CS 434 Implementation Assignment4 Report

Wenbo Hou Zhi Jiang

May 21, 2017

Introduction

In this assignment, we are going to explore two clustering algorithms: K-means Algorithm and Hierarchical agglomerative clustering. For K-means Algorithm, we tested two properties: how SSE changes during the training process and how SSE changes with different values of K(different number of clusters). For Hierarchical Agglomerative Clustering, we tested effects of two distance measurements: Single Link and Complete Link, and analyzed the most suitable cluster number for the given data.

We used MATLAB as the data analysis tool and generated several scripts for this assignment. There are two scripts for the K-means algorithm. `Kmean_1.m` explored who SSE changes during the learning process with K = 2; `Kmean_2.m` explored how SSE changes with different K values (from 2 to 10). `HAC_single.m` used single link to generate dendrogram, and `HAC_complete.m` used complete link to generate dendrogram.

Part 1. K-means Algorithm

Generally, the K-means algorithm will classify the given data into K clusters. It starts with K random cluster centers chosen from the given data set. Then, it will keep classifying data into those clusters and updating cluster centers until the SSE get converged. The first problem of this part requires us to plot how SSE changes during the learning process when K = 2. Here is the plot:

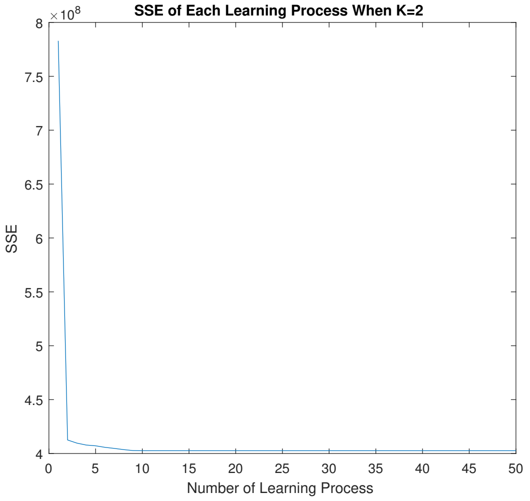

Figure 1: SSE changes until converged

From above plot, we can see that the SSE gets converged very fast, within 5 learning iterations. However, the SSE we get for K = 2 is very large, which means K = 2 is not conclusive.

The second problem now requires us to explore how SSE changes versus different K values. As the problem statements suggested, we choose k value from 2 to 11, and iterate the learning 10 times for each k to find the smallest SSE (Best Model).

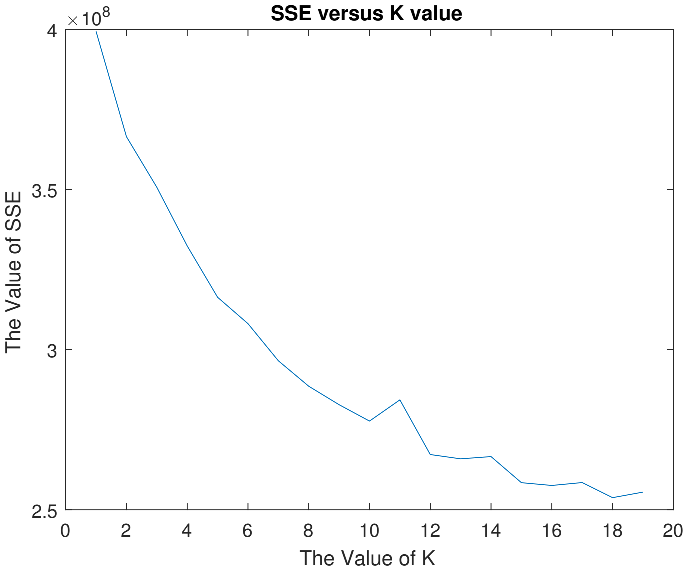

Figure 2: SSE changes with different K values

From above plot, we can see that SSE keeps decreasing when K gets greater. As we discussed in the lecture, K-means algorithm is a flat clustering method. A larger K value will provide a nearer cluster for each sample than a smaller K value does. Consequently, SSE and the value of K has a negative relationship mathematically. From the plot, we can see that the SSE does not decrease largely, when K is greater than 10. So, K = 10 is a critical model for the K-means algorithm. Although K-means algorithm always prefer large K value, there is no need to put extra efforts on larger K values when their benefits do not deserve efforts on them. Consequently, we choose 10 as the best K value for this given data set.

Part 2. Hierarchical agglomerative clustering (HAC)

Single Link

First of all, we implemented the HAC algorithm with single link. The single link represents the distance of two closest members of clusters, thus the single link is famous for its chining effects. In figure 3, it is clear that the clusters are spread out and not compact enough. We found the there are small distances between first five clusters and there is an obvious "chaining" effect, thus we cut the dendrogram at hight = 2060 and eventually we got 6 clusters because the first five clusters are one cluster.

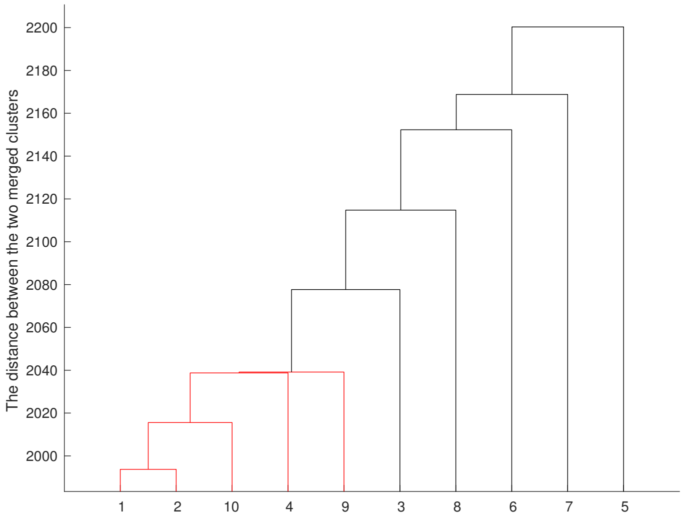

Figure 3: Dendrogram with Single Link

Complete Link

On the other hand, we still used HAC algorithm with complete link to generate dendrogram. The complete link represents the distance of two furthest members of clusters. The complete link can avoid chaining but suffers from crowding. In the other words, clusters can be compact, but not far enough apart. Moreover, outliers in complete-link clustering still make negative influence on solution. We chose to cut dendrogram where the gap between two successive combination distance is largest, and eventually we got 3 clusters.

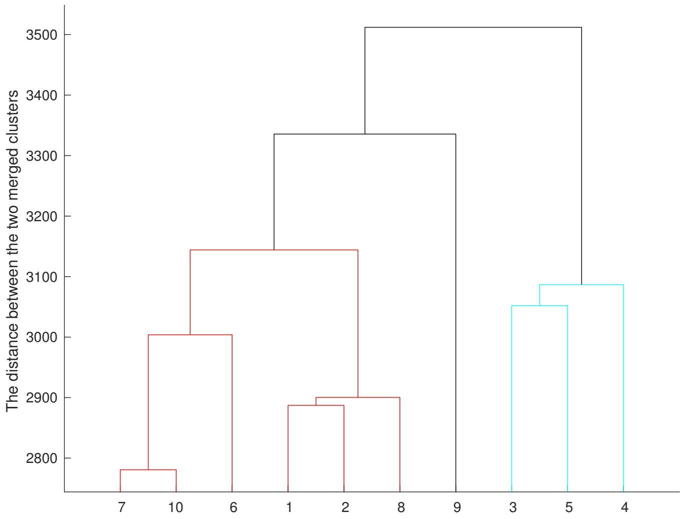

Figure 4: Dendrogram with Complete Link