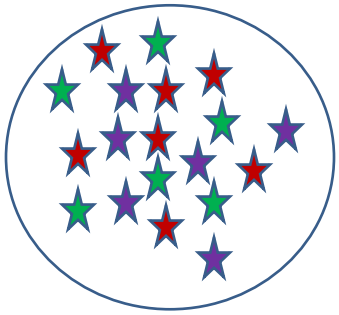
Cluster Analytics

Definition : Given a collection of data objects group them so that

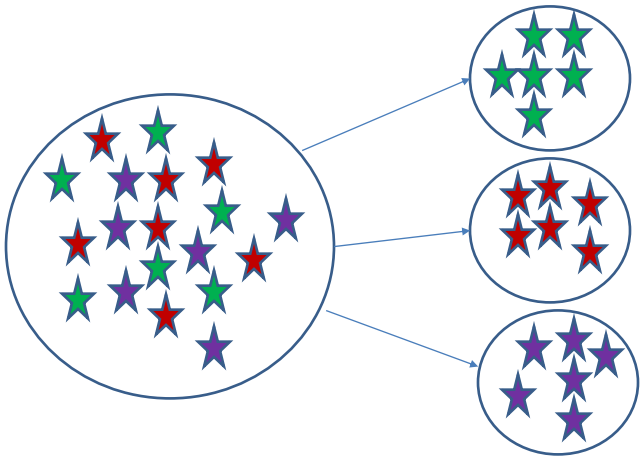
- Similar to the objects within the same cluster(group)
- Dissimilar to the objects in other clusters(groups)

Data objects can be set of web pages, set of emails or set of states in India

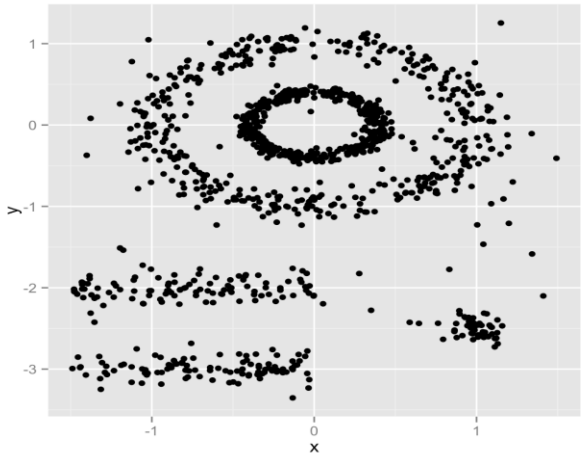


Definition : Given a collection of data objects group them so that

- Similar to the objects within the same cluster(group)
- Dissimilar to the objects in other clusters(groups)

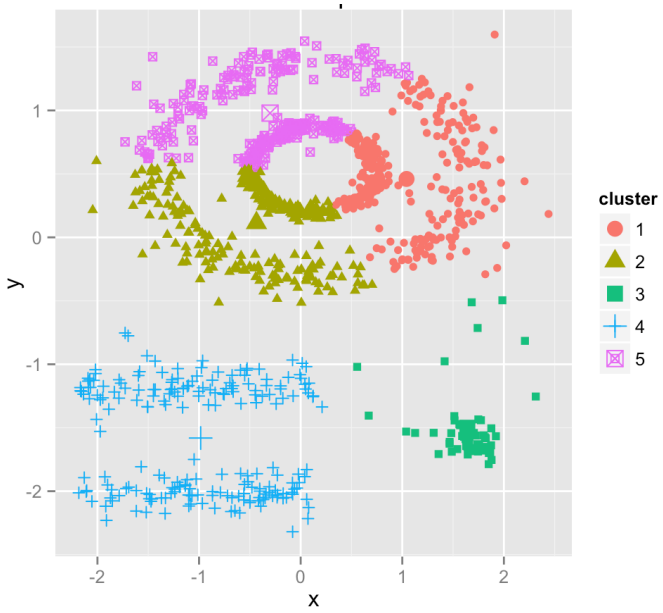


DBSCAN : Density-based algorithm



The plot above contains 5 clusters and outliers

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



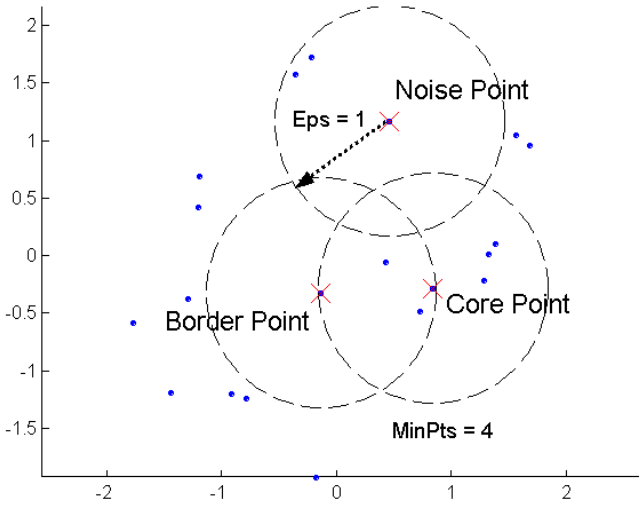
Density Based Clustering

- Clusters are dense regions in the data space, separated by regions of lower object density
- A cluster is defined as a maximal set of density connected points
- Discovers clusters of arbitrary shape and size

- We need to provide two parameters *Eps, MinPts* for this algorithm
- Density of a point is defined as the number points within a specified radius(Eps).
- This algorithm divides the points into three groups based on density

- Core point : A point is a core point if it's density is more than or equal to specified number of points (MinPts)
- Border point : A point is a border point if it's density is less than MinPts, but is in the neighborhood of a core point
- Noise point : A point is a noise point, if it is not a core point or a border point.

Core, Border, and Noise Points



DBSCAN Algorithm

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.

DBSCAN Algorithm

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.

Determining EPS and MinPts

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbor at farther distance
- So, plot sorted distance of every point to its k^{th} nearest neighbor

Point Number	Distance
1	4
2	3
3	8
4	40
5	6
6	36
7	8
8	30
9	4
10	3

Determining EPS and MinPts

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbor at farther distance
- So, plot sorted distance of every point to its k^{th} nearest neighbor

Point Number	Distance	X Value
2	3	1
10	3	2
1	4	3
9	4	4
5	6	5
3	8	6
7	8	7
8	30	8
6	36	9
4	40	10

Determining EPS and MinPts

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbor at farther distance
- So, plot sorted distance of every point to its k^{th} nearest neighbor

Point Number	Distance	X Value
2	3	1
10	3	2
1	4	3
9	4	4
5	6	5
3	8	6
7	8	7
8	30	8
6	36	9
4	40	10