SENG 474
Yidan Long
V00898887
July 30 2020

Assignment 3

I.Lloyd's Algorithm (k-means)

1.1 Uniform Random Initialization

I tried the different k values which are starting from 2 to 7 with an increment of 1 for both uniform random initialization and k means ++ initialization.

1.1.1 Dataset1 Graph and Analysis

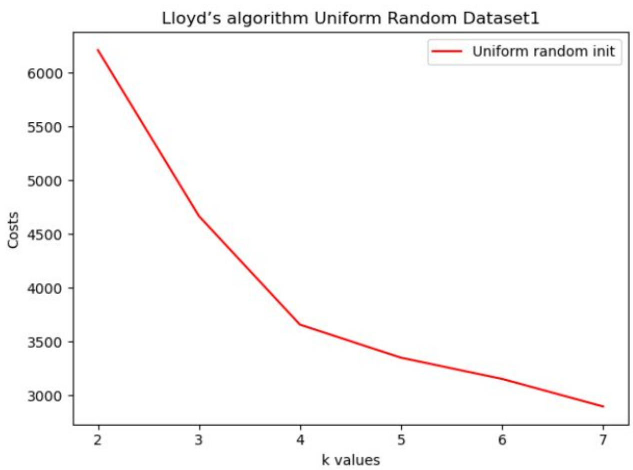

Figure 1: Costs using uniform random initialization of dataset 1 with varied k

Figure 1 indicates that the costs decrease as k increases from 2 to 7. When k increases from 2 to 4, the costs dramatically fall from about 6200 to 3600, however starting from k of 4, the line trend transformed to a gradual decrease.

I decide to use 4 clusters for dataset 1 since that is where a bend in the plot happens, which is an indicator of an appropriate number of clusters.

1.1.2 Dataset2 Graph and Analysis

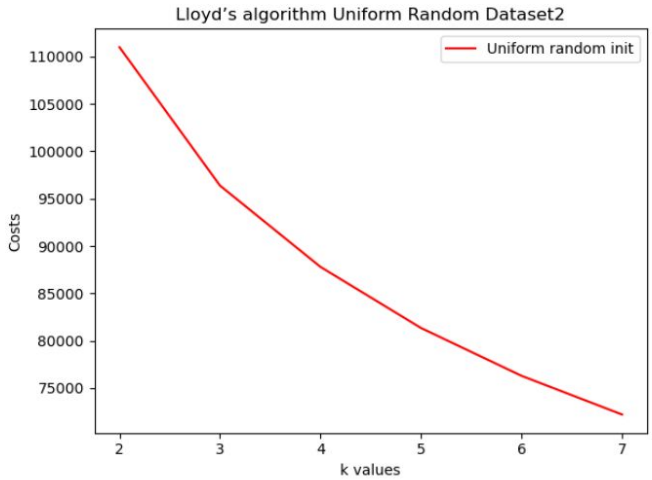

Figure 2: Costs using uniform random initialization
of dataset 2 with varied k

As shown in Figure 2, the trend is similar to the trend in Figure 1 which is the costs decreases as k increases. When k increases from 2 to 4, the costs decrease fastly, however starting from k values 4, the costs decrease slowly to about 86000.

Therefore I decide to use 4 clusters for dataset 2, also the reason is that the bend in the plot happens, from that point, the decreasing trend changed to a more flat falling shape.

1.2 K-means++ Initialization

1.2.1 Dataset1 Graph and Analysis

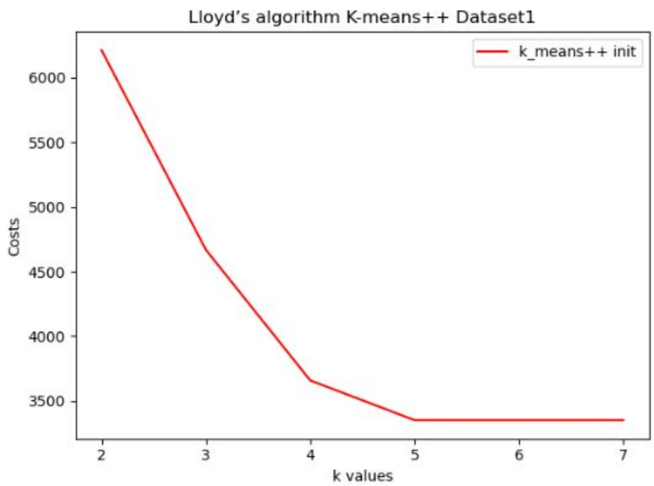

Figure 3: Costs using k means ++ initialization
of dataset 1 with varied k

The plot in Figure 3 indicates that the costs still decreases as k increases. As K increases to 4 and costs decreased to about 3600, there is a bend in the plot, and the line tendency transformed to a steep fall. So I chose 4 clusters to be the optimal choice of k.

1.2.2 Dataset2 Graph and Analysis

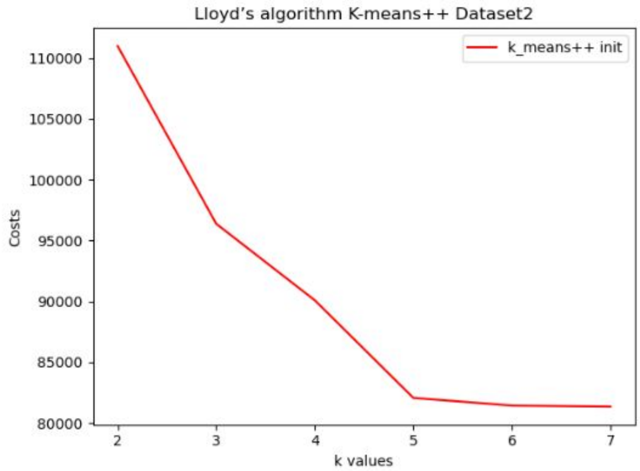

Figure 4: Costs using k means ++ initialization
of dataset 2 with varied k

The whole trend of the line in Figure 4 is similar to the line trend in Figure 3, The costs keep decreasing as k increases. Starting from the cost of 2 clusters about 111,500, the costs fastly falls to about 87,000 when k is 5. As k keeps increasing, the costs slowly decrease.

Since the bend happens when using 5 clusters, I chose 5 to be the optimal number of clusters.

2 Hierarchical Agglomerative Clustering

To get a better configuration, I chose 150 to be the p parameter for truncate_mode when creating the dendrogram.

2.1 Graph and Analysis

2.1.1 Dataset 1

Single Linkage:

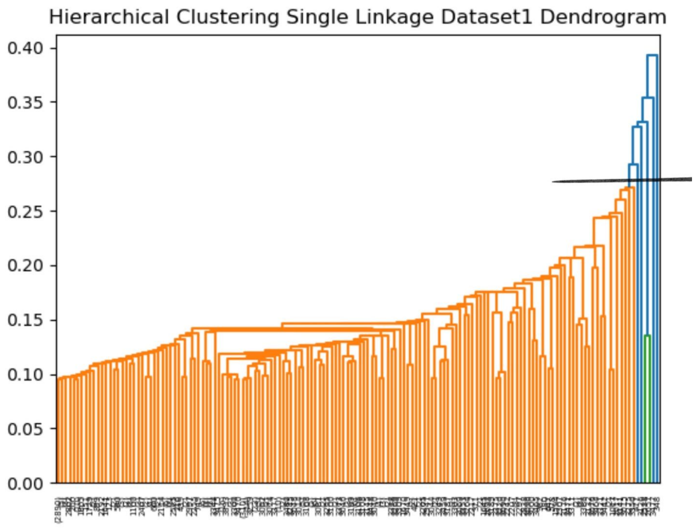

Figure 5: Dendrogram of Hierarchical agglomerative clustering in single linkage of Dataset 1

The black line in Figure 5 is the cut that I chose, there are six clusters as cutting the black line. The reason I chose this cut is the further clustering distance is too long which is unreasonable.

Average Linkage:

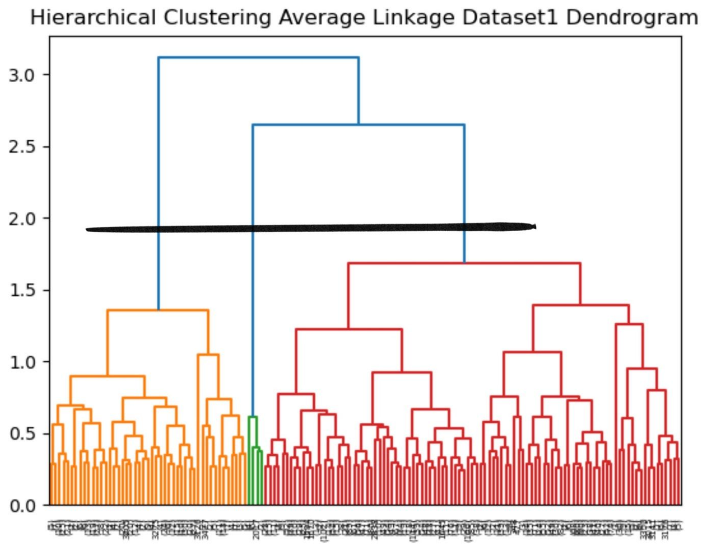

Figure 6: Dendrogram of Hierarchical agglomerative clustering in average linkage of Dataset 1

The black line in Figure 6 is the cut that I chose, there are three clusters as cutting the black line. The reason I chose this cut is the distance of the last two clusterings is large enough to not consider that. So I cut when there are three clusters.

2.1.2 Dataset 2

Single Linkage:

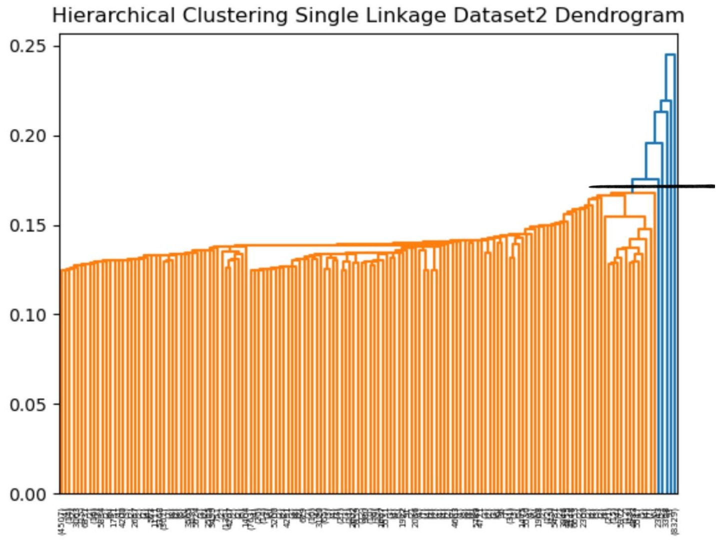

Figure 7: Dendrogram of Hierarchical agglomerative clustering in single linkage of Dataset 2

The black line in Figure 7 is the cut that I chose, there are six clusters as cutting the black line. The reason I chose this cut is comparing to previous clustering, the last four clustering takes too much distance which is not ideal enough, so I chose the cut before that happens.

Average Linkage:

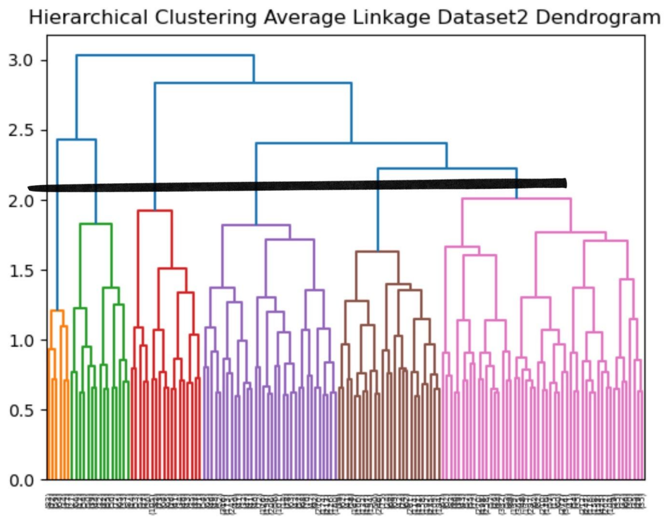

Figure 8: Dendrogram of Hierarchical agglomerative clustering
in average linkage of Dataset 2

The black line in Figure 8 is the cut that I chose, there are six clusters as cutting the black line. I cut the dendrogram before the clustering distance is too large so that the clustering would be more reasonable.