SENG474 Assignment 3

Jonathan Skinnider V00207396

March 23 2020

1 Lloyd's Algorithm

We implemented Lloyds algorithm and two initialization algorithms (fully random and k-means++) and ran both models on dataset 1 and dataset 2. We varied the value of k, that is the number of clusters used, and ran each algorithm 3 times for each value of k ranging from 2 to 20. We plotted the lowest cost value found amongst these three trials. Our results are as follows:

On dataset 1, both algorithms found their minimum cost when using 20 clusters. The trends of both graphs are very similar, however the k-means++ algorithm has a smoother curve. This is likely because it is a much more consistent algorithm, since it's initial clusters are already intuitively 'pretty good'. That means that after it converges, it is more likely to be to closer to optimal. Additionally, due to the structure imposed by the initial clusters, each repeated attempt is more likely to be roughly similar. Perhaps if we had run the random initial cluster's algorithm 5, 10 or 20 times and picked the lowest cost we would have a similarly smooth graph.

Lloyds algorithm with random initialization and k-means++ initialization found the lowest cost to be 1648 and 1403 respectively when using 20 clusters.

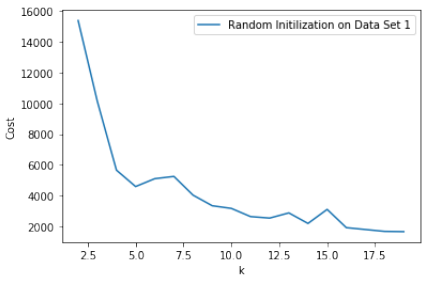

Figure 1: Random Initialization on Dataset 1

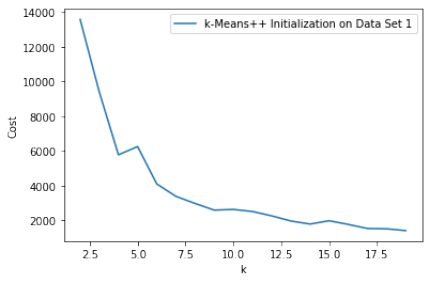

Figure 2: k-means++ Initialization on Datset 1

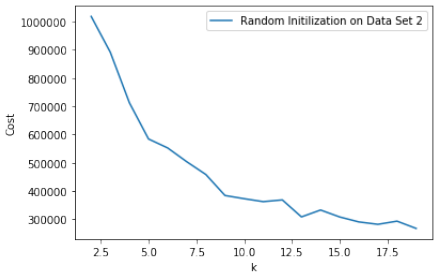

Figure 3: Random Initialization on Dataset 2

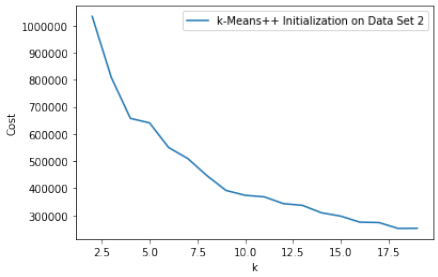

Figure 4: k-means++ Initialization on Datset 2

On the much larger dataset 2, the algorithm took much longer, regardless of initialization method. We similarly ran each algorithm three times and achieved the following results.

Again both plots show a rough decrease in the cost as the number of clusters increase. Again both methods found the lowest cost clustering using k=20. The lowest cost of each algorithm is 267649 for randomized initialization and 252287 for k-means++ initialization.

Both datasets show a significant decrease in cost when using k-means++ initialization over fully random initialization. On dataset 1, the cost was reduced by 14.8%, and on dataset 2 the cost was reduced by 5.8%. This illustrates the disadvantage when using naive random initialization.

2 Hierarchical Agglomerative Clustering

On the two datasets we performed hierarchical agglomerative clustering using both single linkage and average linkage to determine when to merge clusters. The denodrograms from the results are plotted bellow. The general resutls seems to be that average linkage works much better for both data sets.

First lets look at the single linkage. Both denodrograms show a consistent small step size when going from cluster to cluster. Additionally, for the most part almost every initial cluster is a proper subset of every cluster formed with points to the left of it. Since the distance between points on a dendrogram tell you how dissimilar the clusters are, and since the points for both single linkage dendrograms are all very, very close, this implies that single-linkage hierarchical agglommerative clustering is not effective on either dataset.

Average linking clustering, however, performed remarkably well on both data sets. For the first data set, it obviously found 2 main clusters (Figure 7). This is apparent since there is a massive drop from the top point to the second two

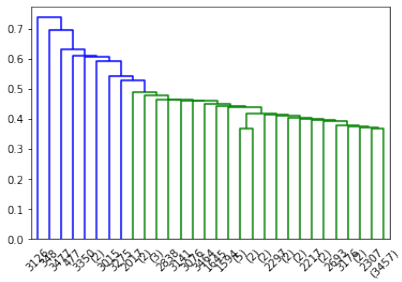

Figure 5: Single Linkage on Dataset 1

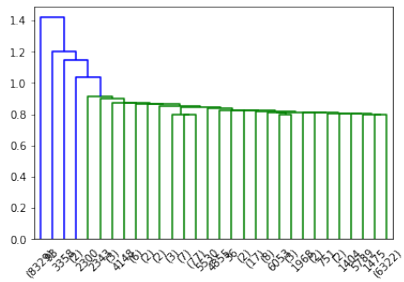

Figure 6: Single Linkage on Dataset 2

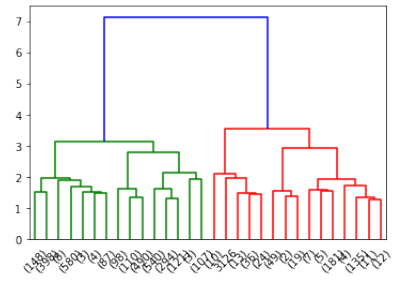

Figure 7: Average Linkage on Dataset 1

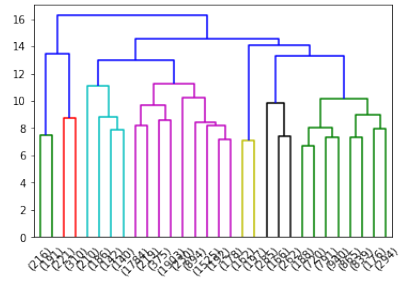

Figure 8: Average Linkage on Dataset 2

points of the dendrogram. This means that there is a large difference between the two clusters. Hence it makes the most intuitive sense that dataset 1 should be partitioned into two clusters. Given the fact that the data has been derived from a gaussian mixture model, this makes sense since it was created by two normal distribution, with two means. Hence clusters centered around these means should be optimal.

What is odd, however, is the fact that this is not related to what we found when using Lloyds algorithm. Lloyds informed us that using more clustered decreased the cost, which is in contradiction to what the dendrogram is saying: that two clusters effectively classify the data.

When analyzing the dendorgram for dataset 2, however, it is less clear. When looking at the left most points, one can see there is a significant drop (from 13 to 7 or 8), however there is a much less significant drop in the middle section (from 13 to 11). It would appear that 7 clusters classify the data well (represented by the colors shown).

Although the three clusters (colors) on the right once combined are quite similar, on their own they are quite dissimilar (as shown by the jump from 10 to 13 with the brown and green color). In addition the significant and consistent

jumps for the clusters on the left steer us towards deciding to use more clusters. Again, hierarchical clustering yielded different results than Lloyds algorithm. Although the difference is less pronounced on dataset 2 than dataset 1, it is still significant - Lloyds recommended 20 clusters, whereas here it appears that 7 works very well.