SENG 474: Data Mining

Asignment 3

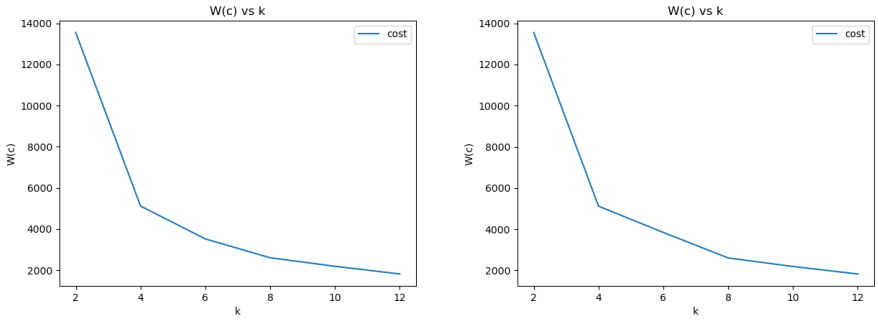
Diego Aquino Chavez - V00892482

March 24, 2020

1

1 Lloyd's algorithm

Both datasets were tested on implementations of *kmeans* with uniform random initial clusters and `kmeans++`. The cost of the clusterings against the number of clusters is represented in Figure 1 for the first dataset.



(a) Uniform random (b) kmeans++

Figure 1: Costs vs number of clusters for dataset 1

For the best number of clusters k, from the graph it was decided for $k = 8$ as the cost starts to reduce slowly after this number of clusters for both methods. To illustrate this decision, the data was clustered for $k = 8$

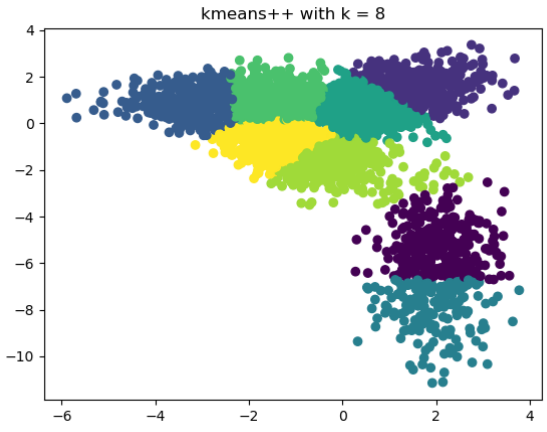


Figure 2: Clustering with $k = 8$ using `kmeans++`

For dataset 2, Figure 3 shows the graphs of costs vs number of clusters for this dataset.

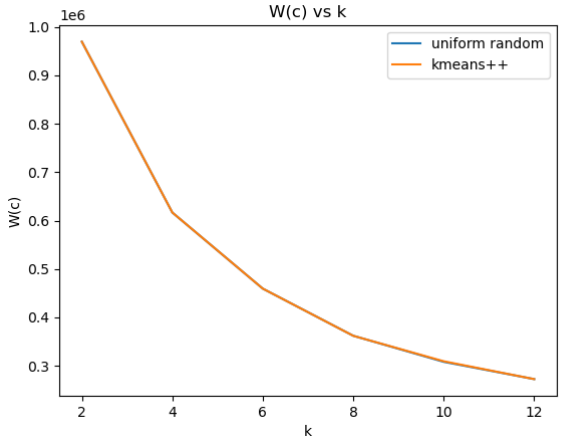
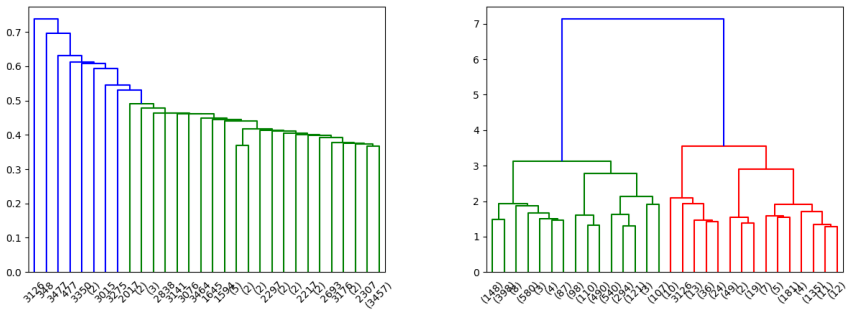


Figure 3: Costs vs number of clusters for dataset 2

For this dataset, following from the graph. An appropiate value of k would be $k = 8$ as the cost value decreases slowly from this point.

2 Hierarchical Agglomerative Clustering

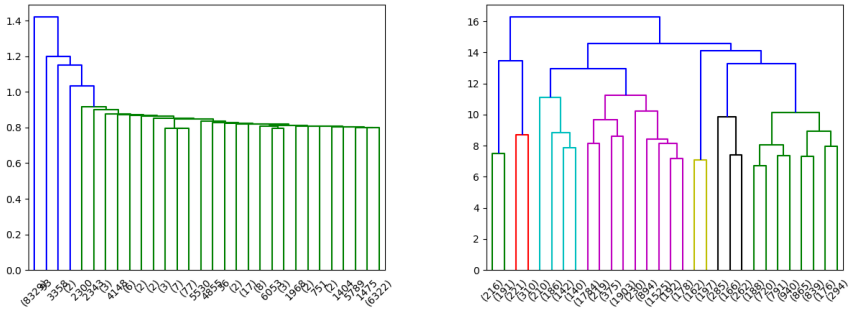
Figure 4 presents the dendograms for dataset 1:



(a) Single Linkeage (b) Average linkeage

Figure 4: Dendograms for dataset 1

Figure 5 presents the dendograms for dataset 2:

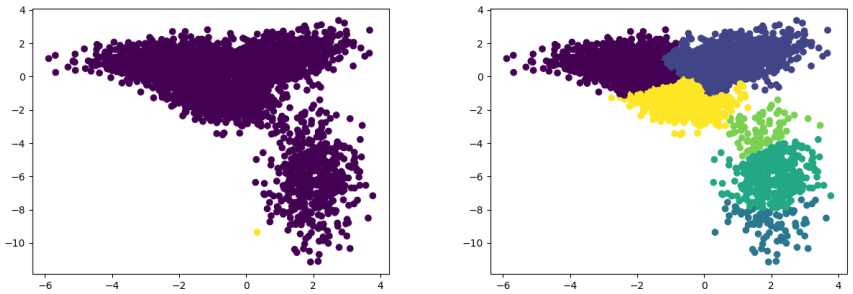


(a) Single Linkeage (b) Average linkeage

Figure 5: Dendograms for dataset 2

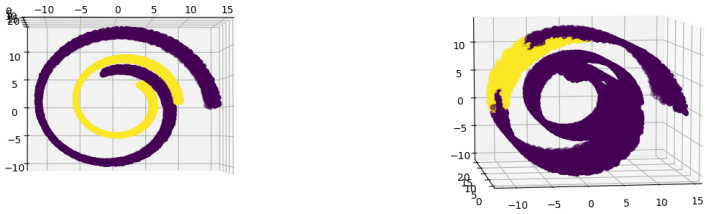
From the dendograms, for dataset 1 k was selected to be 2 for Single Linkeage and 6 for Average Linkeage. For dataset 2 k is 2 for Single Linkeage and 2 for Average Linkeage. This values of k were selected due that it was the number of horizontal lines that could fit between the height differences.

The folowing were the results for the models implemented using sckit-learn AgglomerativeClustering



(a) Single Linkeage (b) Average linkeage

Figure 6: Scatter plots for dataset 1



(a) Single Linkeage (b) Average linkeage

Figure 7: Scatter plots for dataset 2

3 Discussion

For Lloyd's algorithm, both initilizations obtained about the same cost vs k graphs. It was not expected that `kmeans++` were to obtained higher costs at incrementing the number of clusters compared to uniform random. But, this could be partly explained by the shape of the data. It is important to mention that this algorithm performed poorly for the second dataset and made clusters that could not be correctly interpreted.

For Hierarchical Agglomerative Clustering, the decision for k was not simple for the second dataset due that the dendograms were not as clear as the ones from dataset 1. Single Linkeage performed poorly for the first dataset but was highly effective for the second dataset. On the other hand, Average Linkeage performed poorly on the second dataset but as effective as kmeans on the first dataset.