p both are core objects then they both are density reachable from each other But if q is core object & p is not then p is density reachable from q but q is not from p.

- Two objects $p_1$ & $p_2$ are density connected, if there is an object q such that both $p_1$ & $p_2$ are density reachable from q, w.r.t. $\epsilon$ & MinPts




- A subset C of dataset D is a dense region & so a cluster, if (i) for any two objects $O_1$ & $O_2 \in C$, $O_1$ & $O_2$ are density connected & (ii) there exists no object $O \in C$ & another object $O' \in (D - C)$ such that O & O' are density connected.

- The algo works as follows:
(i) Initially all the objects in the given set D as "unvisited".
(ii) An unvisited object p is selected randomly.
(iii) p is now marked visited
(iv) If p is not a core object then it is marked as noise point.
(v) If p is a core object then a new cluster C is created for p & all the objects in $\epsilon$-neighborhood of p are added to a candidate set N.

(vi) for each object p' in N, which is unvisited, algo labels it visited & checks if it is core object.
(vii) if p' is core object then all the objects in ∈-neighborhood p' are added to N.
(viii) algo adds to c, to objects of N that are not already added to some cluster. such objects are removed from N.
(ix) The process is repeated until N is empty. The cluster c is completed.
(x) To find a next cluster, algo randomly selects an unvisited object from the remaining objects.
(xi) The clustering process is continued until all the points are visited.

i/p: D // set of objects
∈ // radius
Minpts // density threshold
o/p: set of clusters.
steps: (i) mark all points as unvisited
(ii) do
(iii) randomly select an unvisited pt p
(iv) mark p as visited
(v) if p is core point
(vi) create new cluster c & add p to c
(vii) let N be set of objects in ∈-neighborhood of p
(viii) for each p' in N
(ix) if p' is unvisited
(x) mark p' as visited

(xi) if p' is a core object
(xii) add all objects in ε-neighborhood of p' to N
(xiii) if p' is not member of any cluster then add p' to C
(ixv) end for
(xv) output C
(xvi) else mark p as noise point
(xvii) until all points are marked visited

- The time complexity of the algo, without using index is $O(n^2)$ where n is no. of objects.

// Solving example is not possible, as it is for large no. of objects. //