CLUSTERING

Identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known

Algorithms

- KMeans
- DBSCAN

KMeans

- Divide into k number of clusters

Algorithm

- Choose k random centroids for cluster
- For each point
  - Find the distance between each clusters
  - Assign the point to the nearest cluster
  - Recompute new centroid for the cluster

  Iterate until convergence

  Return when all the points are computed.

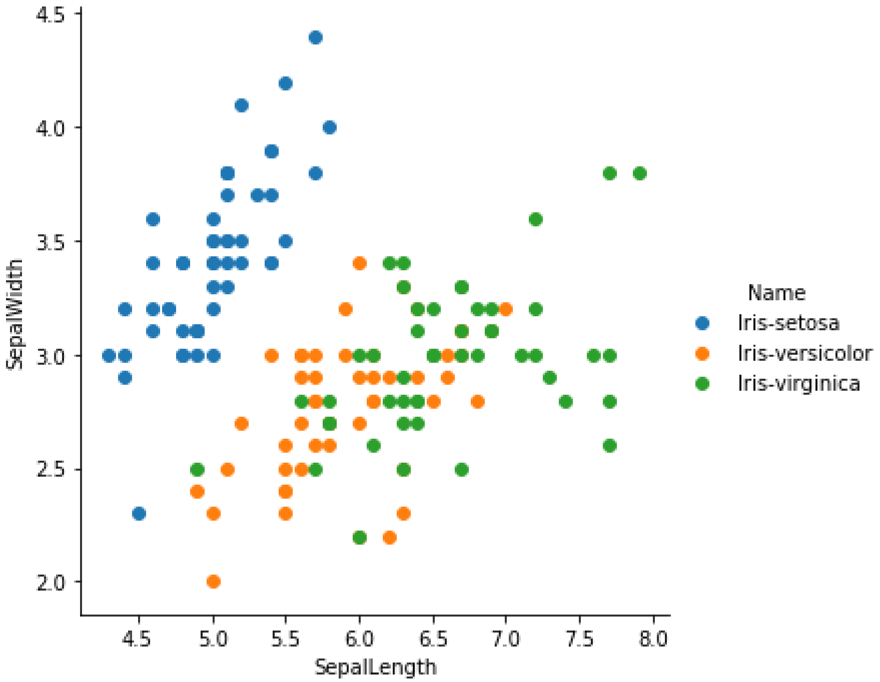
Advantages

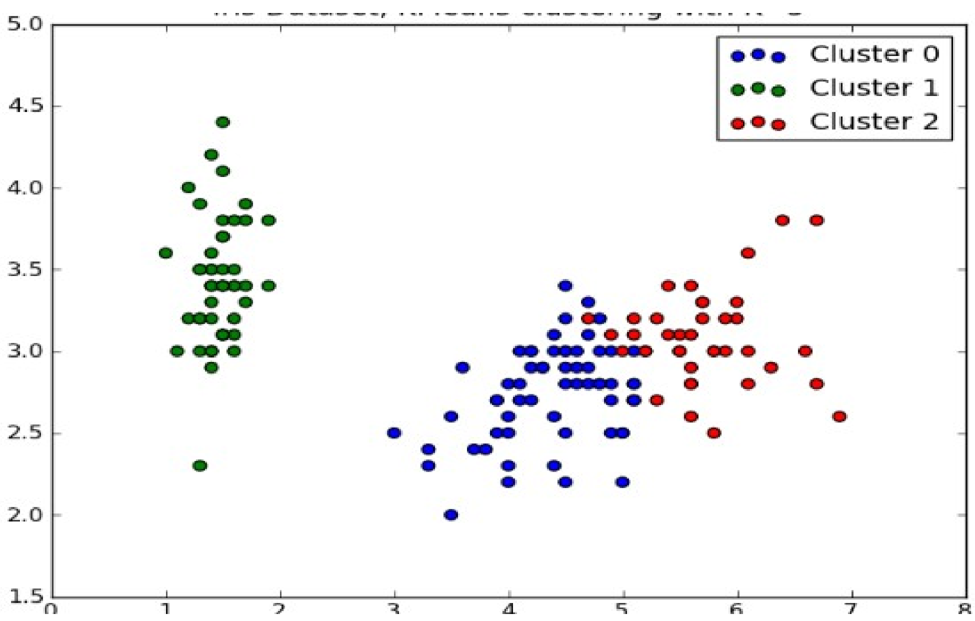
- Efficient for small dataset
- Simple algorithm
- Does not depend upon
- Gives exactly k clusters
- Final result does not depend upon the order of data

Disadvantages

- Calculation heavy
- No notion of noise. So includes outliers.

Result


4.5
4.0
3.5
3.0
2.5
2.0
SepalWidth
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
SepalLength
Name
Iris-setosa
Iris-versicolor
Iris-virginica


5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
1
2
3
4
5
6
7
8
Cluster 0
Cluster 1
Cluster 2

DBSCAN

- Density-based spatial clustering of applications with noise
- Winner of test of time award 2014

Algorithm

- Find the points in the ε (eps) neighborhood of every point, and identify the core points with more than minPts neighbors.
- Find the connected components of core points on the neighbor graph, ignoring all non-core points.
- Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbor, otherwise assign it to noise.

Advantages

- No need to specify number of clusters
- Can find arbitrarily shaped clusters
- Can identify outliers as noise
- Mostly insensitive to order points
- Eps and minpts can be set by domain experts

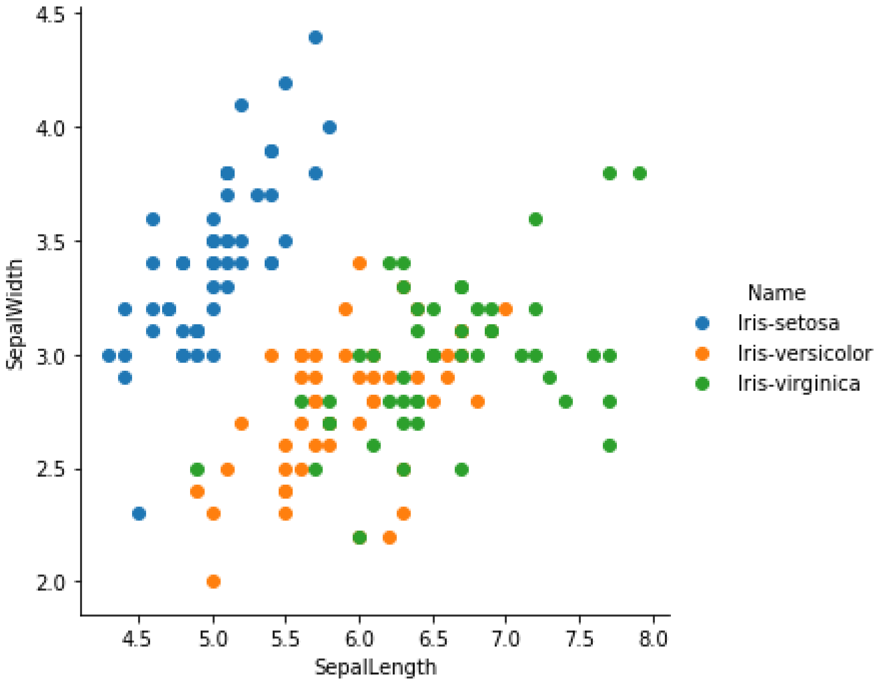
Disadvantages

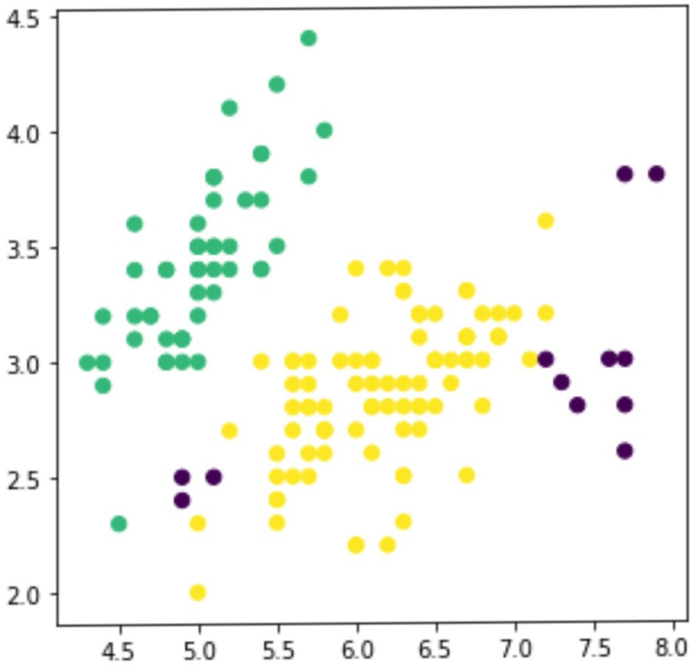
- Not deterministic: Border points can switch clusters
- Depends on distance measure. Mostly euclidean distance.
- W/o understanding data well, choosing eps and minpts is quite daunting.

Dataset

- Famous Iris Dataset
- Parameters: SepalLength, SepalWidth, PetalLength, PetalWidth, Name
- Number of records: 150
- minPts $\geq$ Dimensions + 1 (at least 5)
- Eps: choose small value for better clustering

Result


4.5
4.0
3.5
3.0
2.5
2.0
SepalWidth
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
SepalLength
Name
Iris-setosa
Iris-versicolor
Iris-virginica


4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0