Introduction to Machine Learning
Assignment 5
Hierarchical Clustering and Linkage Measures

Group 05
Bora Yilmaz (s3903125) & Felix Zailskas (s3918270)

October 20, 2021

Contents

Hierarchical Clustering and linkage measures

1. Introduction

Given a (considerably small) data set, a scatter plot to view its data points and how they spread or cluster is possible. It will be easy to identify clusters if there are any, simply by looking at the plot of the data points. For example, closely packed data points will make themselves clear, distant and empty spaces between possible clusters will be clear to the eye. However, as the data sets grow larger and dimensions of data increase, this is not a plausible way to both identify and validate clusters. At this point, clustering algorithms, cluster linkage measures and cluster validation techniques come in to help. They allow the clustering to be done systematically, validated accurately, with all sizes of data.

In this report, the agglomerative hierarchical clustering algorithm is implemented and its result are discussed. This particular algorithm is hierarchical, meaning that it results in a hierarchical order of the clustering steps, and since it is agglomerative the clustering goes from small clusters (singletons) to the largest cluster (all points). Multiple linkage methods will be used, to analyze the different results they all give, for different number of clusters. Single, average, complete and ward linkage is considered in this report and their methodical background is referred to in section 2. After a clustering process is done, silhouette scores are calculated, which is a validation method to ensure that the formed final clusters are plausible.

2. Method

To implement the hierarchical clustering algorithm on our data set, we have to first apply the appropriate linkage method to the data.

A linkage method analyzes the distance between all data points and formed clusters in the data set and arranges new clusters based on the distances observed. This means that the clustering starts out with all data points and no formed clusters hence, each data point is treated as a cluster with one data point in it. It then joins the two closest cluster together into a single cluster. Note that evaluation of the closest clusters can vary across different linkage methods. After combining the clusters, they will be treated as a single cluster. Now this process will be repeated, where data points that are part of a cluster are evaluated together within that cluster.

In our analysis we investigate four linkage methods to generate clusters. These methods are:

1. Single Linkage:
   In this linkage method the distance between two clusters is determined by the shortest distance between any two data points within those clusters. This can be expressed using Equation 1.

   $$D(C_1, C_2) = \min_{x_i \in C_1, x_j \in C_2} D(x_i, x_j) \tag{1}$$

2. Complete Linkage:
   In this linkage method the distance between two clusters is determined by the largest distance between any two data points within those clusters. This can be expressed using Equation 2.

   $$D(C_1, C_2) = \max_{x_i \in C_1, x_j \in C_2} D(x_i, x_j) \tag{2}$$

3. Average Linkage:
   In this linkage method the distance between two clusters is determined by the average distance between all combination of data points from the two clusters. This can be expressed using

Equation 3.

$$D(C_1, C_2) = \frac{1}{|C_1|} \frac{1}{|C_2|} \sum_{x_i \in C_1} \sum_{x_j \in C_2} D(x_i, x_j) \tag{3}$$

4. Ward Linkage:
In this linkage method the distance between two clusters is determined by the increase of within cluster variance of all clusters in the system. This can be expressed using Equation 4.

$$\Delta D(C_1 \cup C_2) = \sum_{x_i \in C_1 \cup C_2} D(x_i, m_{C_1 \cup C_2})^2 - \sum_{x_i \in C_1} D(x_i, m_{C_1})^2 - \sum_{x_i \in C_2} D(x_i, m_{C_2})^2 \tag{4}$$

To compare the results of these different linkage methods, we will present different amount of clusters $K = 2, 3, 4$ for each linkage method. We will then show the resulting clusters in a scatter plot, a dendrogram and their silhouette scores.

3. Results

In this section we will present the results of our analysis. We used three ways to report our findings. Firstly, we used dendrograms, which are graphs that represents which clusters are combined at which proximity value. Secondly, we used scatter plots in which the different cluster associations can be seen by different colors. Thirdly, we present a table with silhouette scores, which is a score that indicates the goodness of the clustering.
The distribution of all data points can be seen in Figure 1

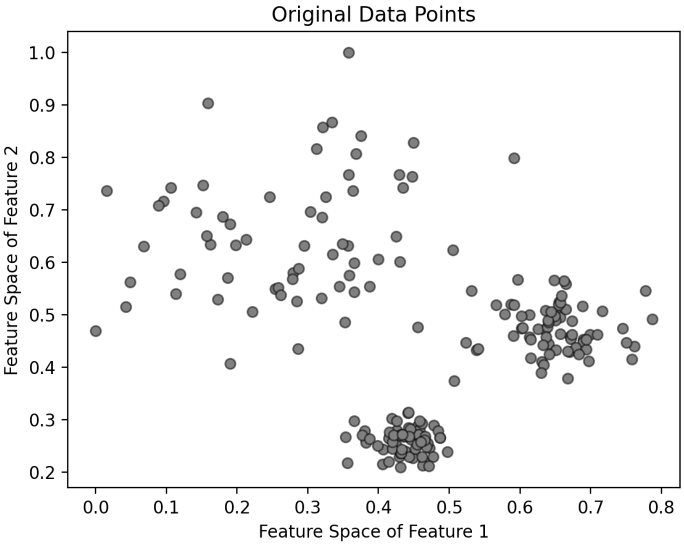


Figure 1: *All Data Points not organized in clusters.*

3.1. Single Linkage

For the single linkage method the following results have been obtained.

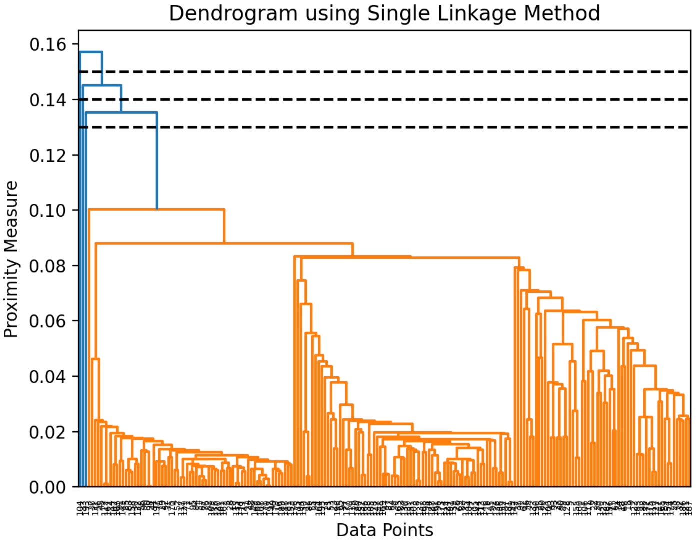
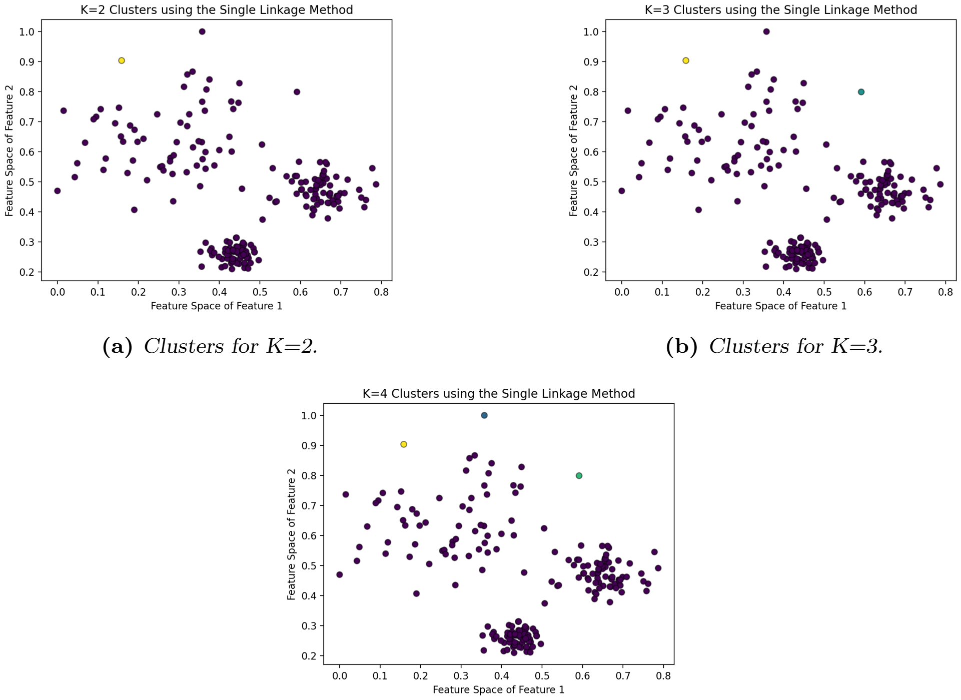


Figure 2: *Dendrogram for the Single Linkage Method. Cut-off thresholds are at 0.15, 0.14 and 0.13*

In Figure 2 we can see that $K = 2, 3, 4$ clusters are observed for cut-off thresholds of 0.15, 0.14 and 0.13, respectively. It can be seen that one big cluster is combined with a single data point for the last 3 clustering steps.



(a) *Clusters for K=2.*

(b) *Clusters for K=3.*

(c) *Clusters for K=4.*

Figure 3: *Clusters for K=2,3,4 determined using the Single Linkage Method.*

In Figure 3 we can see that for each value of K almost all data points are in a single cluster.

For each K value there is $K-1$ singleton clusters and 1 cluster with $N-K+1$ data points.

3.2. Average Linkage

For the average linkage method the following results have been obtained.

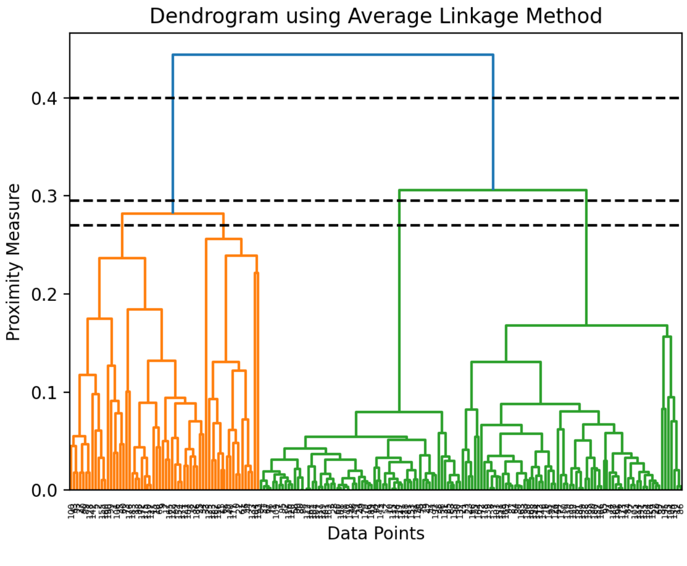
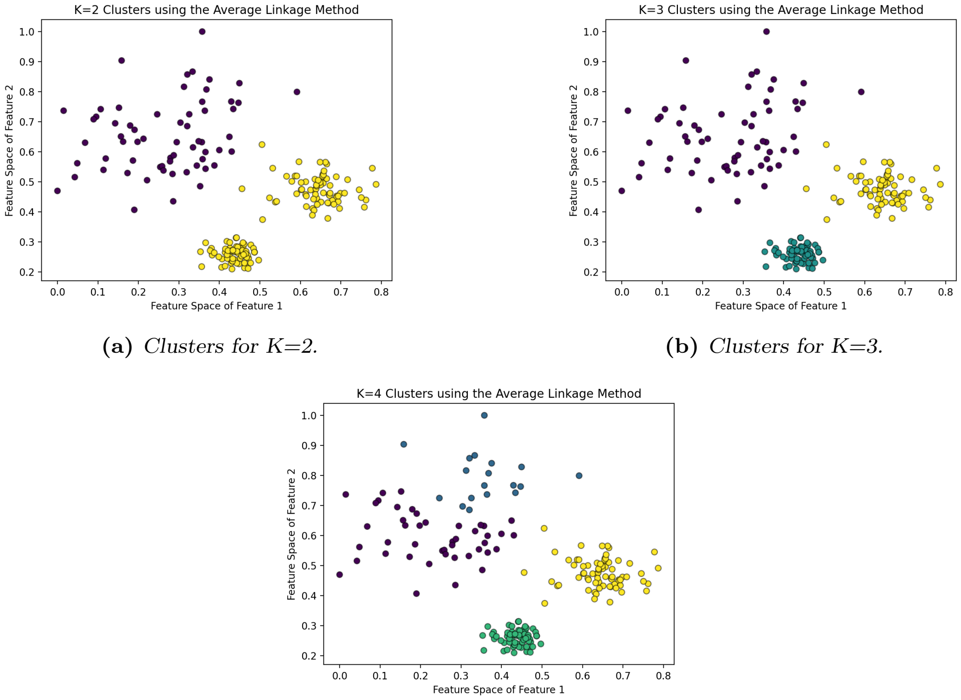


Figure 4: *Dendrogram for the Average Linkage Method. Cut-off thresholds are at 0.40, 0.295 and 0.27*

In Figure 4 we can see that $K = 2, 3, 4$ clusters are observed for cut-off thresholds of 0.40, 0.295 and 0.27, respectively. It can be seen that multiple big clusters are combined with each other for the last 3 clustering steps.



(a) *Clusters for K=2.*

(b) *Clusters for K=3.*

(c) *Clusters for K=4.*

Figure 5: *Clusters for K=2,3,4 determined using the Average Linkage Method.*

In Figure 5 we can see that for each value of K the distribution of data points among the clusters seems to be quite equal. We see that the separation into clusters for $K = 2$ distinguishes between points at the top left and bottom right. Separation for $K = 3$ splits the data points in the bottom right into two separate clusters. Finally, with $K = 4$ the data points on the top left are also split into two separate groups.

3.3. Complete Linkage

For the complete linkage method the following results have been obtained.

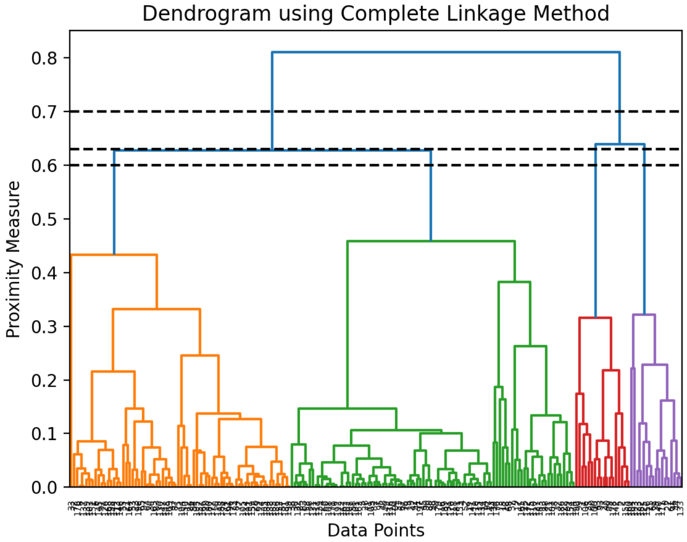
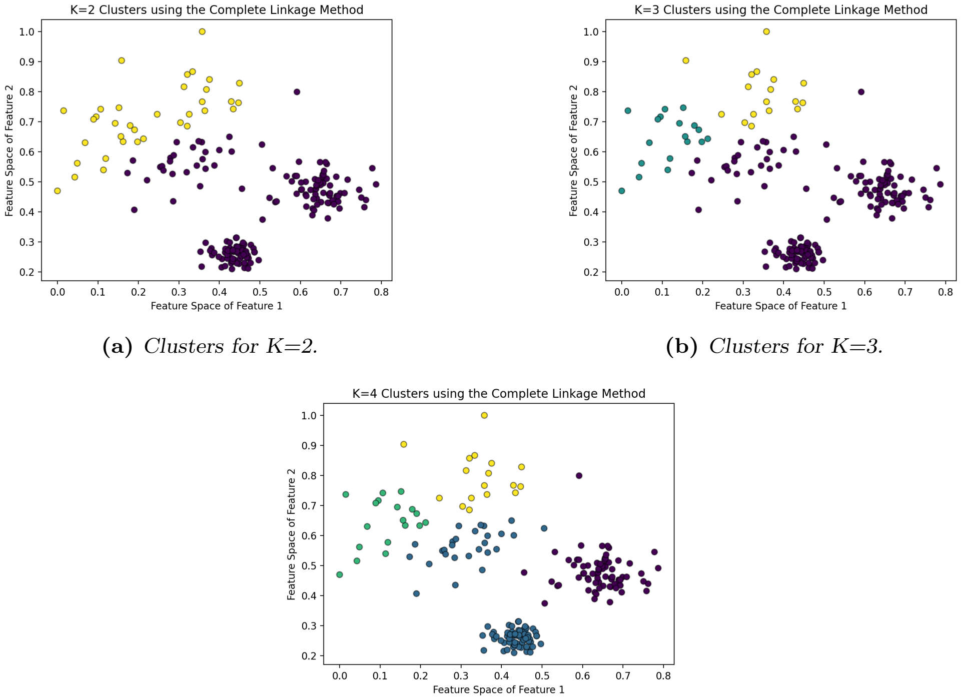


Figure 6: *Dendrogram for the Complete Linkage Method. Cut-off thresholds are at 0.70, 0.63 and 0.60*

In Figure 6 we can see that $K = 2, 3, 4$ clusters are observed for cut-off thresholds of 0.70, 0.63 and 0.60, respectively. It can be seen that multiple big clusters are combined with each other for the last 3 clustering steps.



(a) *Clusters for K=2.*

(b) *Clusters for K=3.*

(c) *Clusters for K=4.*

Figure 7: *Clusters for K=2,3,4 determined using the Complete Linkage Method.*

In Figure 7 we can see that for each value of K the distribution of data points among the clusters seems to be somewhat equal. For $K = 2, 3$ the cluster on the bottom right contains a majority of the data points. We see that the separation into clusters for $K = 2$ distinguishes between points at the top left and bottom right. Separation for $K = 3$ splits the data points in the top left into two separate clusters. Finally, with $K = 4$ the data points on the bottom right are also split into two separate groups.

3.4. Ward Linkage

For the ward linkage method the following results have been obtained.

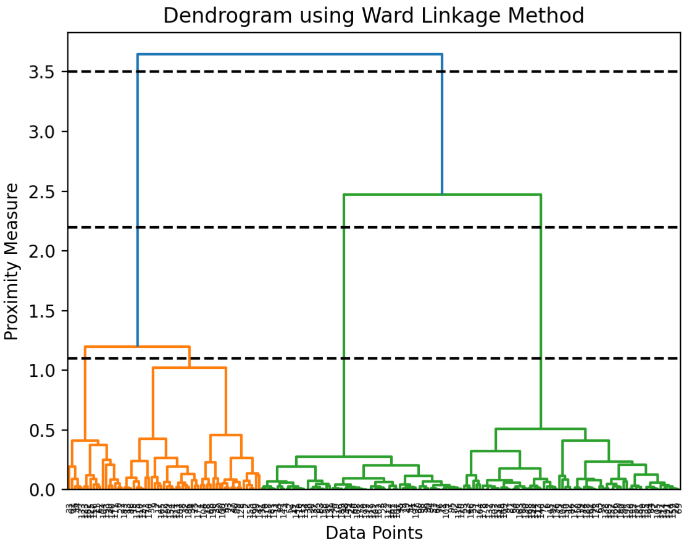
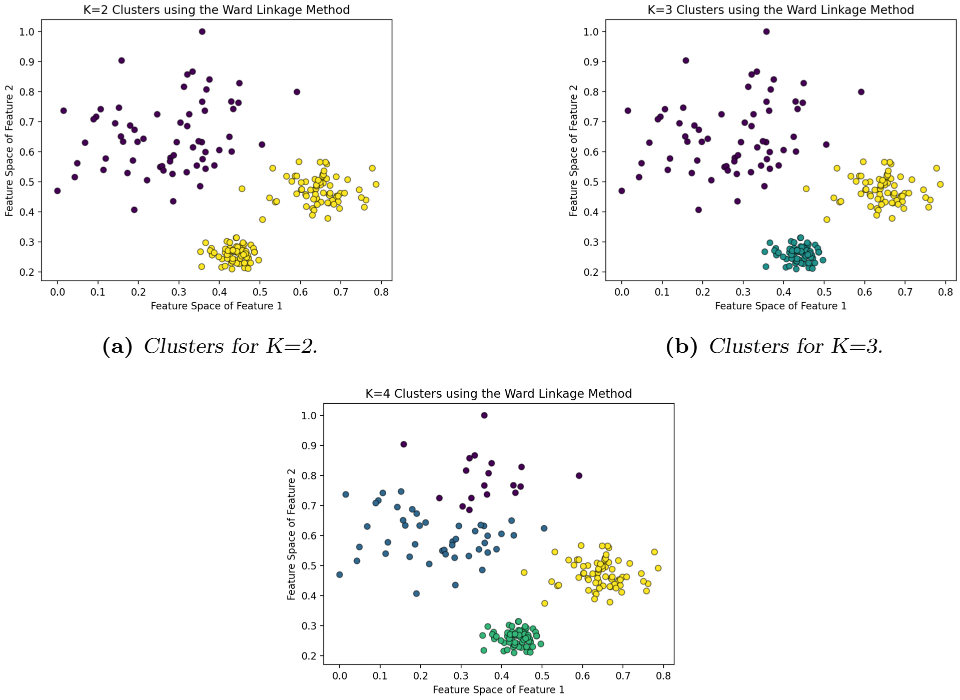


Figure 8: *Dendrogram for the Ward Linkage Method. Cut-off thresholds are at 3.5, 2.2 and 1.1*

In Figure 8 we can see that $K = 2, 3, 4$ clusters are observed for cut-off thresholds of 3.5, 2.2 and 1.1, respectively. It can be seen that multiple big clusters are combined with each other for the last 3 clustering steps.



(a) *Clusters for K=2.*

(b) *Clusters for K=3.*

(c) *Clusters for K=4.*

Figure 9: *Clusters for K=2,3,4 determined using the Ward Linkage Method.*

In Figure 9 we can see that for each value of K the distribution of data points among the clusters seems to be quite equal. We see that the separation into clusters for $K = 2$ distinguishes between points at the top left and bottom right. Separation for $K = 3$ splits the data points in the bottom right into two separate clusters. Finally, with $K = 4$ the data points on the top left are also split into two separate groups.

3.5. Silhouette Scores

The silhouette score of a clustering can give an insight on the effectiveness and goodness of the clustering. It's value is between -1 and 1, where a higher value indicates better matching clusters.

K Linkage Method	2	3	4
Single	0.377	0.180	0.185
Average	0.541	0.651	0.627
Complete	0.465	0.421	0.490
Ward	0.541	0.651	0.625

Table 1: *Silhouette Scores for the different Linkage Methods for K=2,3,4 Clusters.*

In Table 1 we can see that for each value of K the average and the ward linkage method have the highest silhouette values. While the silhouette values are somewhat stable or increasing for higher values of K for the average, complete and ward linkage measure, the silhouette score for the single linkage measure decreases with a higher values of K.

4. Discussion

There are some general observations which can be made without considering any specification. Just by looking at Figure 1, it is clear to the human eye that there are 3 to 4 clusters, the uncertainty caused by the data points between $0.0 \leq x \leq 0.5$ and $0.4 \leq y \leq 0.9$ which is the left top cluster, while the two closely packed data points around $0.3 \leq x \leq 0.5$ and $0.6 \leq x \leq 0.8$ are very clear. What is meant by these "obvious" clusters are more clear to the eye in Figure 5b. So a result of 3 or 4 clusters identified correctly can be strongly argued to be the best clustering, compared to other possible results. Still, all of these observations are not the ground truth and all depends on the application and context of data but these ideas are the way which our discussion will lean towards. For each linkage measure in this section, systematic cluster validation through silhouette scores is done and they are all mentioned in this discussion section. Each linkage measure and its results are discussed separately, before combining them in an overall conclusion.

4.1. Single Linkage

Single linkage is the most efficient measure yet it gave the worst results amongst all other linkages, because it performs poorly against noise and outlying data points. As seen from Figure 3, almost all data points are in a single cluster (color), thus the cluster scatter plot does not really help with identifying any clusters. The dendrogram is also unable to provide a ground for meaningful observations since it adds only a single data point to the cluster towards the end steps at K=2,3,4 as seen in Figure 2, so the dendrogram collapses towards the left side. The silhouette scores of single linkage are also low, especially when compared to ward linkage. So all three deductions made on single linkage are negative in a sense that it did not provide a good clustering and the clustering validations also support this. Thus, single linkage is not preferable in this case. It's efficiency is not a reason to prefer it because it provided bad results.

4.2. Average Linkage

Average measurement provided good and expected results. Visually, as seen by the color of clusters in figures Figure 5, the clusters are distinct from each other in a sense and are as expected. It's dendrogram shows how the clustering process happened with very distinct and further clusters, eventually meeting larger clusters towards the end, unlike the single linkage. As seen by the threshold points, the last clustering steps combine multiple large clusters, which is a good sign that it has done a good job at identifying the 3-4 plausible clusters. The silhouette scores of it is the best along with Ward method, further supporting the idea that it is the best clustering measure for this case and / or in general.

4.3. Complete Linkage

Complete linkage yielded mixed results. Seen in Figure 7a, for K=2, two plausible clusters are formed so that is good. Yet for K = 3 and K = 4, towards the bottom right and top left, clusters that do not make much sense are formed because they are not distinct clusters and look intertwined. The middle points of the complete linkage always seem to be within the bottom clusters. Whether or not this is good or bad for forming the clusters would depend on the context of the data itself, so this method will not be judged negatively on this. The silhouette scores are not exceptionally good or bad, but in the middle compared to other methods. So overall, the complete linkage measure has performed better than single linkage, yet worse than the other methods.

4.4. Ward Linkage

The ward method, even though not the most efficient, is known as the best assumed method overall as it also takes possible errors into consideration, so the best results are expected compared to all other methods. From Figure 9, it can be seen how the clusters make sense visually for all K values, there are no intertwined clusters and all clusters are plausible through observation. The dendrogram is similar to the average linkage, so just like it was for average linkage, it is an indicator that the ward linkage process performs well as it forms large, distinct clusters. These good results is accompanied by high silhouette scores which are the same as the average methods silhouette scores, the best silhouette scores amongst all the other methods.

4.5. Conclusion on Linkage Choice and K

For the best linkage choice, it is reasonable to rely on the silhouette scores. Ward and average linkage gave the best silhouette scores. Their scores are the same with average method, except for K = 4, in which there is a very small ignorable difference of 0.002, so it can be said that the two methods performed very similar. So, for this data set, without any knowledge on the context also, the average and ward methods yielded the best hierarchical clustering. Thus, ward or average are the best choices as a linkage measure. They performed well even with the noise and spread data in the data set working against them.

When it comes to K, the best number of clusters for this data set is $K = 3$. This is based on multiple facts deduced from both systemic and intuitive validation of clusters. In a systemic way, the maximum silhouette score achieved is when $K = 3$, with a significant difference between $K = 4$ score. Thus, it supports that $K = 3$ is the most plausible number of clusters. Meanwhile, as mentioned before, an intuitive observation on the data set (Figure 1) by the team, made expectations that either $K = 3$ or $K = 4$ is the best number of clusters and other numbers would not be plausible. By the human eye, the 3 clusters identified are mentioned in the start of section 4. So in conclusion, when $K = 3$ and the linkage measure selected is ward or average, the best results were obtained.

5. Contribution

Both team members attended the lab session for this assignment and both agree that the contribution to this assignment was equal and fair.

5.1. Code

The development of the code for the assignment was done in a lab session. Therefore, the contribution to the code base was entirely equal for both group members.

5.2. Report

Sections were split in half and thus the report was worked on by both members. After looking over the report together one last time, it was finalized and ready.

6. Code Appendix

6.1. clustering.py

```
import numpy as np
from calculation import *
from plotting import *
from matplotlib import pyplot as plt


def plot_dendrograms(link_single, link_average, link_complete,
    link_ward):
    plot_dendrogram(link_single, title="Dendrogram using Single
    Linkage Method", cutoffs=[0.15, 0.14, 0.13])
    plot_dendrogram(link_average, title="Dendrogram using Average
    Linkage Method", cutoffs=[0.4, 0.295, 0.27])
    plot_dendrogram(link_complete, title="Dendrogram using Complete
    Linkage Method", cutoffs=[0.7, 0.63, 0.6])
    plot_dendrogram(link_ward, title="Dendrogram using Ward Linkage
    Method", cutoffs=[3.5, 2.2, 1.1])


def plot_clusters(clusters_single, clusters_average, clusters_complete,
     clusters_ward):
    for i in range(3):
        plot_data_points(data, clusters=clusters_single[i], title=f"K
    ={i + 2} Clusters using the Single Linkage Method")
    for i in range(3):
        plot_data_points(data, clusters=clusters_average[i], title=f"K
    ={i + 2} Clusters using the Average Linkage Method")
    for i in range(3):
        plot_data_points(data, clusters=clusters_complete[i], title=f"
    K={i + 2} Clusters using the Complete Linkage Method")
    for i in range(3):
        plot_data_points(data, clusters=clusters_ward[i], title=f"K={i
     + 2} Clusters using the Ward Linkage Method")


def print_silhouette_scores(scores_single, scores_average,
    scores_complete, scores_ward):
    print("Silhouette Scores Single Linkage:")
    for i in range(3):
        print(f"K={i + 2}:", scores_single[i])
    print("Silhouette Scores Average Linkage:")
    for i in range(3):
        print(f"K={i + 2}:", scores_average[i])
    print("Silhouette Scores Complete Linkage:")
    for i in range(3):
        print(f"K={i + 2}:", scores_complete[i])
    print("Silhouette Scores Ward Linkage:")
    for i in range(3):
        print(f"K={i + 2}:", scores_ward[i])


if __name__ == '__main__':
    # Reading in the Data File
    data = np.loadtxt(open("data/data_clustering.csv", "r+"),
   delimiter=",")

    # apply different linkage methods
    link_single = single_linkage(data)
    print(link_single)
    link_average = average_linkage(data)
    link_complete = complete_linkage(data)
    link_ward = ward_linkage(data)

    # creating the different clusters for the different methods
    clusters_single = [
        clusters(link_single, 2),
        clusters(link_single, 3),
        clusters(link_single, 4)
    ]
    clusters_average = [
        clusters(link_average, 2),
        clusters(link_average, 3),
        clusters(link_average, 4)
    ]
    clusters_complete = [
        clusters(link_complete, 2),
        clusters(link_complete, 3),
        clusters(link_complete, 4)
    ]
    clusters_ward = [
        clusters(link_ward, 2),
        clusters(link_ward, 3),
        clusters(link_ward, 4)
    ]

    # compute silhouette scores
    scores_single = [
        silhouette_score(data, clusters_single[0]),
        silhouette_score(data, clusters_single[1]),
        silhouette_score(data, clusters_single[2])
    ]
    scores_average = [
        silhouette_score(data, clusters_average[0]),
        silhouette_score(data, clusters_average[1]),
        silhouette_score(data, clusters_average[2])
    ]
    scores_complete = [
        silhouette_score(data, clusters_complete[0]),
        silhouette_score(data, clusters_complete[1]),
        silhouette_score(data, clusters_complete[2])
    ]
    scores_ward = [
        silhouette_score(data, clusters_ward[0]),
        silhouette_score(data, clusters_ward[1]),
        silhouette_score(data, clusters_ward[2])
    ]

    # print report of silhouette scores
    print_silhouette_scores(scores_single, scores_average,
   scores_complete, scores_ward)
    # plot original data points
    plot_data_points(data, title="Original Data Points")
    # plotting the dendrograms with fitting cut-off thresholds
    plot_dendrograms(link_single, link_average, link_complete,
   link_ward)
    # plotting different clusters for the different linkage methods
    plot_clusters(clusters_single, clusters_average, clusters_complete,
    clusters_ward)
    plt.show()
```

6.2. CALCULATION.PY

```
from scipy.cluster.hierarchy import linkage, fcluster
from sklearn.metrics import silhouette_score


def single_linkage(data):
    return linkage(data, method='single', metric='euclidean')


def average_linkage(data):
    return linkage(data, method='average', metric='euclidean')


def complete_linkage(data):
    return linkage(data, method='complete', metric='euclidean')


def ward_linkage(data):
    return linkage(data, method='ward', metric='euclidean')


def clusters(data, amt):
    return fcluster(data, amt, criterion='maxclust')


def silhoutte_scores(data, fclusters):
    return silhouette_score(data, fclusters)
```

6.3. PLOTTING.PY

```
from scipy.cluster.hierarchy import dendrogram
from matplotlib import pyplot as plt
```

```
def plot_dendrogram(data, x_lab="Data Points", y_lab="Proximity
   Measure", title="", cutoffs=[]):
    plt.figure()
    plt.title(title)
    plt.xlabel(x_lab)
    plt.ylabel(y_lab)
    for val in cutoffs:
        plt.axhline(val, linestyle='--', color='black')
    dendrogram(data)


def plot_data_points(data, x_lab="Feature Space of Feature 1", y_lab="
   Feature Space of Feature 2", title="", clusters=None):
    plt.figure()
    plt.title(title)
    plt.xlabel(x_lab)
    plt.ylabel(y_lab)
    edge_color = (0, 0, 0, 0.5)
    if clusters is None:
        plt.scatter(data[:, 0], data[:, 1], color='gray', edgecolors=
   edge_color)
    else:
        plt.scatter(data[:, 0], data[:, 1], c=clusters, cmap='hsv',
   edgecolors=edge_color)
```