Homework 5

Tylman Michael
CSE 546 Machine Learning

4/5/2023

1 Problem 1:

1.1 Part a:

For problem 1a, we were tasked with finding the optimal number of clusters by finding the elbow point. Now, since I have a mathematics and physics background, I couldn't simply be happy with eyeballing it. So, I did a little bit of research on how to mathematically find an elbow point, and I found an interesting paper on detecting knee points in system behavior which I will link to HERE. The algorithm represented in this paper is implemented in the kneed package in python, and it is what I used to automatically detect the optimal number of clusters.

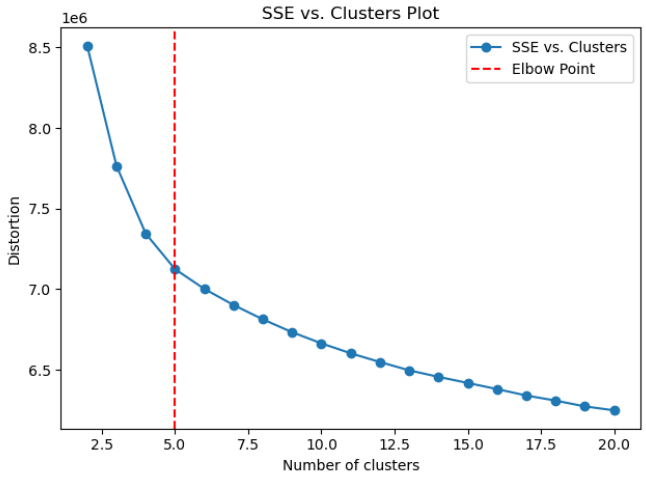
The SSE vs. No. clusters plot can be seen in the first part of Figure ??, where we can see the point decided by the kneed method marked as a red vertical line.

1.2 Part b:

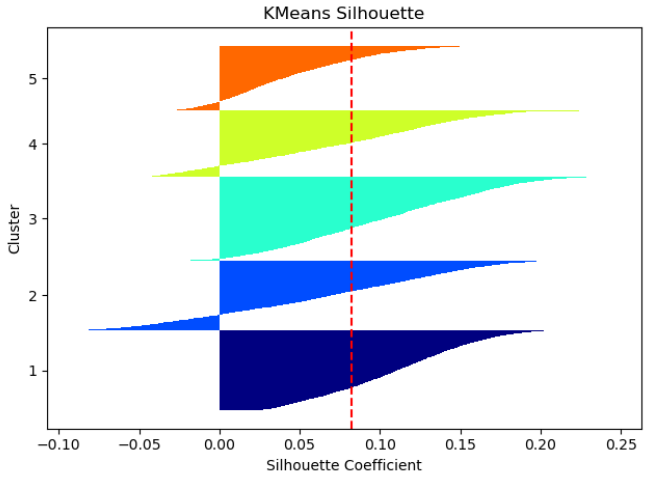
For problem 1b, we were tasked with generating the Silhouette plot for the optimal choice of clusters. This plot is shown in the second part of Figure ??. In this plot we can see that cluster number 1 performed the very best, with no elements past 0. However, we can still see that the overall average Silhouette score is quite low, clocking in at less than .10 with some clusters (cluster 2 in particular) showing non-trivial amounts of negative values.

1.3 Part c:

For problem 1c, we were tasked with finding images at the core of their clusters and on the edge of their clusters. To find the images at the core of their clusters, I took the top 5 highest scoring sample



(a) SSE vs. No. Clusters



(b) Silhouette Plot

from each cluster. To find the images at the edge of their clusters, I took the bottom 2 lowest absolute value samples from each cluster. All clusters except for number 1 had members very close to 0, but cluster 1 had it's lowest members hovering around .2 which was orders of magnitude higher than the other clusters. Given that behavior, I decided to not include cluster 1 results in the analysis of the edge cases given the stipulation of the assignment to only include border cases if they exist.

Looking at the images at the center of clusters, it's clear that the clusters are boats, cars, birds, horses, and planes, respectively. This will be important to keep in mind as we analyze the images at the edge of clusters in Figure 3.

Right away in cluster 2 (cars), we can see that these images are predominantly green and sky blue, respectively. I think that these backgrounds biased the values to appear more like horses and planes than they would normally seem. Additionally, the car on the right is a metallic color, which is unlike other cars in its class that are usually painted but more typical of planes. I think it is reasonable that these images are bordering two clusters, even though they were correctly clustered.

In cluster 3 (birds), we can see that the image of an ostrich was barely clustered with the birds. I think this is a totally fair way to treat the ostrich. If there were any creature I'd put on the border of birds and horses, it would be the ostrich. The next image on the border is a car which was incorrectly clustered as a bird, but notably it's the first image of a car I've come across which is both facing directly into the camera, and has a door open. Additionally, this car is in a field and does not have colors which are not readily found in nature. With these facts in mind, I think that this picture is likely reasonable for miss-clustering, but would warrant an analysis of the feature space to understand what features exactly caused this error.

In cluster 4 (horses), we can see that the first image has a smaller horse which is the darkest color in the image, along with having it's tail up. Additionally, the front legs are a bit blurred, which causes the image to evoke a similar shape and coloration to a rooster nearby a barn. I think that these factors caused this image to be placed on the edge of the horse cluster. The right image of the horse cluster is what I believe to be the most intriguing image in this analysis. This image appears to clearly be a horse, and the only features I can find that could cause it to struggle are the inclusion of the person, and the tri-layer background of alternating light-green-dark that was also present in the left horse image.

Finally, in cluster 5 (planes), we can see that the first image is of a plane taken at the same altitude as plane. This angle only leaves the tail fin and the cockpit as defining features, and most importantly removes any information about the plane's wings. I think that perfectly symmetric wings angled towards a point is likely one of the key differentiating features of a plane vs. a boat, given their similar blue shifted backgrounds lacking foliage. On the right image we see a duck resting in choppy water. Most importantly, we can see the ducks reflection in the water, which is giving it a symmetric shape (the heads) angled towards a point (the tail). I think this symmetry along with the cloud-colored water is what caused this image to be clusterd on the edge of the plane cluster.

2 Problem 2:

2.1 Part a:

When it came time to analyze the dendrogram to select the optimal amount of clusters, it was difficult to remove my inherent bias towards selecting 5 clusters. Given the fact that the KMeans gave 5 clusters and the ground truth had 5 classes I had strong prompting to select 5 clusters again. But I wanted to look at this through an objective lens and find a feature of the dendrogram which could justify my choice.

After spending some time looking at Figure 4, I realized that visually the optimal choice of 5 clusters minimized the ratio of the variance of the cluster distance over the average cluster distance weighted by the inverse of the number of clusters. This is my first interaction with a dendrogram for picking number of clusters, so I was unable to develop and test an automated function which matches my hypothesis, and I could not find online others who may have calculated this in the same way. Likely, the fact that this method is not easily findable online means that my hypothesis is flawed and may be overly influence by the behavior of this toy dataset which is well-balanced. Regardless, it is the reasoning

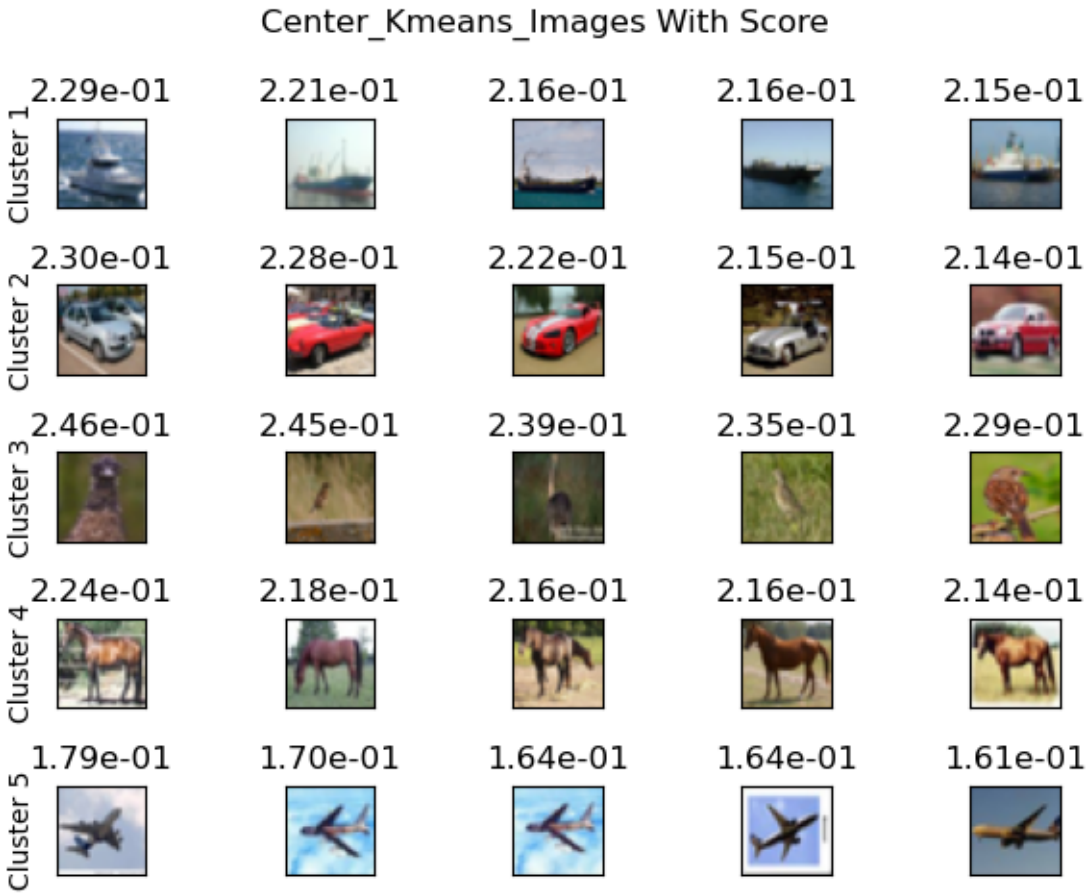


Figure 2: Images at the Center of Clusters Kmeans

Figure 3: Images at the Edge of Clusters Kmeans

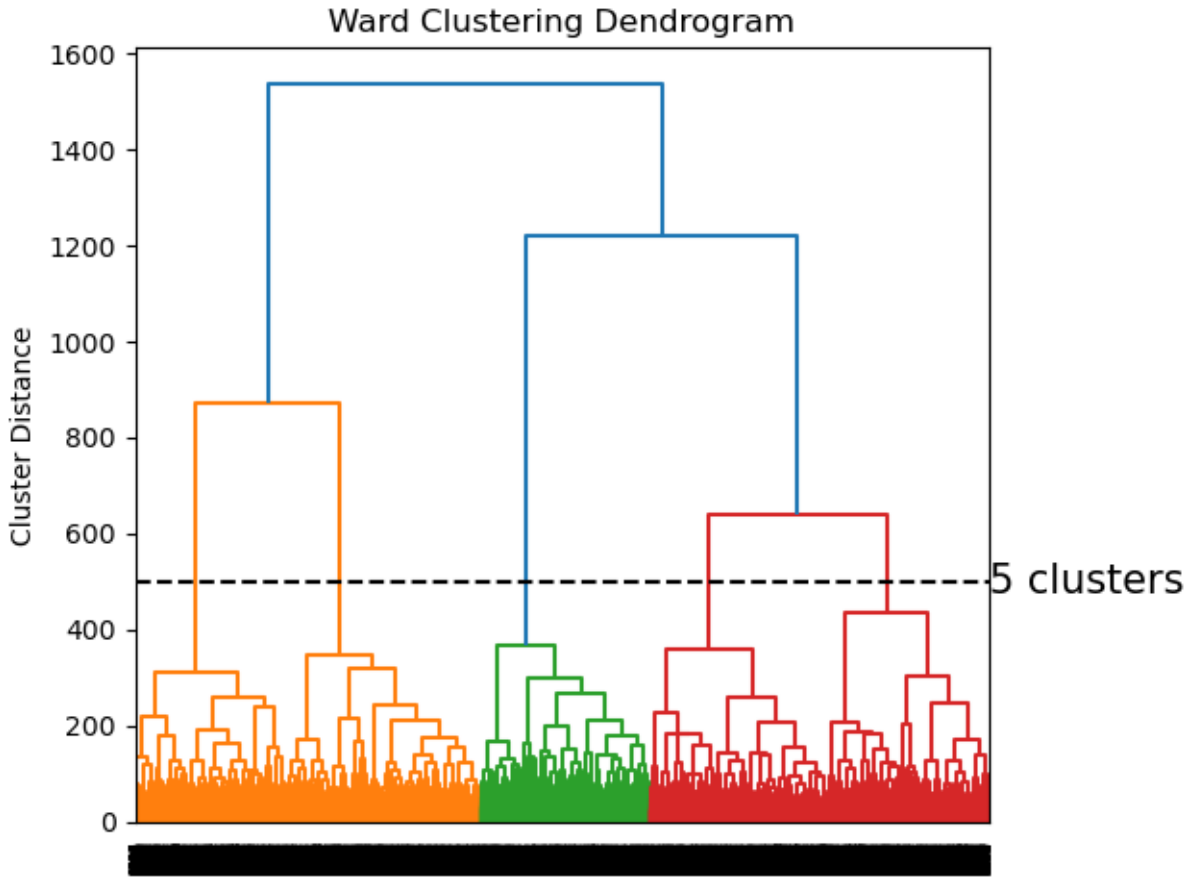


Figure 4: Ward Dendrogram

which I found to best capture why I believe the dendrogram shows 5 clusters to be the optimal number of clusters for the ward method.

2.2 Part b:

The Silhouette plot for the ward method is in Figure 5. This figure is evidence that the agglomerative method may not be better than the KMeans algorithm for this dataset. Previously, the clusters showed far fewer amounts of points with a negative Silhouette score, even at a similar overall average score. More in-depth discussion is to follow in part c:

2.3 Part c:

The repeated plots of parts a and b for single linkage and complete linkage are shown in figures ?? and ??, respectively.

Starting off with the odd one out, the single linkage clustering algorithm appears to have completely missed the mark. It seems to have piece-by-piece grown one cluster a single sample at a time until it encompassed the whole group. This plot and it's accompanying Silhouette plot show beyond a shadow of a doubt that single linkage is not the direction to go with this data. Even when I forced the number of clusters to be 5, the single linkage clustering method caused every item to be placed in one group, and no items to be placed in any other clusters leaving 4 empty clusters.

Next, the complete clustering algorithm appears much more reasonable. In an attempt to be fair, I am going to try and eyeball my same reasoning I used before. Looking around the line provided in the image, we can see that it appears like 6 clusters are the optimal number of clusters for this algorithm. However, we can also see that the variance between the clusters (and cluster balance, but that must imply a balanced dataset) is much worse than the ward algorithm managed. in the Silhouetteplot, we can see that cluster 5 is nothing but an extremely tiny sliver, and should likely be merged in with

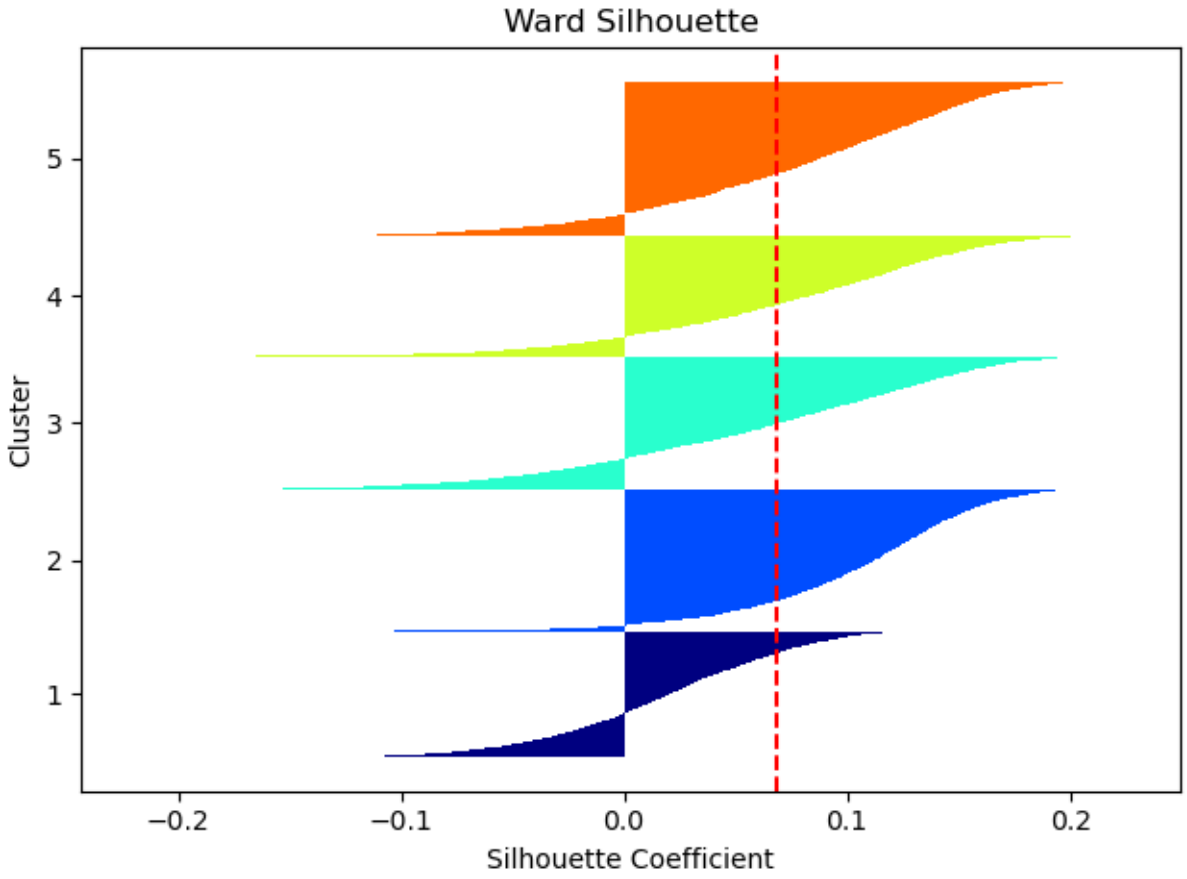
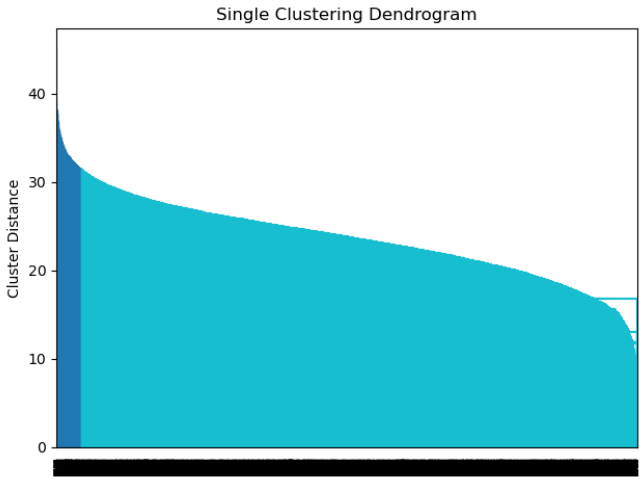
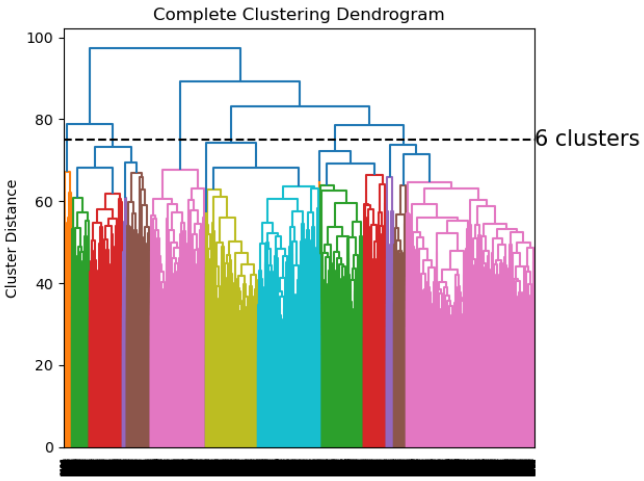


Figure 5: Ward Silhouette

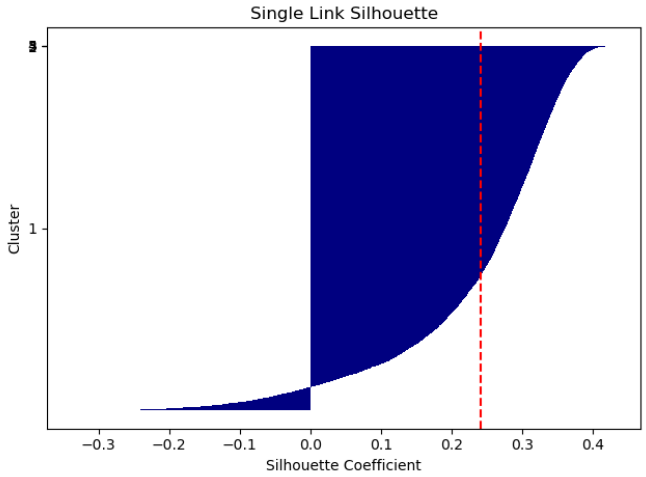
another cluster.



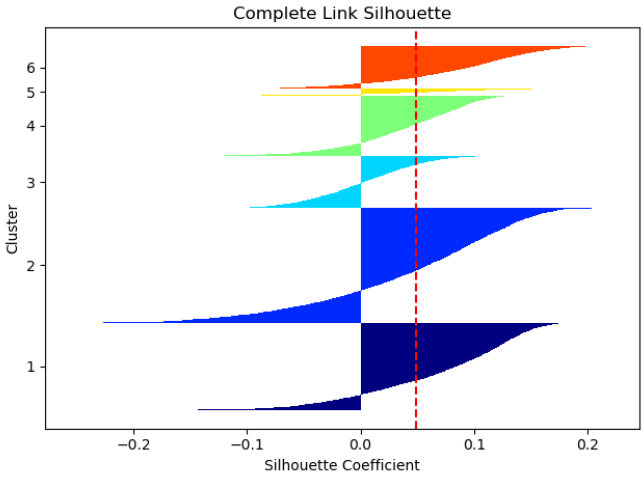
(a) Single Dendrogram



(b) Complete Dendrogram



(a) Single Silhouette



(b) Complete Silhouette