What's clustering

- Group data such that within group variance < between group variance

- Group such that objects in the same group are more similar to each other in some sense than to objects of different groups.

- These groups are known as clusters and each cluster gets distinct label called cluster ID and the centroid of cluster.

- Clustering types:

  - **Centroid based(K-Means)**

  - **Hierarchical clustering (Agglomerative)**

  - **Density based (DBSCAN)**

K-means

- need to specify the number of clusters-K

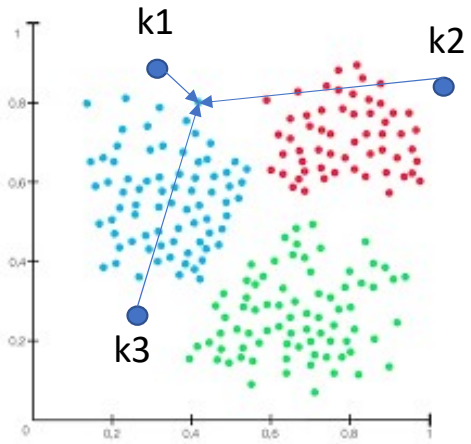
- Step-1: create arbitrarily ‘k’ number of centroids, assign labels k1,k2,…

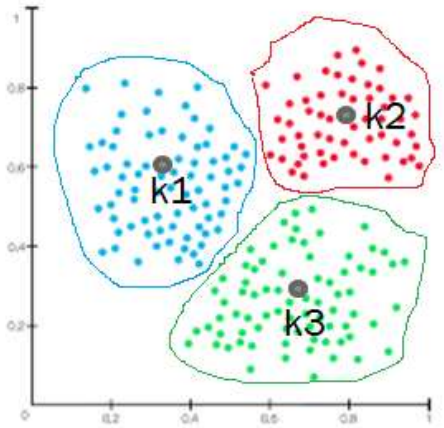
- Step-2: calculate the distance of each of the ‘n’ points from each of the ‘k’ centroids

- Step-3: assign each point to nearest centroid (based on distance)

- Step-4: calculate the new centroids based on the records belonging to k*i*

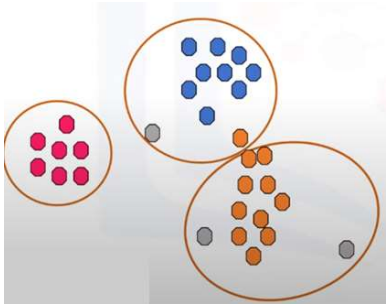
- *Step-5:* check did the cluster index of the points change or iterations less than max: yes→ go to step-2. no→ stop





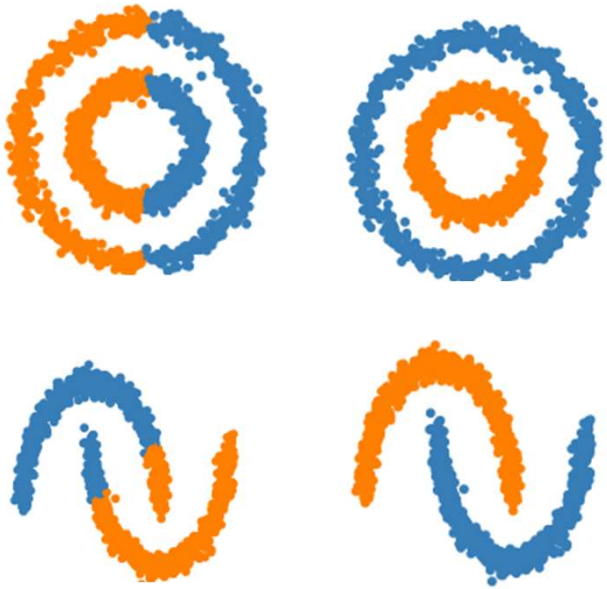
Problem with K-means

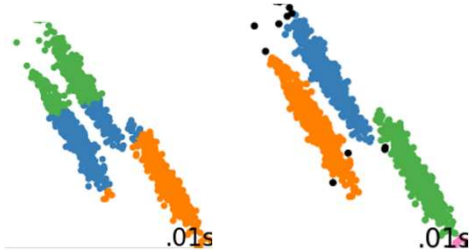
- Convex shape
- Isotropic variance
- outliers



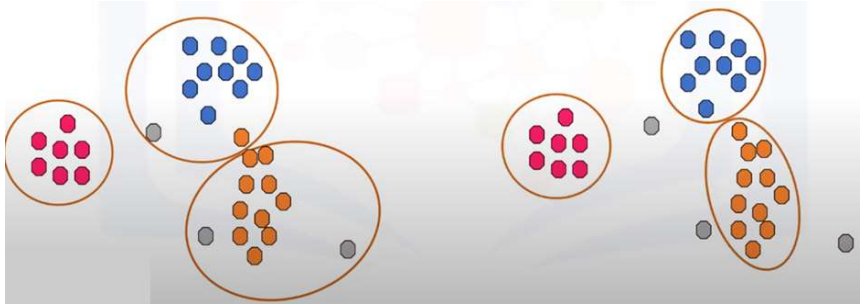
DBSCAN

- Density Bases Spatial Clustering of Applications with Noise





Better at outlier handling



DBSCAN

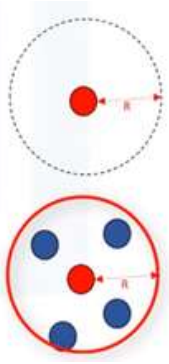
- Density Bases Spatial Clustering of Applications with Noise
- Eps: Radius of neighborhood
- n: Number of neighbors

(**R**adius of neighborhood)

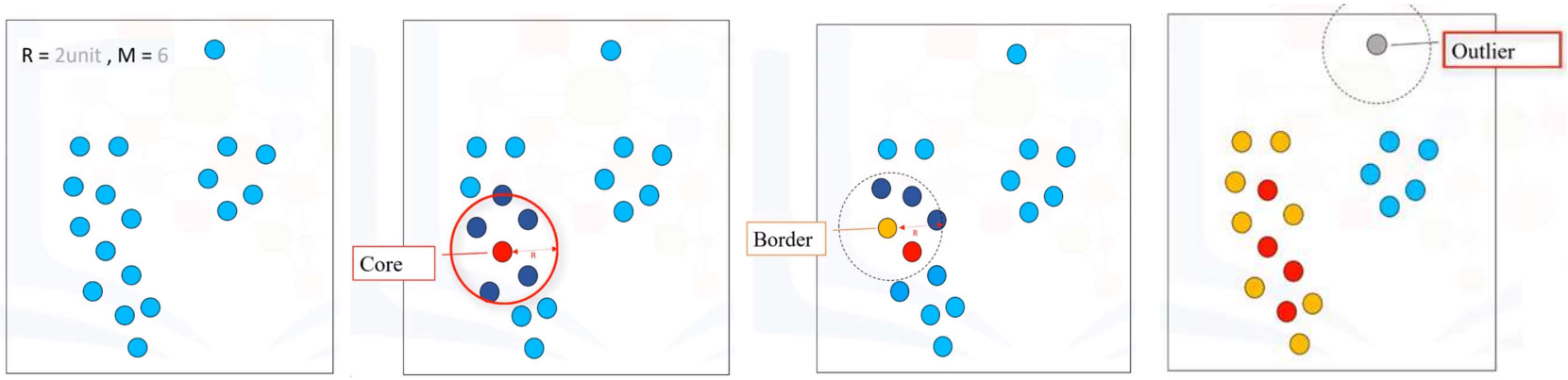
- Radius (R) that if includes enough number of points within, we call it a dense area

(**M**in number of neighbors)

- The minimum number of data points we want in a neighborhood to define a cluster



DBSCAN



Each point is either:
- *core point*
- *border point*
- *outlier point*

Number of points =M in R

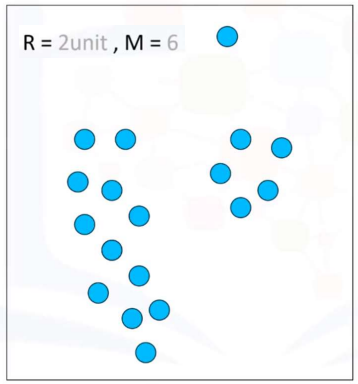
Number of points <M or Reachable from some core point

Number of points in R=0 or not Reachable from some core point

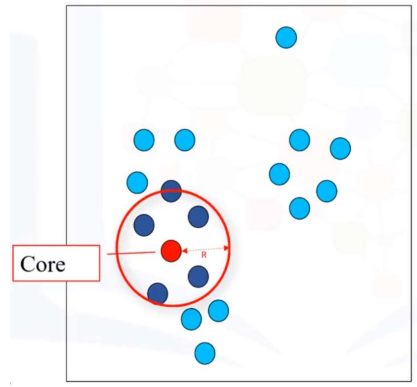
Group points as clusters based on there types

https://www.youtube.com/watch?v=6jl9KkmgDIw

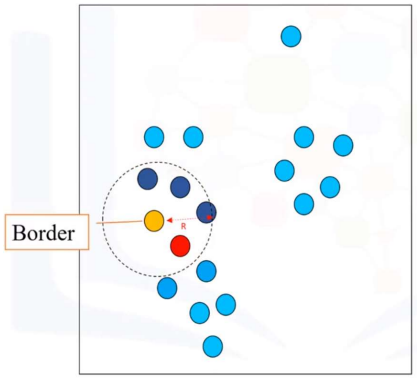
DBSCAN



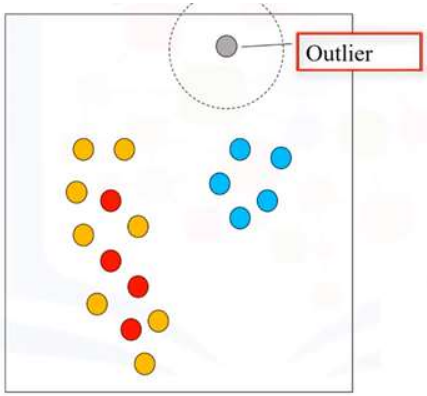
Each point is either:
- *core point*
- *border point*
- *outlier point*



Number of points =M in R

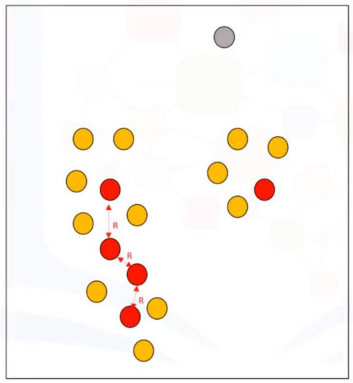


Number of points <M or Reachable from some core point

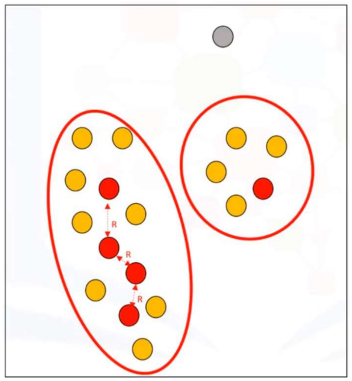


Number of points in R=0 or not Reachable from some core point

Group points as clusters based on there types

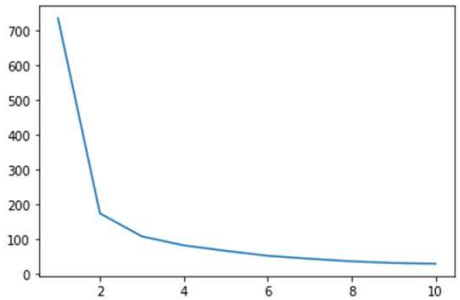


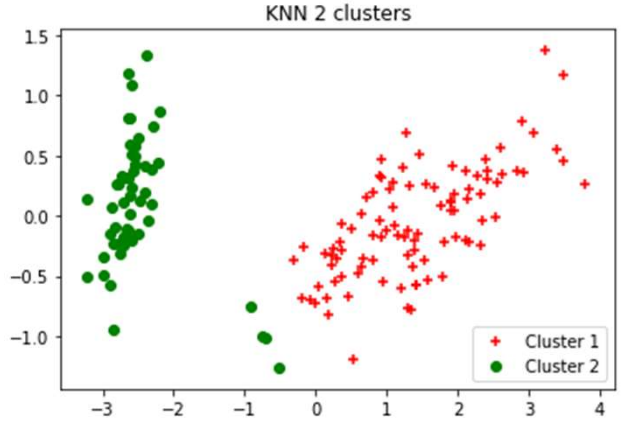
All core reachable core points in a cluster

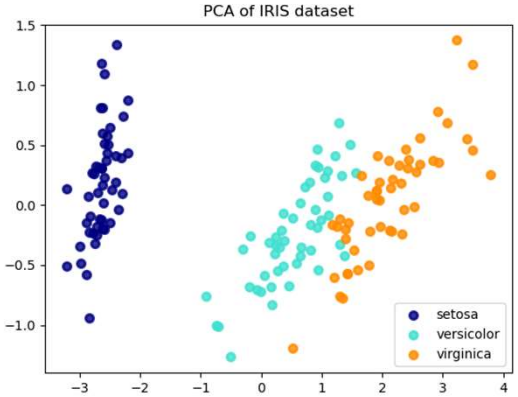


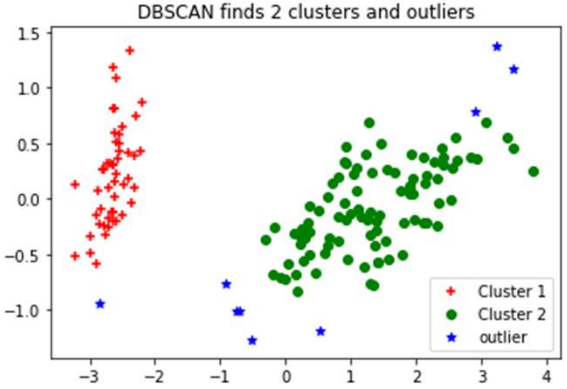
1 core point + reachable core points+ border points= 1 cluster

https://www.youtube.com/watch?v=6jl9KkmgDIw


700
600
500
400
300
200
100
0
2
4
6
8
10


KNN 2 clusters
Cluster 1
Cluster 2


PCA of IRIS dataset
setosa
versicolor
virginica


DBSCAN finds 2 clusters and outliers
Cluster 1
Cluster 2
outlier