Introduction to Machine Learning Homework 2 Report

Nuray Akar 2380004

November 2022

1 Part 1

1.1 Cosine

Configuration: k=1, cosine
Average Accuracy %91.46666666666667
Confidence Interval [0.8981219755141148,0.9312113578192187]

Configuration: k=5, cosine
Average Accuracy %93.73333333333333
Confidence Interval [0.9219124521364338,0.9527542145302329]

Configuration: k=10, cosine
Average Accuracy %95.19999999999999
Confidence Interval [0.9376671114487614,0.9663328885512383]

Configuration: k=15, cosine
Average Accuracy %93.99999999999999
Confidence Interval [0.9233688485064923,0.9566311514935074]

Configuration: k=19, cosine
Average Accuracy %93.6
Confidence Interval [0.9212352448490784,0.9507647551509215]

1.2 Minkovski

Configuration: k=1, Minkovski
Average Accuracy %92.93333333333332
Confidence Interval [0.9139124521364337,0.9447542145302328]

Configuration: k=5, Minkovski
Average Accuracy %94.66666666666666
Confidence Interval [0.9318834208946598,0.9614499124386733]

Configuration: k=10, Minkovski
Average Accuracy %95.06666666666666
Confidence Interval [0.9330155393120503,0.968317794021283]

Configuration: k=15, Minkovski
Average Accuracy %95.6
Confidence Interval [0.9404907968827108,0.9715092031172892]

Configuration: k=19, Minkovski
Average Accuracy %95.06666666666666
Confidence Interval [0.9385153691959345,0.9628179641373988]

1.3 Mahalanobis

Configuration: k=1, Mahalanobis
Average Accuracy %86.4
Confidence Interval [0.8448102308739499,0.8831897691260503]

Configuration: k=5, Mahalanobis
Average Accuracy %88.8
Confidence Interval [0.8676093002463269,0.9083906997536731]

Configuration: k=10, Mahalanobis
Average Accuracy %87.6
Confidence Interval [0.8523323744044035,0.8996676255955963]

Configuration: k=15, Mahalanobis
Average Accuracy %84.4
Confidence Interval [0.8207398652148742,0.867260134785126]

Configuration: k=19, Mahalanobis
Average Accuracy %83.86666666666667
Confidence Interval [0.8119927380554638,0.8653405952778698]

1.4 How did I have picked the best-performing hyperparameter values?

I have analyzed the average accuracies above. Thus, as one can see, the best accuracy belongs to the configuration with k=15 and Minkovski similarity.

2 Part 2

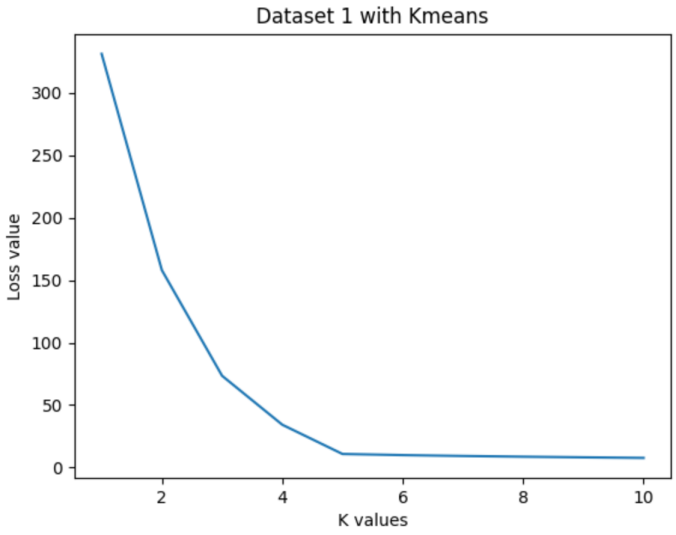
In this part, I have run the algorithm with the following k values:

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

2.1 Kmeans

2.1.1 What is the most suitable cluster number for dataset 1?

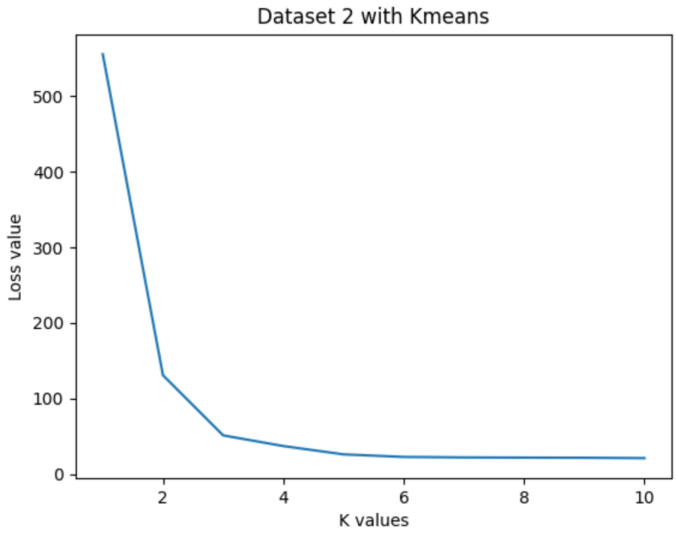
K = 5




```
Dataset 1 with Kmeans
K:1, Average Loss:331.05829266175095, Confidence Interval:[331.0582926617509,331.058292661751]
K:2, Average Loss:157.99417117169423, Confidence Interval:[157.99417117169423,157.99417117169423]
K:3, Average Loss:73.47528105274182, Confidence Interval:[73.47528105274182,73.47528105274182]
K:4, Average Loss:34.3102810112205, Confidence Interval:[34.13079826933532,34.489763753105684]
K:5, Average Loss:10.893421106088477, Confidence Interval:[10.893421106088475,10.893421106088478]
K:6, Average Loss:10.007220978970064, Confidence Interval:[9.976118386103948,10.03832357183618]
K:7, Average Loss:9.316219664202142, Confidence Interval:[9.269591802572684,9.3628475258316]
K:8, Average Loss:8.72100298102161, Confidence Interval:[8.601991975439113,8.840013986604106]
K:9, Average Loss:8.212605941395523, Confidence Interval:[8.062466168056853,8.362745714734194]
K:10, Average Loss:7.7563082705451745, Confidence Interval:[7.552617324227487,7.959999216862862]
```

2.1.2 What is the most suitable cluster number for dataset 2?

K = 3

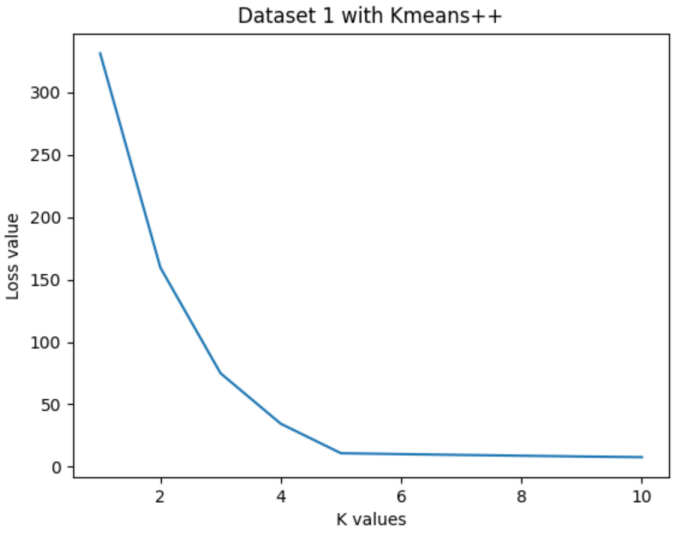



```
Dataset 2 with Kmeans
K:1, Average Loss:555.8067130382981, Confidence Interval:[555.8067130382981,555.8067130382981]
K:2, Average Loss:130.6921524807779, Confidence Interval:[130.6921524807779,130.6921524807779]
K:3, Average Loss:51.026447953780306, Confidence Interval:[51.0264479537803,51.02644795378031]
K:4, Average Loss:36.868321265542335, Confidence Interval:[28.67702651300917,45.0596160180755]
K:5, Average Loss:25.8679750603117, Confidence Interval:[20.994371500615156,30.741578620008244]
K:6, Average Loss:22.468243065625316, Confidence Interval:[22.135880777740432,22.8006053535102]
K:7, Average Loss:21.827589703407547, Confidence Interval:[21.469456411288192,22.185722995526902]
K:8, Average Loss:21.57071873713726, Confidence Interval:[21.22154925479967,21.91988821947485]
K:9, Average Loss:21.38874436439735, Confidence Interval:[21.029519803959623,21.747968924835078]
K:10, Average Loss:20.869008163514252, Confidence Interval:[20.496220309759934,21.24179601726857]
```

2.2 Kmeans++

2.2.1 What is the most suitable cluster number for dataset 1?

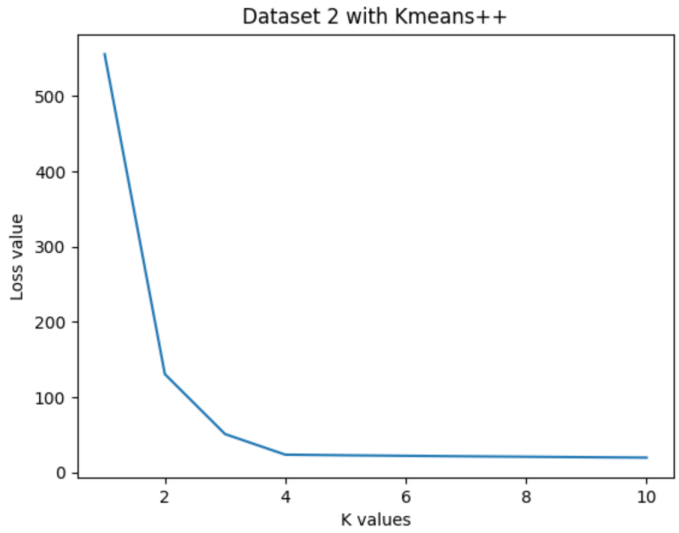
K = 5




```
Dataset 1 with Kmeans++
K:1, Average Loss:331.05829266175095, Confidence Interval:[331.0582926617509,331.058292661751]
K:2, Average Loss:159.5629632890701, Confidence Interval:[158.64959640447756,160.47633017366263]
K:3, Average Loss:74.92998559293672, Confidence Interval:[73.79667833551431,76.06329285035912]
K:4, Average Loss:34.52374424668652, Confidence Interval:[34.390996448410014,34.656492044963024]
K:5, Average Loss:10.89541914476444, Confidence Interval:[10.892942350361333,10.89789593916755]
K:6, Average Loss:10.17217335706026, Confidence Interval:[10.132462593452574,10.211884120667944]
K:7, Average Loss:9.506497090183299, Confidence Interval:[9.433224773687241,9.579769406679356]
K:8, Average Loss:8.892620193325047, Confidence Interval:[8.845363827071473,8.939876559578622]
K:9, Average Loss:8.261323118133244, Confidence Interval:[8.217471489685533,8.305174746580956]
K:10, Average Loss:7.733704313253857, Confidence Interval:[7.628781404751198,7.838627221756516]
```

2.2.2 What is the most suitable cluster number for dataset 2?

K = 3



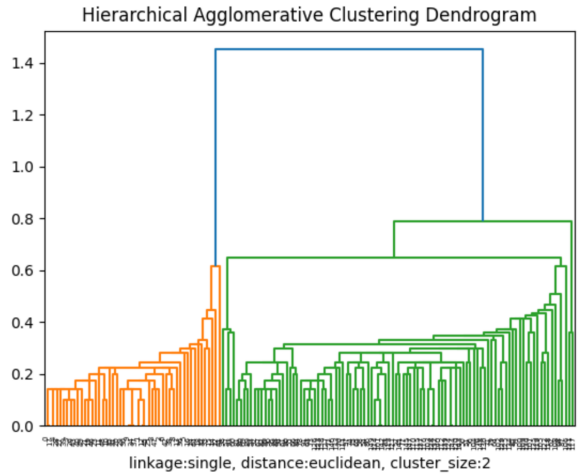

```
Dataset 2 with Kmeans++
K:1, Average Loss:555.8067130382981, Confidence Interval:[555.8067130382981,555.8067130382981]
K:2, Average Loss:130.6921524807779, Confidence Interval:[130.6921524807779,130.6921524807779]
K:3, Average Loss:51.026447953780306, Confidence Interval:[51.0264479537803,51.02644795378031]
K:4, Average Loss:23.65243120476044, Confidence Interval:[23.65243120476044,23.65243120476044]
K:5, Average Loss:22.82470651282056, Confidence Interval:[22.804731519019807,22.84468150662131]
K:6, Average Loss:22.110927736581424, Confidence Interval:[22.06801929346905,22.153836179693798]
K:7, Average Loss:21.481942383921353, Confidence Interval:[21.395606538968497,21.56827822887421]
K:8, Average Loss:20.89719167461552, Confidence Interval:[20.82993447447728,20.96444887475376]
K:9, Average Loss:20.293504499464593, Confidence Interval:[20.191436907825274,20.39557209110391]
K:10, Average Loss:19.73355873110764, Confidence Interval:[19.635067398875492,19.832050063339786]
```

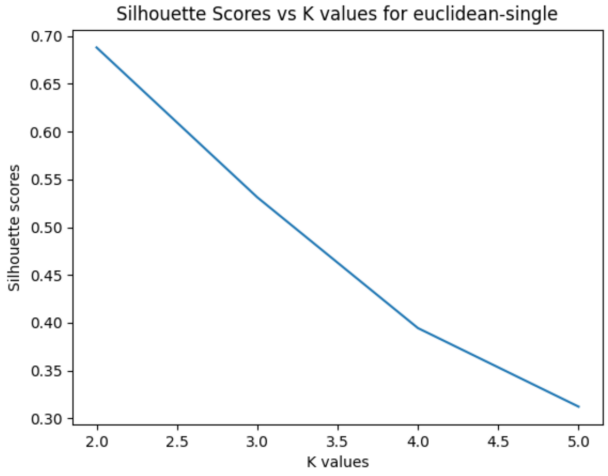
2.3 A worst-case running time analysis for Kmeans

A worst-case running time analysis for Kmeans with respect to the number of data points (N), data sample vector dimension (d), cluster number (K), and the number of iterations (I) is O(NdKI). In an iteration, we traverse every data instance(N) and for every cluster(K) we find the difference for every dimension(d) in order to find the distance which is O(NdK), after that we update the centers which is also O(NdK). We are doing this for every iteration(I). Overall it is O(NdKI).

3 Part 3

3.1 Single Linkage - Euclidean





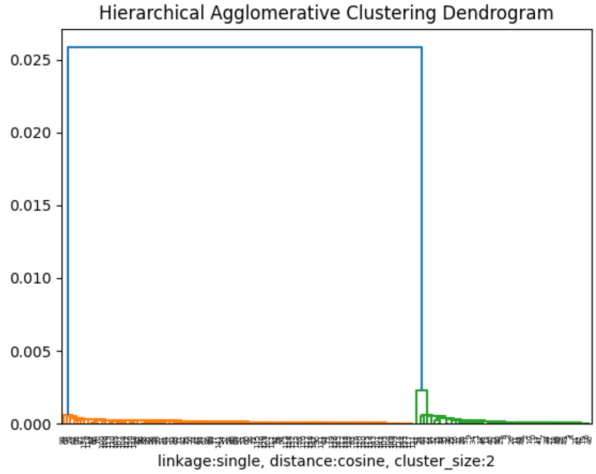
Silhouette Scores for k=[2, 3, 4, 5] respectively:

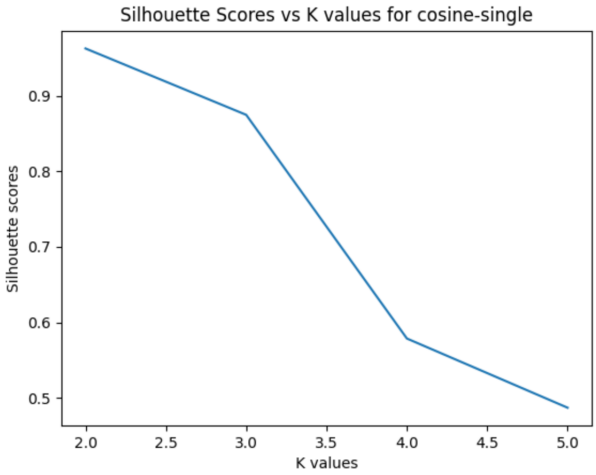
[0.68810517, 0.53133893, 0.39444217, 0.31220323]

3.1.1 Best Configuration

As one can see the silhouette scores above, k=2 is the best K value, since it has the highest score which is 0.68810517.

3.2 Single Linkage - Cosine





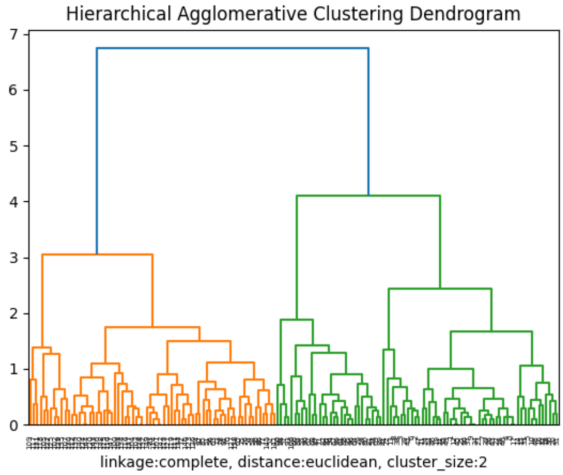
Silhouette Scores for k=[2, 3, 4, 5] respectively:

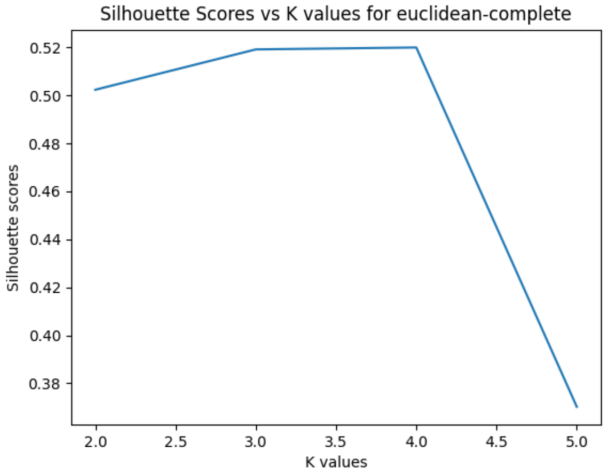
[0.96254456, 0.87470824, 0.5786437, 0.4870524]

3.2.1 Best Configuration

As one can see the silhouette scores above, k=2 is the best K value, since it has the highest score which is 0.96254456.

3.3 Complete Linkage - Euclidean





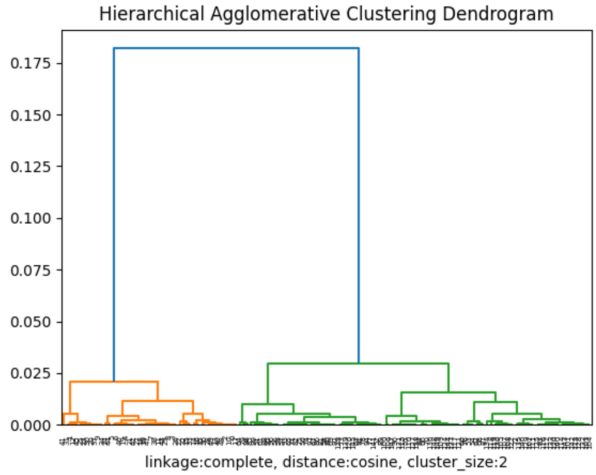
Silhouette Scores for k=[2, 3, 4, 5] respectively:

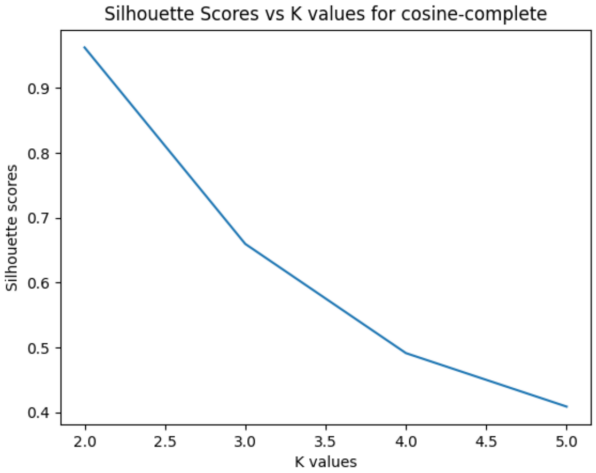
[0.5023174, 0.5191346, 0.51993114, 0.3702072]

3.3.1 Best Configuration

As one can see the silhouette scores above, k=4 is the best K value, since it has the highest score which is 0.51993114.

3.4 Complete Linkage - Cosine





Silhouette Scores for k=[2, 3, 4, 5] respectively:

[0.96254456, 0.65968406, 0.49126318, 0.40880266]

3.4.1 Best Configuration

As one can see the silhouette scores above, k=2 is the best K value, since it has the highest score which is 0.96254456.

3.5 Among 4 best configurations the one that attains the highest average silhouette score

I have indicated the best K values and their silhouette scores. Among these 4 value, the highest score is 0.96254456. Both Complete Linkage - Cosine and Single Linkage - Cosine has this score with k=2.

3.6 A worst-case run time analysis for HAC

At the beginning we found distance between every data instance which is O(N^2). After that, until there is one cluster, in other words from n to 1, we are finding distances, thus it is O(N^3). Finally, finding difference for distance for d dimension is O(d). Overall, it is O(N^3d).

4 Which clustering method (Kmeans or HAC) you would prefer to use with a dataset consisting of 1 million data points each of which has a dimension of 120000?

Since the dimension is too high, using HAC would be more logical. Because, as the number of dimensions increases, a distance-based similarity measure converges to a constant value between any given examples. However, as I found, "Hierarchical clustering is extensively used to organize high dimensional objects such as documents and images into a structure which can then be used in a multitude of ways.".