CENG499 Assignment2

Uygar Yaşar 2310613

December 2022

1 Part1

```
Confidence interval 0. configuration: 95.333 +- 0.370
Confidence interval 1. configuration: 93.867 +- 0.234
Confidence interval 2. configuration: 95.467 +- 0.234
Confidence interval 3. configuration: 93.600 +- 0.701
Confidence interval 4. configuration: 95.067 +- 0.596
Confidence interval 5. configuration: 94.533 +- 0.437
Confidence interval 6. configuration: 92.800 +- 0.437
Confidence interval 7. configuration: 92.400 +- 0.596
Confidence interval 8. configuration: 88.267 +- 1.259
Confidence interval 9. configuration: 85.600 +- 1.203
Confidence interval 10. configuration: 81.867 +- 1.192
Confidence interval 11. configuration: 90.000 +- 0.739
Configuration 2 has best mean accuracy & (if equal) standard deviation combination with 95.467 +- 0.267
```

Figure 1: KNN Confidence Intervals

```
configurations = [(3, Distance.calculateCosineDistance), (5, Distance.calculateCosineDistance),
                  (10, Distance.calculateCosineDistance), (20, Distance.calculateCosineDistance),
                  (5, Distance.calculateMinkowskiDistance), (20, Distance.calculateMinkowskiDistance),
                  (30, Distance.calculateMinkowskiDistance), (35, Distance.calculateMinkowskiDistance),
                  (5, Distance.calculateMahalanobisDistance), (15, Distance.calculateMahalanobisDistance),
                  (25, Distance.calculateMahalanobisDistance), (3, Distance.calculateMahalanobisDistance)]
```

Figure 2: Corresponding configs

Now, as it can be seen, configuration 0, 2, 4 gave better performance. So, the best fitting distance metric is cosine distance. Worst fitting is Mahalanobis. Also choosing k, 3 or 10 gives better performance using cosine distance. The best configuration is 2 since it gives better mean accuracy with smaller variance.

2 Part2

2.1 K-means

```
(dataset1) Confidence interval k=2: 321.318 +- 4.103
(dataset1) Confidence interval k=3: 169.619 +- 21.784
(dataset1) Confidence interval k=4: 79.702 +- 17.133
(dataset1) Confidence interval k=5: 41.212 +- 11.714
(dataset1) Confidence interval k=6: 23.432 +- 4.289
(dataset1) Confidence interval k=7: 19.850 +- 0.625
(dataset1) Confidence interval k=8: 20.174 +- 3.939
(dataset1) Confidence interval k=9: 16.801 +- 0.166
(dataset1) Confidence interval k=10: 15.802 +- 0.115
(dataset2) Confidence interval k=2: 261.384 +- 0.000
(dataset2) Confidence interval k=3: 112.969 +- 13.596
(dataset2) Confidence interval k=4: 63.006 +- 14.872
(dataset2) Confidence interval k=5: 46.713 +- 1.610
(dataset2) Confidence interval k=6: 44.401 +- 0.192
(dataset2) Confidence interval k=7: 43.049 +- 0.188
(dataset2) Confidence interval k=8: 41.878 +- 0.183
(dataset2) Confidence interval k=9: 40.637 +- 0.191
(dataset2) Confidence interval k=10: 39.261 +- 0.222
```

Figure 3: k values intervals for both

By looking at the confidence intervals and elbow plots. It seems that the most appropriate k values are 4 and 6 for dataset1, 4 and 5 for dataset2. I would prefer 6 for dataset1 and 4 for dataset 2. The complexity is O(d * K * N * I) for k-means algorithm. (It can change if any dimensionality reduction method has been used.) I would expect lower coss in k-means++ algorithm but I could not implement it.

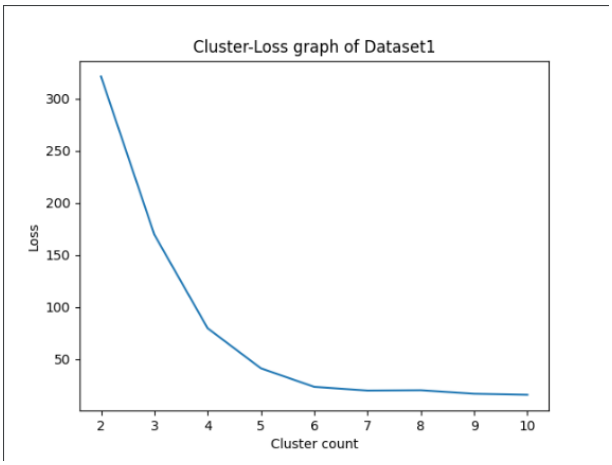

Figure 4: Elbow plot for dataset1

Figure 5: Elbow plot for dataset2

3 Part3

HAC worst case complexity is O(n^3) which is worse than k-means. So, if the dataset is large, I would prefer k-means.
From documentation of scikit learn,
1- The thickness of plot gives information about cluster size.
2- The values close to 1 indicate that the sample is far away from the neighboring clusters.
3- A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters
4- negative values indicate that those samples might have been assigned to the wrong cluster. (I will give references them with numbers)

3.1 Single Euclidean Dendrogram and silhouettes

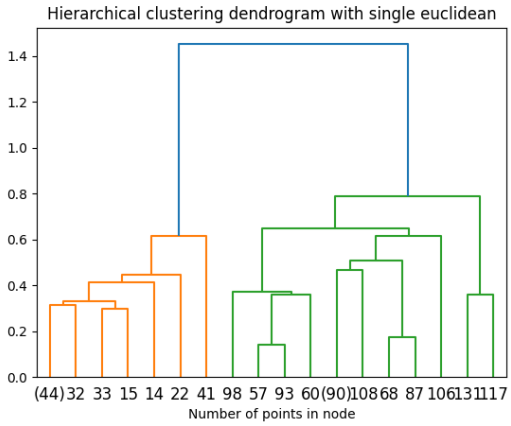

Figure 6: Single euclidean dendrogram

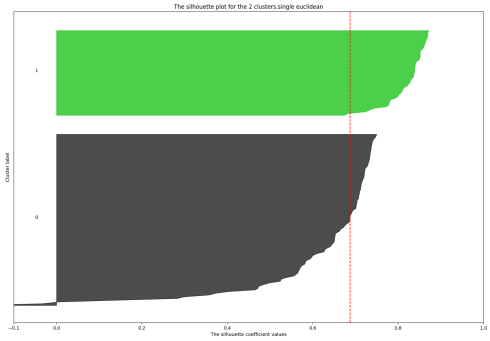

Figure 7: Silhouette graph for k=2

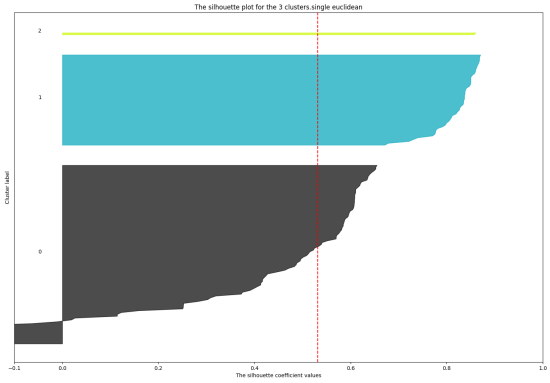

Figure 8: Silhouette graph for k=3

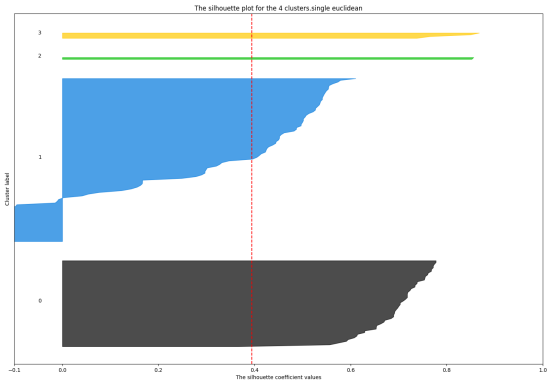

Figure 9: Silhouette graph for k=4

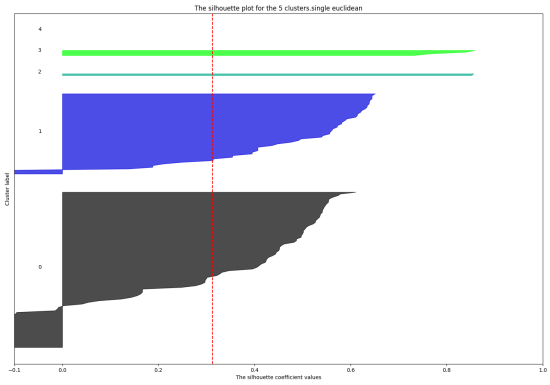

Figure 10: Silhouette graph for k=5

The most appropriate choose of k is 2 for single linkage euclidean distance. For the reasons (1), (2), in k=3, k=4 and k=5 some of the cluster sizes are too small also there are some negative values which may indicate misassigns.

3.2 Complete Euclidean Dendrogram and silhouettes

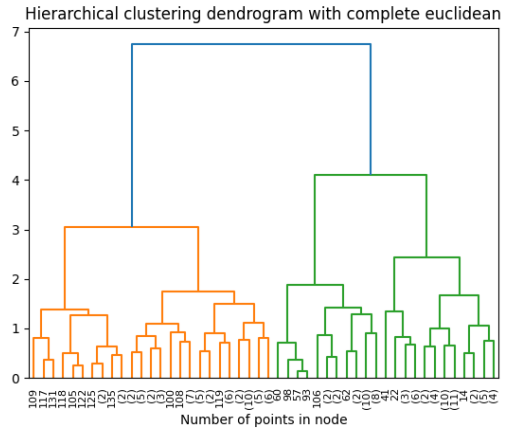

Figure 11: Complete euclidean dendrogram

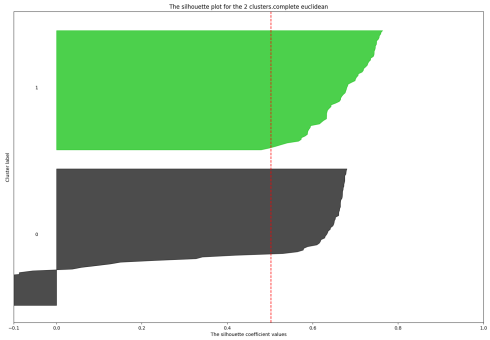

Figure 12: Silhouette graph for k=2

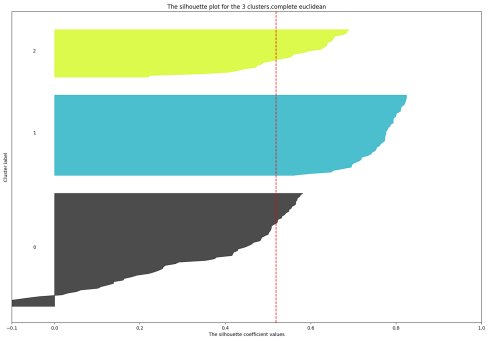

Figure 13: Silhouette graph for k=3

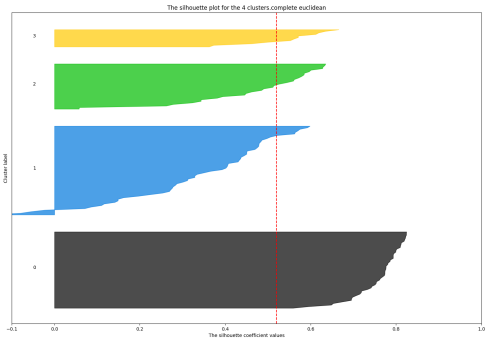

Figure 14: Silhouette graph for k=4

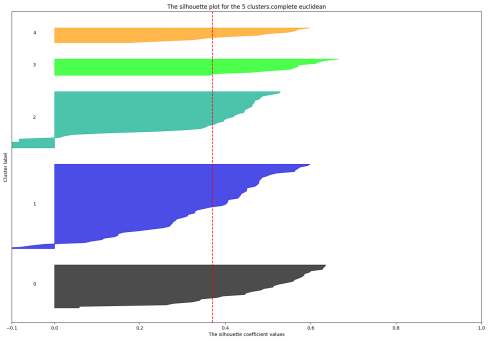

Figure 15: Silhouette graph for k=5

The most appropriate choose of k is 3 and 4 for complete linkage euclidean distance. In k=2 misassignment possibility is too high since it goes through -1. and k=5 some of the cluster sizes are too small and more negative values than 3 and 4.

3.3 Single Cosine Dendrogram and silhouettes

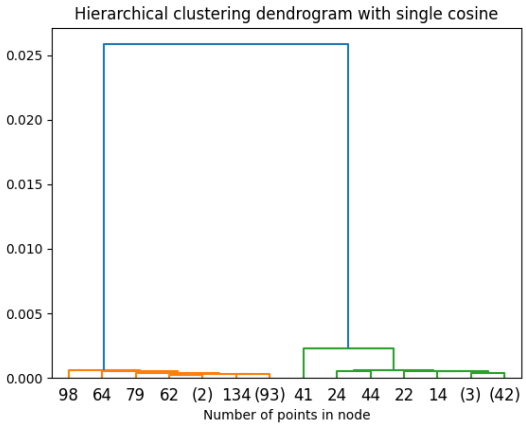

Figure 16: Single cosine dendrogram

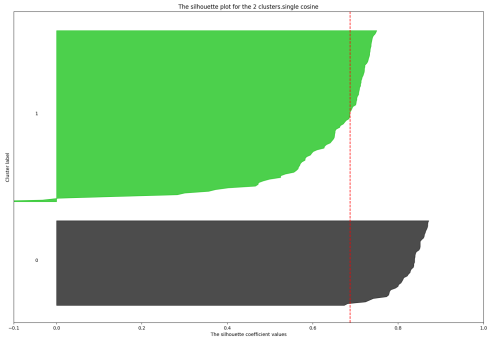

Figure 17: Silhouette graph for k=2

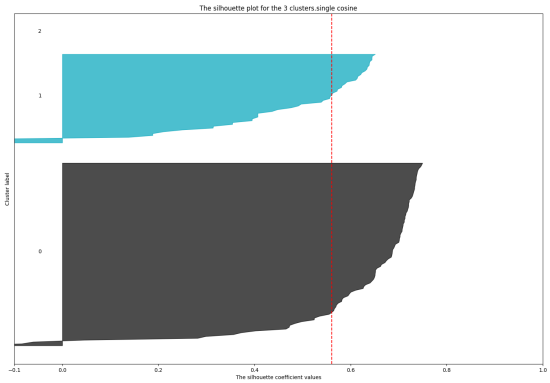

Figure 18: Silhouette graph for k=3

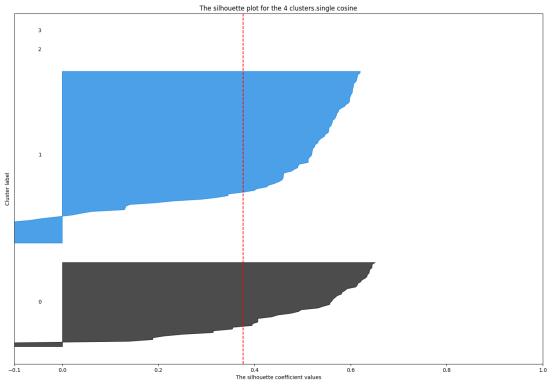

Figure 19: Silhouette graph for k=4

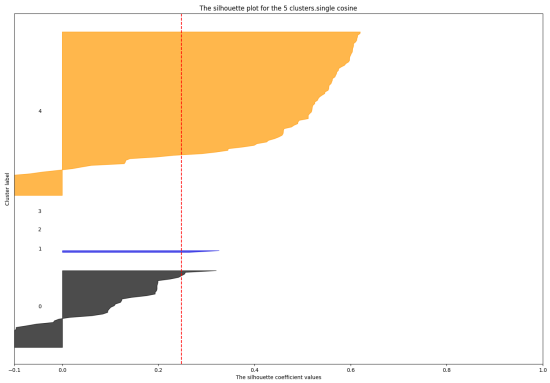

Figure 20: Silhouette graph for k=5

The best k value is 2 even it has some negative values. In k=3, there are more negative values and there is an empty cluster. In k=4 and k=5 there are more empty clusters and wrongly clustered point count increases. This configuration (single-cosine) seems the worst one.

3.4 Complete Cosine Dendrogram and silhouettes

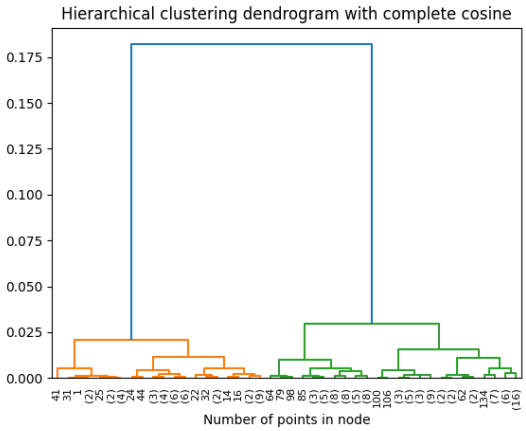

Figure 21: Single euclidean dendrogram

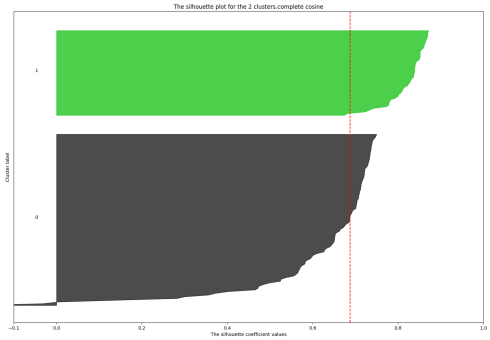

Figure 22: Silhouette graph for k=2

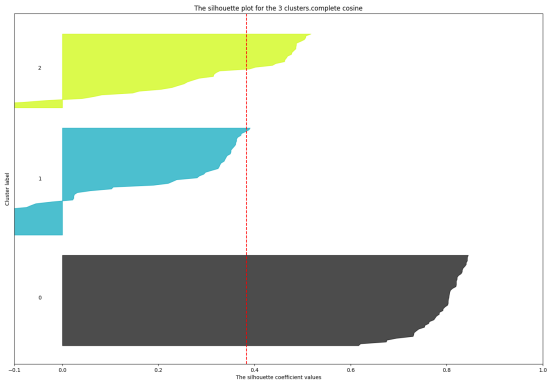

Figure 23: Silhouette graph for k=3

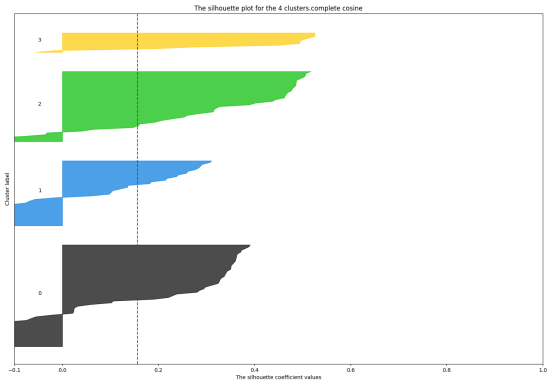

Figure 24: Silhouette graph for k=4

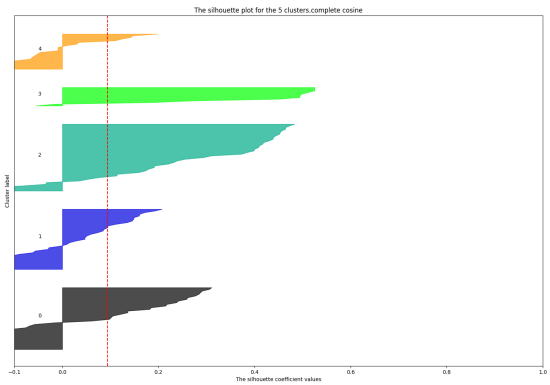

Figure 25: Silhouette graph for k=5

The best k value is 2 even it has some negative values (just a few). In k=3, there are more negative values. In k=4 and k=5 wrongly clustered point count increases. This configuration is slightly better than single cosine but still not appropriate for this dataset.