CSE 590-04 Homework 4

Aliaa Elshamekh

March 21, 2021

1 Introduction

Unsupervised learning means you have a data set that is completely unlabeled. We don't know if there are any patterns hidden in the data,so we leave it to the algorithm to find anything it can.There are a lot of different unsupervised learning techniques,like neural networks,reinforcement learning,and clustering. The specific type of algorithm we want to use is going to depend on what the data looks like.That's where clustering algorithms come in. It's one of the methods we can use in an unsupervised learning problem.Using a clustering algorithm means we are going to give the algorithm a lot of input data with no labels and let it find any groupings in the data it can.Those groupings are called clusters.A cluster is a group of data points that are similar to each other based on their relation to surrounding data points.Clustering is used for things like feature engineering or pattern discovery.Moreover, might want to use clustering when we're trying to do anomaly detection to try and find outliers in the data. It helps by finding those groups of clusters and showing the boundaries that would determine whether a data point is an outlier or not.Clustering is especially useful for exploring data we know nothing about. It might take some time to figure out which type of clustering algorithm works the best, but when we do,we will get invaluable insight on your data.

There are different types of clustering algorithms that handle all kinds of unique data.In this assignment,we will work with just two types of them.

- Centroid-based
  It is a little sensitive to the initial parameters you give it,but it's fast and efficient.These types of algorithms separate data points based on multiple centroids in the data.Each data point is assigned to a cluster based on its squared distance from the centroid.This is the most commonly used type of clustering.We are going to use the K-Means algorithm.
- Hierarchical-based
  It is typically used on hierarchical data,like you would get from a company database or taxonomies. It builds a tree of clusters so everything is organized from the top-down.This is more restrictive than the other clustering types,but it's perfect for specific kinds of data sets.We are going to use Agglomerative Hierarchy clustering algorithm.

We are going to use CIFAR-10 dataset which contain(32,32) colored 10K images,wherease the data we are going to work on is 128 features extracted for all the 10K images.

2 K-means Clustering

This algorithm is the most commonly used clustering algorithm.It's a centroid-based algorithm and the simplest unsupervised learning algorithm.This algorithm tries to minimize the variance of data points within a cluster.It's also how most people are introduced to unsupervised machine learning.K-means is best used on smaller data sets because it iterates over all of the data points.That means it'll take more time to classify data points if there are a large amount of them in the data set.Generally,since this is how k-means clusters data points, it doesn't scale well.In our dataset since the features are just pixel intensities ,all of them has almost the same range.Hence, we won't scale the features.

The Elbow Method is one of the most popular methods to determine this optimal value of k.We are going to it based on the "Distortion" which is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.

In figure 1,in order to determine the optimal number of clusters,we have to select the value of k at the "elbow" i.e.

the point after which the distortion start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 5.

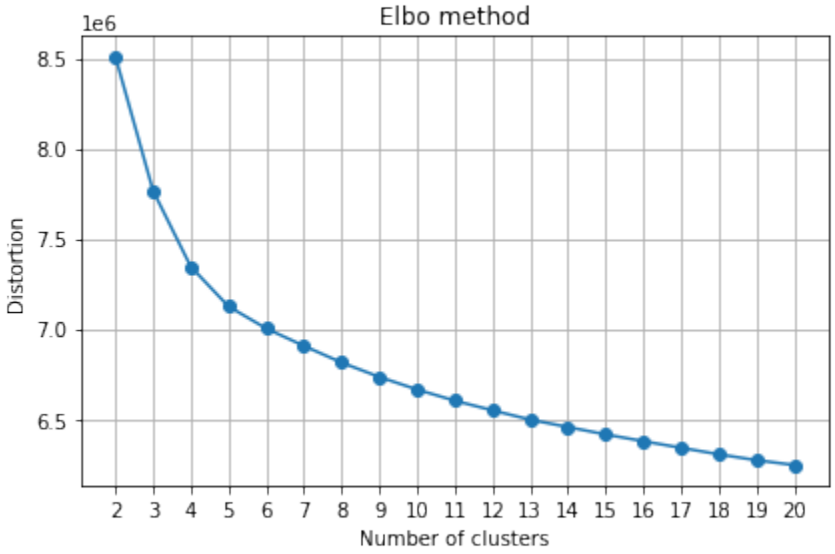


Figure 1: Elbow graph for k-Means clustering.

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other.The Silhouette score is calculated for each sample of different clusters.To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters:

- mean intra-cluster distance denoted as **a**
  Mean distance between the observation and all other data points in the same cluster.
- mean nearest-cluster distance denoted as **b**
  Mean distance between the observation and all other data points of the next nearest cluster.

Silhouette score, **S**, for each sample is calculated using the following formula:

$$\frac{\mathbf{b}-\mathbf{a}}{\max(\mathbf{b}-\mathbf{a})}$$

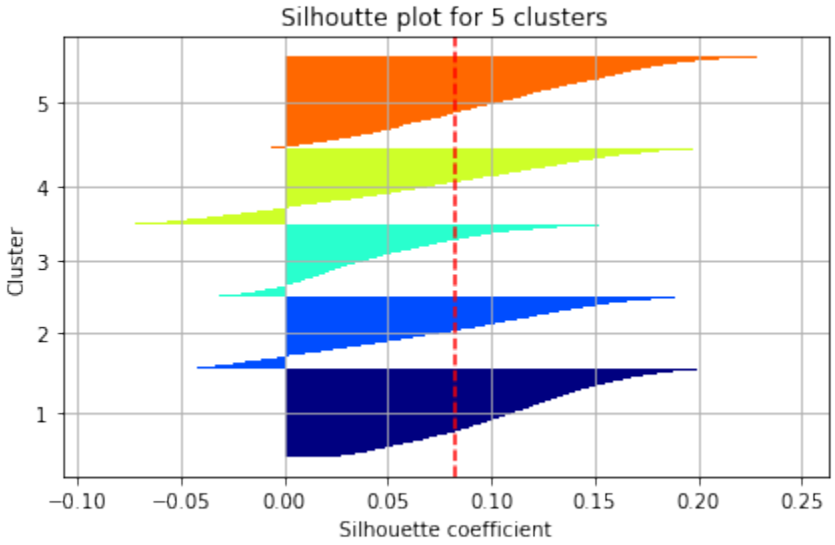


Figure 2: silhouette plot for $k = 5$.

The silhouette score falls within the range $[-1, 1]$.The silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping. The score of less than 0 means that data belonging to clusters may be wrong/incorrect The thickness of the silhouette plot representing each cluster also is a deciding point.For the plot in figure 2 with n_cluster 5, the thickness is more uniform than the plot with n_cluster as 4 with one cluster thickness much more than the other. Thus, one can select the optimal number of clusters as 5. The average silhouette score is 0.08.

To sum up the final k-means model is :
Number of clusters is $k = 5$ **and silhouette score is** 0.08

Using the silhouette coefficients we are able to identify samples from the core of each cluster by finding the index of the maximum 5 coeff values among each cluster as shown in 3.

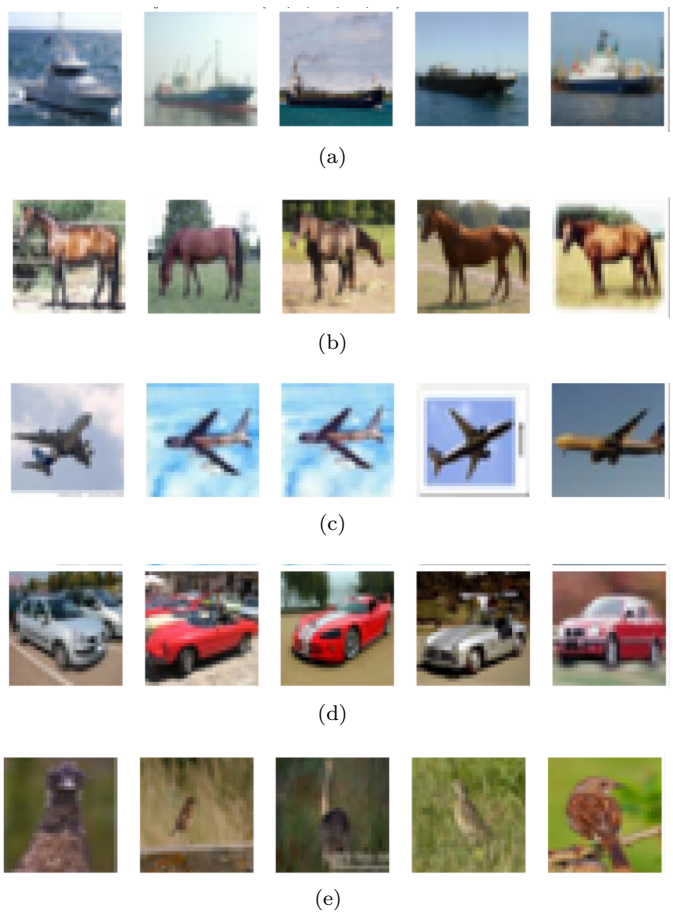


Figure 3: Samples from the core of each cluster (a) Cluster '0' (b) Cluster '1' (c) Cluster '2' (d) Cluster '3' (e) Cluster '4'

and figure 4 shows samples that are on the boundary of each cluster.

Figure 4: Samples at the boundary of each cluster (a) Cluster '0' (b) Cluster '1' (c) Cluster '2' (d) Cluster '3' (e) Cluster '4'

3 Agglomerative Clustering

This is the most common type of hierarchical clustering algorithm.It's used to group objects in clusters based on how similar they are to each other.This is a form of bottom-up clustering,where each data point is assigned to its own cluster.Then those clusters get joined together.At each iteration, similar clusters are merged until all of the data points are part of one big root cluster.Agglomerative clustering is best at finding small clusters.The end result looks like a dendrogram so that you can easily visualize the clusters when the algorithm finishes. This algorithm is based on defining each data point as a cluster and combine existing clusters at each step.Here are three different methods for this approach we are asked to use:

- Ward's method
  This method does not directly define a measure of distance between two points or clusters.It is an ANOVA based approach.One-way univariate ANOVAs are done for each variable with groups defined by the clusters at that stage of the process.At each stage,two clusters merge that provide the smallest increase in the combined error sum of squares.
- Single method
  we define the distance between two clusters as the minimum distance between any single data point in the first cluster and any single data point in the second cluster.On the basis of this definition of distance between clusters, at each stage of the process we combine the two clusters with the smallest single linkage distance.
- Complete method
  we define the distance between two clusters to be the maximum distance between any single data point in the first cluster and any single data point in the second cluster.On the basis of this definition of distance between clusters, at each stage of the process we combine the two clusters that have the smallest complete linkage distance.

In figure 5, we see the Dendogram of the data using Ward's method.The three horizontal lines represents 3,4 and 5 clusters .The highest jump in the dendogram is between 4 and 3 which suggests that the best number of clusters in such case is 4.The silhouette plot in figure 6 shows how the clusters' thickness is uniform.The average silhouette score is 0.08 for this case.

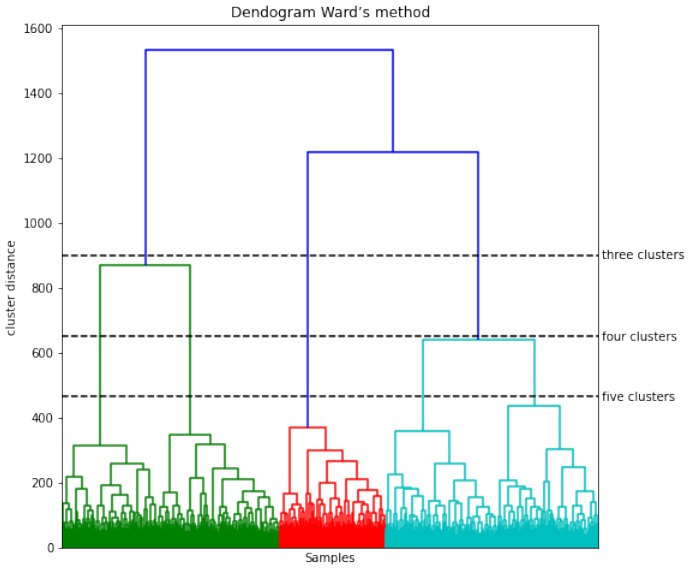


Figure 5: Dendogram for the data using Ward's method.

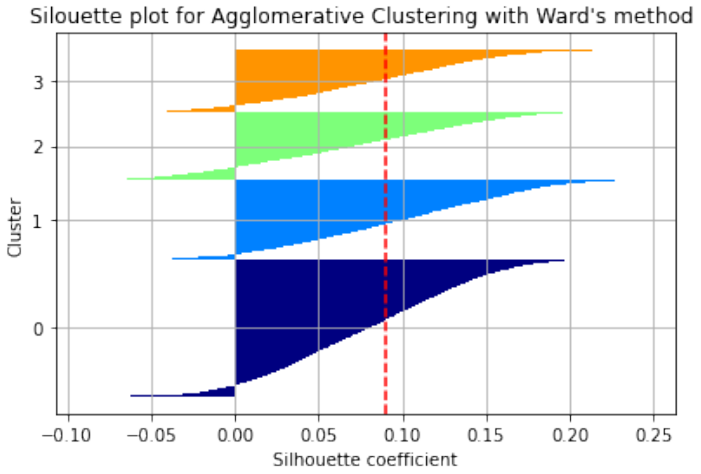


Figure 6: Silhouette plot for Agglomerative Clustering with Ward's method for 4 clusters.

In figure 7,we see the Dendogram of the data using Single-link method.We can see how is the data consists of just two clusters where cluster 0 has 9999 samples and cluster 1 has a single sample.This happens because of the way this technique works since at every updating step we choose the minimum of the two distances and two clusters of objects can be merged when there is a single close link between them,irrespective of the other inter-object distances.In addition the data is a high-dimensional which makes the sample are sparse in the space.

Silhouette plot in figure 8 illustrates how the clusters don't have a uniform thickness.

Therefore,this is not a suitable choice for our dataset,because it leads to clusters that are quite heterogeneous internally,and the usual object of clustering is to obtain homogeneous clusters.The average silhouette score is 0.29 for this case.

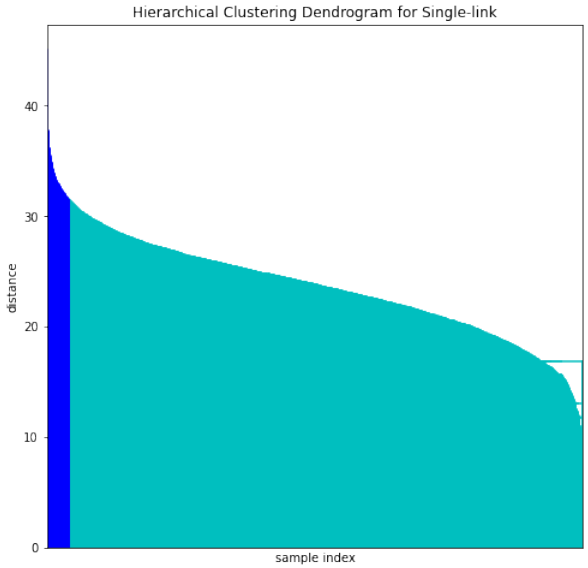


Figure 7: Dendogram for the data using Single-link method.

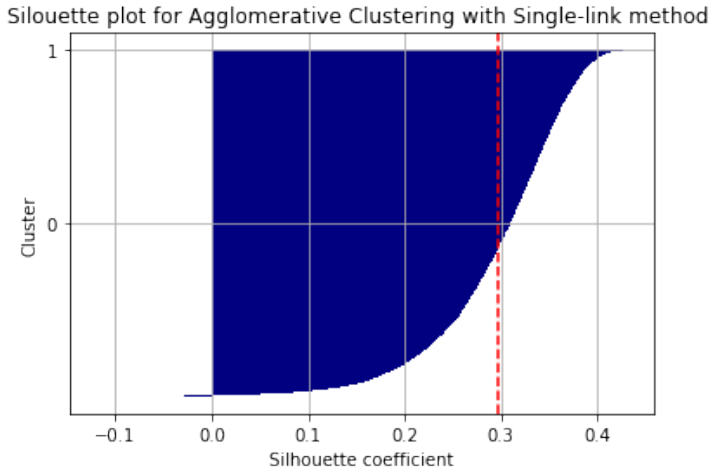


Figure 8: Silhouette plot for Agglomerative Clustering with Single-link method for 2 clusters.

In figure 9, we see the Dendogram of the data using Complete-link method.The three horizontal lines represents 3,4 and 6 clusters .The highest jump in the dendogram is between 6 and 4 which suggests that the best number of

clusters in such case is 4.The silhouette plot in figure 10 shows how the clusters' thickness is uniform.The average silhouette score is 0.05 for this case.

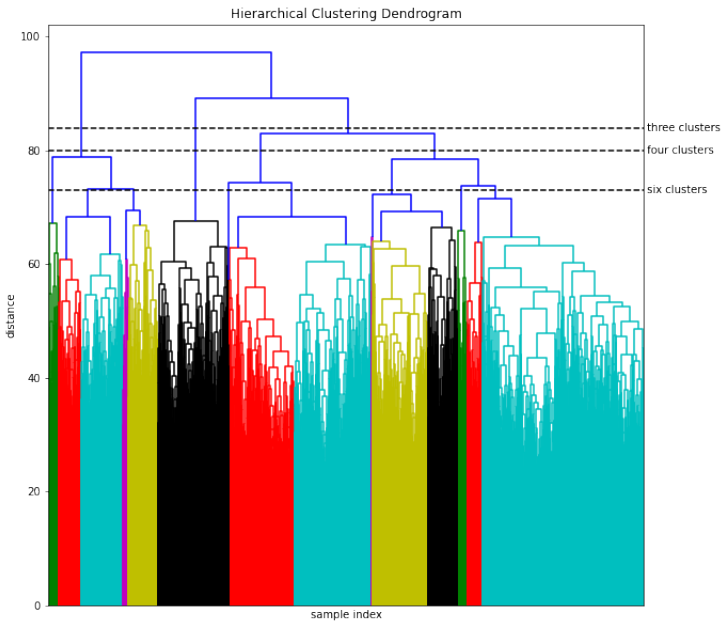


Figure 9: Dendogram for the data using Complete-link method.

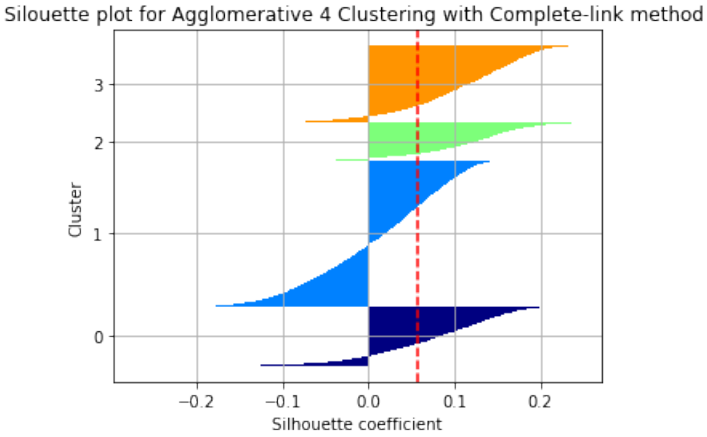
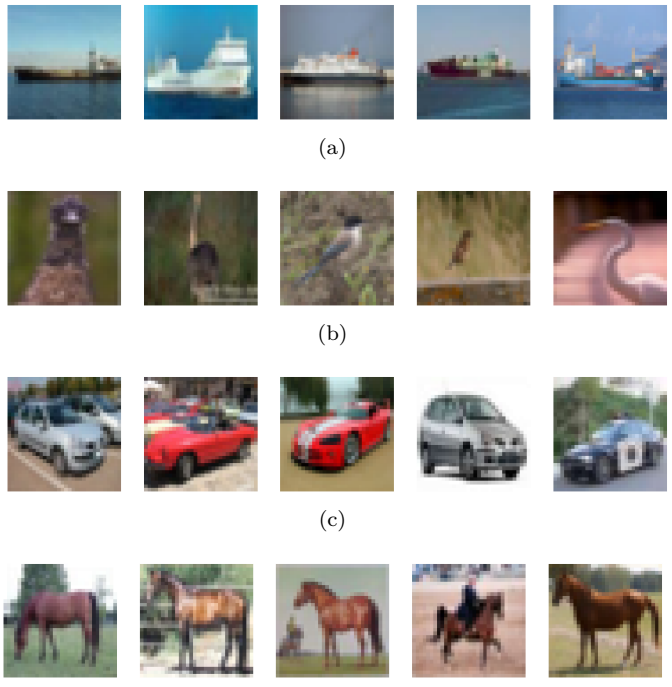


Figure 10: Silhouette plot for Agglomerative Clustering with Complete-link method for 4 clusters.

The best method for hierarchical agglomerative is the Ward's method.The data will be clustered into 4 groups with a silhouette score of 0.08.

Using the silhouette coefficients we are able to identify samples from the core of each cluster by finding the index of the maximum 5 coeff values among each cluster as shown in 11.



(a)

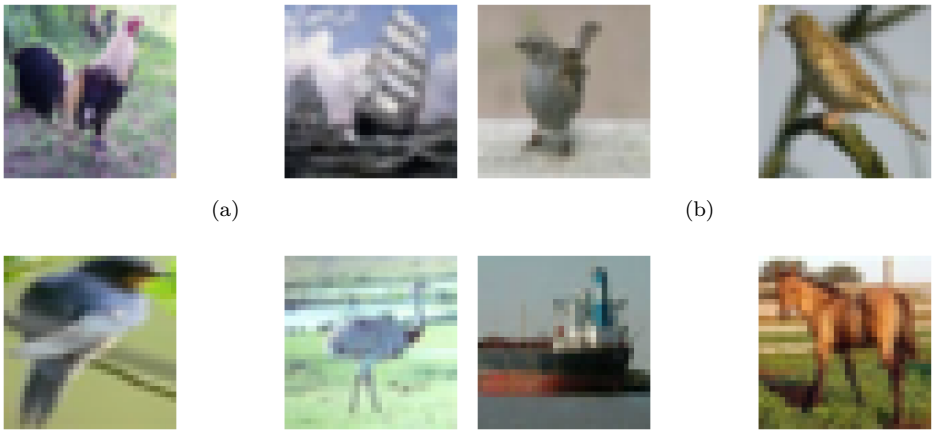
(b)

(c)

(d)

Figure 11: Samples from the core of each cluster (a) Cluster '0' (b) Cluster '1' (c) Cluster '2' (d) Cluster '3'

and figure 12 shows samples that are on the boundary of each cluster.



(a) (b)

(c) (d)

Figure 12: Samples at the boundary of each cluster (a) Cluster '0' (b) Cluster '1' (c) Cluster '2' (d) Cluster '3'

4 Adjusted Rand index Comparison

In table 1,we see the random index score for both k-mean model with 5 clusters and agglomartive model using Ward's method. The better model is the one with higher score.

	K-Means Model	agglomerative with Ward's method Model
Adjusted Rand Index	0.76	0.56
Number of Clusters	5	4
silhouette mean score	0.08	0.08

Table 1: Summary of the various model performance on testing and training data

The k-means model is better than the agglomartive model for my experiments.Since it has the higher random index score.Morover, we can see in the Silhouette plot that mean Silhouette score is equal to the Ward's method,however the clusters has uniform thickness in k-means relative to the aggolmarative model. Also, using kmeans model , we have lower number of outliers in each cluster.