CSE 6363 Machine Learning

Project Report

K means Clustering

Introduction:

K means clustering is a method of unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to cluster with the nearest mean serving as prototype of the cluster. (Source Wikipedia).

The Algorithm:

The K means clustering algorithm is often referred to as Lloyd's algorithm. Since the initial choice of centroids is random in our algorithm each simulation of the K means yields a different result.

K-Means

Algorithm

Input – Desired number of clusters, *k*

Initialize – the *k* cluster centers (randomly if necessary)

Iterate –

1. Decide the class memberships of the *N* objects by assigning them to the nearest cluster centers

2. Re-estimate the *k* cluster centers (aka the centroid or mean), by assuming the memberships found above are correct.

$$\vec{\mu}_k = \frac{1}{C_k} \sum_{i \in C_k} \vec{x}_i$$

Termination –

If none of the *N* objects changed membership in the last iteration, exit. Otherwise go to 1.

The Dataset

Though the K means is an unsupervised learning algorithm which is mostly used on unlabeled data we use **Iris dataset**, strip off the labels and use it to illustrate K means clustering. We then conclude the optimal choice of the number of clusters (K) based on results.

The dataset consists of 4 feature vectors, sepal length, sepal width, petal length, petal width and there are 150 data points.

The Result:

K = 1

It won't make any sense to choose only one cluster because all data points will belong to the same cluster.

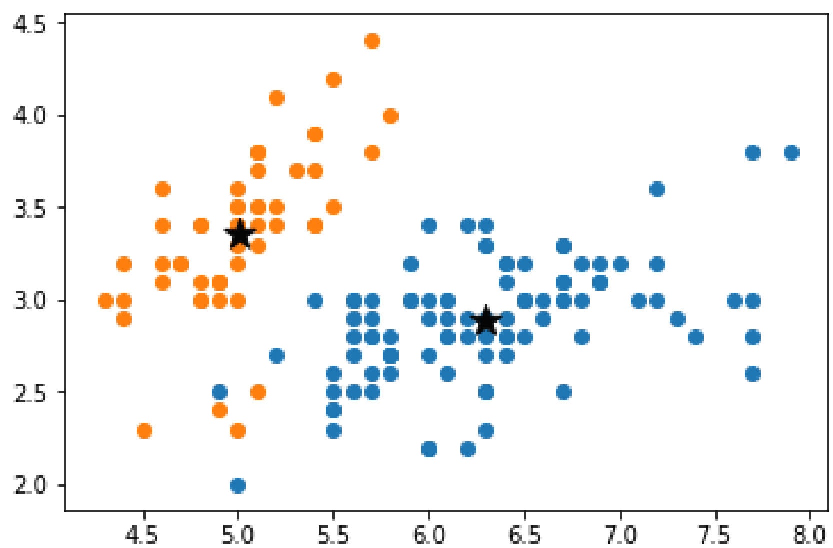
K = 2

The cluster labels (0 to K-1) for the iris dataset is: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

The cluster centers are: [[6.30103093, 2.88659794, 4.95876289, 1.69587629]

[5.00566038, 3.36037736, 1.56226415, 0.28867925]]

-----------------------------The cluster plot---------------------------



K = 3

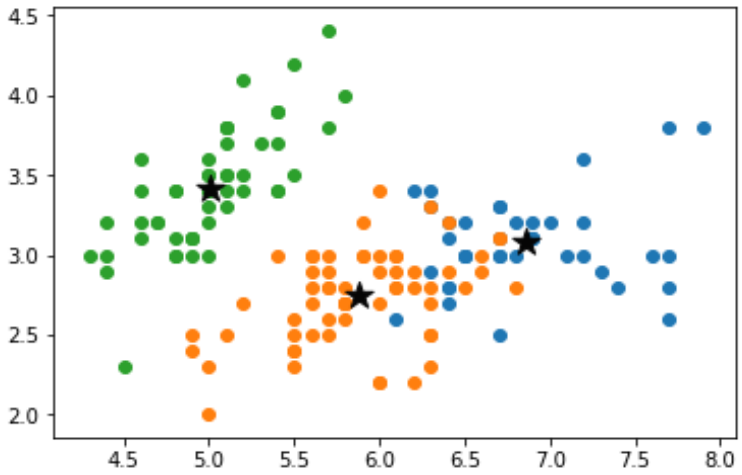
The cluster labels (0 to K-1) for the iris dataset is: [2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1]

The cluster centers are: [[6.85384615, 3.07692308, 5.71538462, 2.05384615]

[5.88360656, 2.74098361, 4.38852459, 1.43442623]

[5.006, 3.418, 1.464 , 0.244]]

----------The cluster plot----------



K = 5

The cluster labels (0 to K-1) for the iris dataset is: [4, 2, 2, 2, 4, 4, 2, 4, 2, 2, 4, 2, 2, 2, 4, 4, 4, 4, 4, 4, 4, 4, 2, 4, 2, 2, 4, 4, 4, 2, 2, 4, 4, 4, 2, 2, 4, 2, 2, 4, 4, 2, 2, 4, 4, 2, 4, 2, 4, 2, 1, 1, 1, 3, 1, 3, 1, 3, 1, 3, 3, 3, 3, 1, 3, 1, 1, 3, 1, 3, 1, 3, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 1, 3, 1, 1, 1, 3, 3, 3, 1, 3, 3, 3, 3, 3, 1, 3, 3, 0, 1, 0, 0, 0, 0, 3, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1]

The cluster centers are: [[6.9125 , 3.1 , 5.846875 , 2.13125]

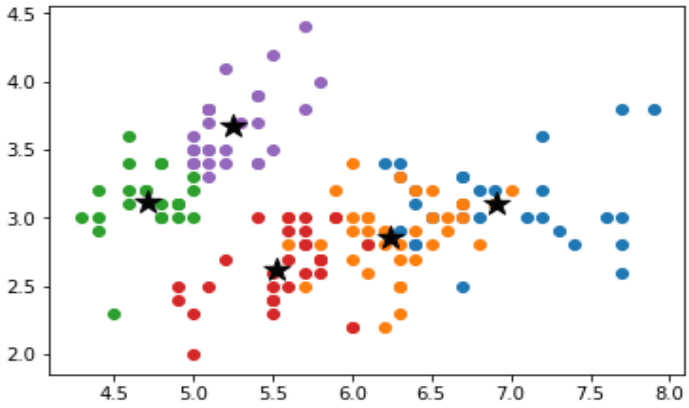
[6.23658537, 2.85853659 , 4.80731707 1.62195122]

[4.71304348 , 3.12173913 , 1.4173913 , 0.19130435]

[5.52962963 , 2.62222222 , 3.94074074, 1.21851852]

[5.25555556 , 3.67037037 , 1.5037037 , 0.28888889]]

----------The cluster plot----------



K = 7

The cluster labels (0 to K-1) for the iris dataset is: [0, 1, 1, 1, 0, 4, 6, 0, 2, 1, 4, 6, 1, 2, 4, 4, 4, 0, 4, 4, 0, 0, 6, 0, 0, 1, 0, 0, 0, 1, 1, 0, 4, 4, 1, 1, 0, 1, 2, 0, 0, 2, 2, 0, 4, 1, 4, 1, 4, 0, 5, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 5, 3, 5, 5, 5, 5, 3, 5, 5, 5, 5, 5, 5, 3, 3, 5, 5, 5, 5, 3, 5, 3, 5, 3, 5, 5, 3, 3, 5, 5, 5, 5, 5, 3, 5, 5, 5, 5, 3, 5, 5, 5, 3, 5, 5, 5, 3, 5, 5, 3]

The cluster centers are: [[5.11764706 3.45294118 1.50588235 0.28823529]

[4.81538462 3.1   1.45384615 0.17692308]

[4.4   2.88   1.28   0.2   ]

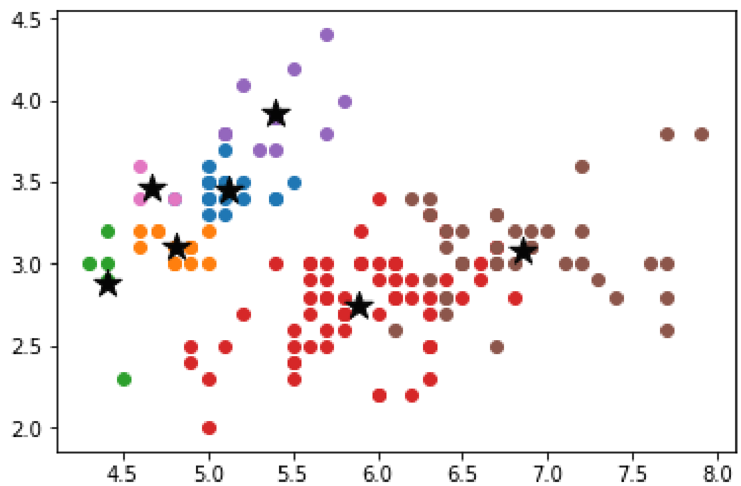
[5.88360656 2.74098361 4.38852459 1.43442623]

[5.39166667 3.925   1.525   0.275   ]

[6.85384615 3.07692308 5.71538462 2.05384615]

[4.66666667 3.46666667 1.33333333 0.23333333]]

------------------------------The cluster plot----------------------------



K = 9

The cluster labels (0 to K-1) for the iris dataset is: [7, 1, 2, 2, 7, 4, 6, 7, 8, 1, 0, 7, 1, 8, 4, 4, 4, 7, 4, 0, 0, 0, 6, 7, 7, 1, 7, 0, 7, 2, 1, 0, 4, 4, 1, 1, 0, 1, 8, 7, 7, 8, 2, 7, 0, 1, 0, 2, 0, 7, 5, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 5, 3, 5, 5, 5, 5, 3, 5, 5, 5, 5, 5, 5, 3, 3, 5, 5, 5, 5, 3, 5, 3, 5, 3, 5, 5, 3, 3, 5, 5, 5, 5, 5, 3, 5, 5, 5, 5, 3, 5, 5, 5, 3, 5, 5, 5, 3, 5, 5, 3]

The cluster centers are: [[5.26   3.63   1.55   0.27   ]

[4.88888889 3.06666667 1.45555556 0.16666667]

[4.6   3.18   1.42   0.2   ]

[5.88360656 2.74098361 4.38852459 1.43442623]

[5.52857143 4.04285714 1.47142857 0.28571429]

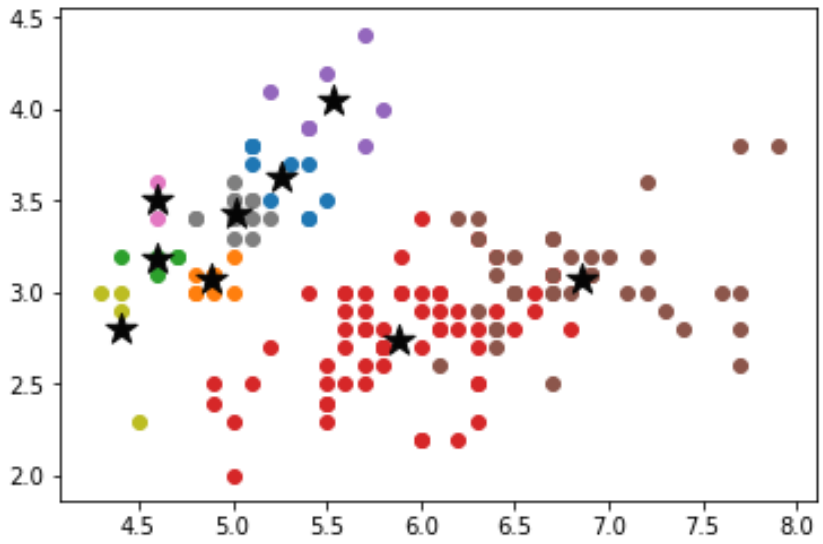
[6.85384615 3.07692308 5.71538462 2.05384615]

[4.6   3.5   1.2   0.25   ]

[5.01538462 3.43076923 1.51538462 0.28461538]

[4.4   2.8   1.275   0.2   ]]

----------The cluster plot----------



Conclusion

For the optimal choice of the number of cluster we are looking for a choice of K in which the intra cluster distance is minimized and inter cluster distance is maximized, From out results we see that K = 3 will be an optimal choice of the number of clusters.