Abdul Rehman Liaqat

Lab Course Machine Learning

Exercise Sheet 11

Report

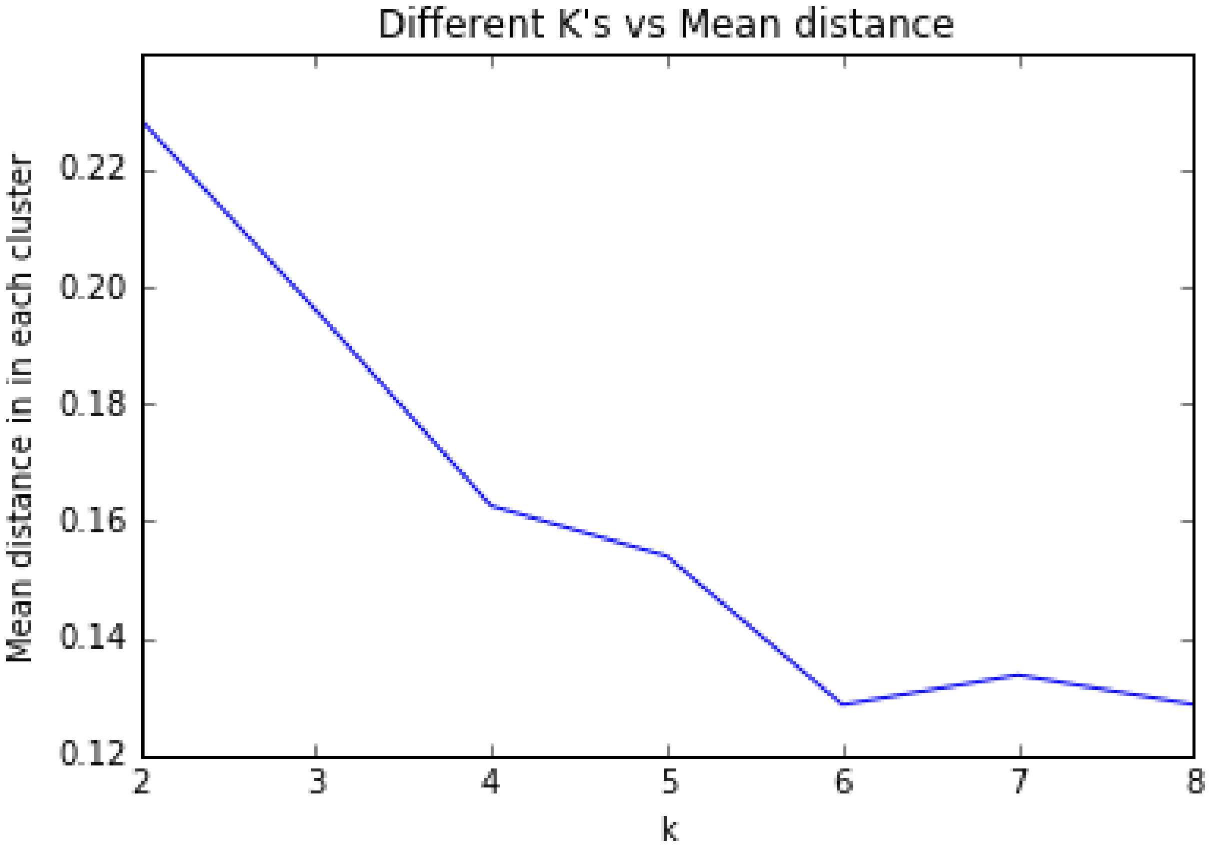
Exercise 1: Implement K-Means clustering algorithm

Following steps were taken to implement K-Means algorithm:

1- Normalized whole dataset using ‘normalize’ function and shuffled it.
2- Picked a uniformly random point as an initial centroid.
3- Calculated k-1 centroid points which are basically at the maximum distance from each other.
4- Classified the dataset into clusters around respective centroids.
5- And calculated the centroids again from the respective clusters.
6- Repeated steps 4-5 till convergence that is till two steps 5 in consecutive iteration results the same value.
7- During this exercise following extra things were considered:
    - For each value of k, 10 times random restart of the whole algorithm was done. This way maximum effort to obtain global optimum was applied.
    - K values of 2,3,4,5,6,7 and 8 were tested.
    - Labels of the dataset were dropped since we are performing unsupervised learning.

Iris Dataset was considered for this part and all the procedure and results are obtained through Iris dataset. Although this algorithm will equally work for the other dataset (sparse) since the written program handles each row separately it will not make any difference if the data is sparse.

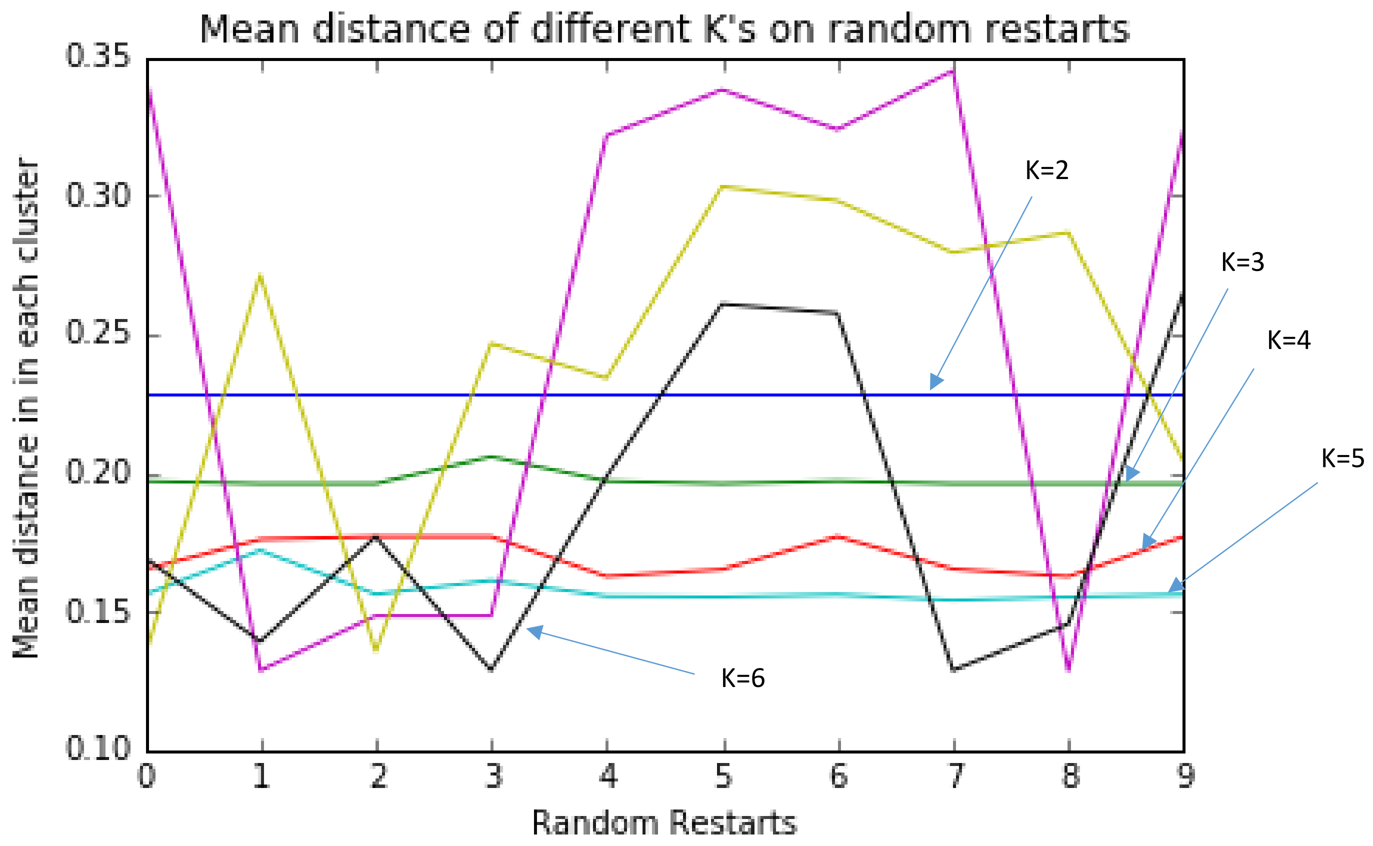
After performing all the steps as described following results were obtained:



This is the standard graph of K-means clustering. Since Iris data set was actually labeled dataset and we know that it was divided into four classes hence it is obvious that k=4 will provide the best results in our unsupervised learning. And the plot above is the proof of it. After k=4 there is no extraordinary decrease in mean distance.

The reason of the above graph not being so uniform is that for each value of k among 10 random restarts least value of mean distance was selected. Hence it is quite possible that for different k’s data was shuffled and totally different initial points were selected.

Abdul Rehman Liaqat



Above graph is another way of comparing in the change of mean distance with increase in k. Each color represents a value of k and the values of k's are increasing from top to bottom. Distance between k=4 and k=3 is far greater that distance between k=4 and k=5.

Also after k=5 further increase in k is producing empty clusters for many values of random restarts. Hence the reason of so much noise in the graph. When an empty cluster is formed the previous mean distance of the cluster is considered and which is obviously greater than the converged lesser k valued algorithms.

Abdul Rehman Liaqat

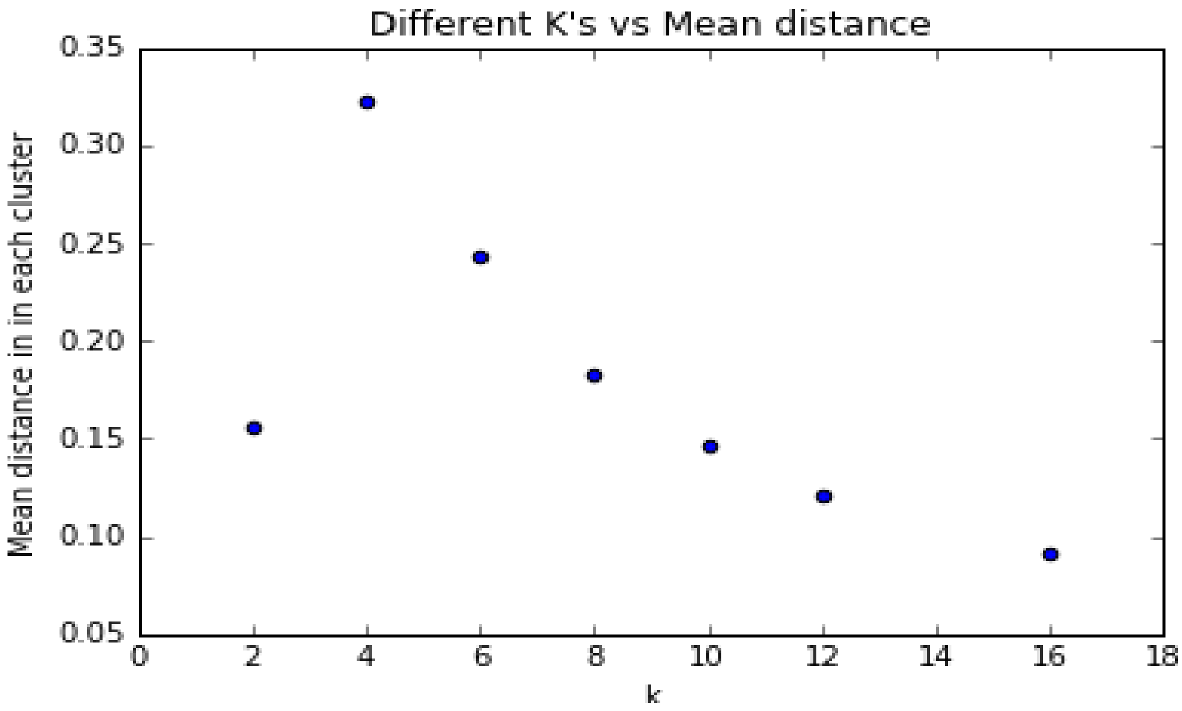
Exercise 2: Cluster news articles

To perform k-means clustering on news articles pre-processing is needed. Pre-processing is done as following:

1- Content of each file in a sub-folder i.e. “alt.atheism” is combined and itdf is performed on the content. Achieved itdf vector is sorted and top 100 words (not considering numeric) are picked.
2- Similar procedure is repeated for each folder and at the end all of these itdf vectors are combined.
3- This way a vocabulary list of top frequency words in each sub-folder is achieved.
4- Next step is to perform itdf on each file i.e. ‘49960’ against the vocabulary list achieved in the previous step. This way a itdf vector of the similar length as vocabulary list is achieved. This itdf vector describes frequency of words from the vocabulary list, inside the respective file.
5- Same procedure is repeated for each file.
6- This way we are able to achieve a data frame whose columns are words from vocabulary lists and each row describe one file.

Above procedure is performed for both test and training data. Hence data now ready to be processed.

In the next step K-Means clustering is performed as done in the previous question. Only difference is the range of K’s tested which are [2,4,6,8,10,16,20]. Means distance for each K provide us following plot showing decrease in mean distance with the increase in K.


Different K's vs Mean distance
Mean distance in in each cluster
k