Master International E3A

PRACTICAL REPORT

FOR

UNSUPERVISED LEARNING - CLUSTERING PRACTICAL

Classification of Handwritten Digits after Exploiting K-means/Hierarchical Clustering

BY:

ABODE DANIEL

SUBMITTED TO:

Prof. Sonia Garcia

DATE: 19/06/2020

Abstract

This practical is about classification of handwritten digits by two different clustering method; K-means and Hierarchical Clustering. The data set use was gotten from UC Irvine ML Repository. The aim is to understand how to build a K-means clustering algorithm and Hierarchical Clustering algorithm, to train it and to test its performance on some test data set. The performance was meant to be compared. Further analysis were conducted by varying some of the hyper parameters of the model, especially the initialization for k-means clustering, and new model was developed for better performance.

Table of Contents

Classification of Handwritten Digits by a K-means and Hierarchical Clustering

1.1 Introduction

Unsupervised learning refers to grouping closely related data set together to define a class or cluster. The objective is to partition the set of data into meaningful clusters, with each clusters containing similar elements. There are 2 major classes of clustering algorithm; the Hierarchical clustering and the Partitional. K-means an example of Partitional clustering that uses square error to group similar elements was implemented in this practical to classify handwritten digits. Its performance was compared to the performance of a Hierarchical-clustering algorithm. Silhouette scores were calculated to determine the best number of clusters in each case.

The laboratory exercise consist of 6 parts;

1. Understanding of the data set
2. Implementation of K-means clustering algorithm
3. Testing and evaluation of the K-means model
4. Implementation of Hierarchical clustering algorithm
5. Testing and evaluation of the Hierarchical clustering algorithm
6. Comparison of the performance of the two models

1.2 Experimental Setup

The experimental setup involve the use of python programming language development environment, for this practical the Jupyter development environment was utilized. Some required API were also installed including;

1. numpy
2. Matplotlib,
3. Sklearn

The Data used was a preprocessed image of handwritten digits, from which normalized bitmaps has been extracted. 30 persons contributed to the training set and 13 persons contributed to the test set. The training data is a matrix of 3823 x 65, the first 64 columns are the features and the 65th column is the label. For the test data, it is 1797 handwritten digits.

1.3 Objective

The practical objectives is as follow;

1. To be introduced to unsupervised learning - clustering techniques with python.
2. To design a model of K-means clustering techniques and a model of hierarchical clustering techniques.
3. To vary the hyper parameters and study the effect on the performance of the model.
4. To compare the performance of K-means clustering and hierarchical clustering and to adjudge the difficulty in realizing them.

1.4 Practical Procedure

1.4.1 Training Procedure for K-means for number of clustering K = 10

1. The required APIs were imported including numpy, matplotlib, seaborn, sklearn and scipy.
2. The raw data (training_data, test_data) were imported using np.loadtxt().
3. The training data was splitted into features 3823x64 and labels 3823x1(column 65).
4. Histogram of the labels of the trainng data and test data were plotted and can be seen in figure 1 and figure 2.
5. The test data was splitted into features 1797x64 and labels 1797x1(column 65).
6. The KMeans class of sklearn API was invoked to create an object of it with the following parameter; n_clusters = 10, init = ‘k-means++’. This was use to fit_predict the training_features.
7. The verbose was set to 1, to visualise the simulation process. The algorithm iterated 10 times and varied the initialization of the kmeans at each time. The computation with the minimum inertia (quantization error J) was chosen.
8. The accuracy of the model was evaluated with accuracy, number of digits per cluster histogram and the training confusion matrix and this is shown jointly in figure 3.
9. Because we were not content with the performance of the previous model, the KMeans model was repeated again with init = ‘random’, this was simulated in a for –loop until we realized an accuracy of 87.8368%. The cluster center variable at that accuracy was saved

as a csv file and imported and use for our model. The performance of this new model was better as shown in figure 4.

1.4.2 Measuring the Quality of Clustering with the Silhouette

1. The value of the attribute n_cluster for the Kmeans class was varied from 10 to 25 and the value of the silhouette was plotted as shown in figure 5.

1.4.3 Testing the best Clustering Model

1. Per cluster we did majority voting to allocate the test data to suitable cluster
2. This was done by finding the closest cluster center to each data digit
3. The histogram, confusion matrix and global performance (accuracy and number of digits in the Test Set correctly classified) was reported.

1.4.4 Comparison with a Hierarchical Clustering (with Ward linkage)

1. The model was made from the object of Agglomerative Clustering of kmeans with distance threshold set to 400 which correspond to 10 clusters.
2. The model was fit on the training data
3. The dendogram was cut at K = 10 and visualize as shown in figure 7.
4. The silhouette at K = 10 was computed and the histogram per clusters was made, also the accuracy was also computed. The training confusion was also provided in figure 8.
5. The model was tested on the test data following the same procedure, the confusion matrix, and test accuracy were computed.
6. To confirm that K=10 has the best silhouette for this method as well, we computed the silhouette for K = 8 to 25 and the plot was computed as shown in figure 11.

1.5 Result and Discussion

1. The histogram of the training data is shown below in figure 1, it is obvious that the data contain more than 350 samples of handwritten digits for each number 0 to 9. And for the test data we have about 175 samples per digits as shown in figure 2.

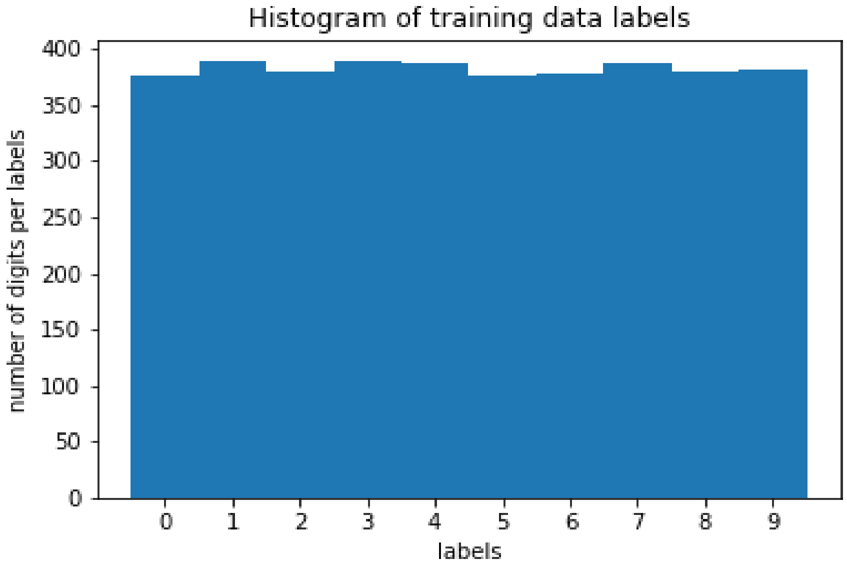


Figure 1: Histogram of Training Data Labels

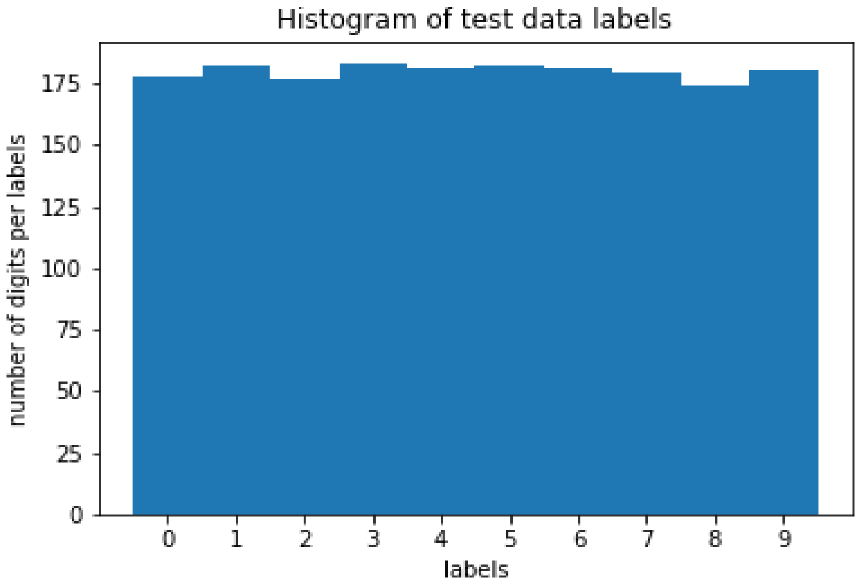
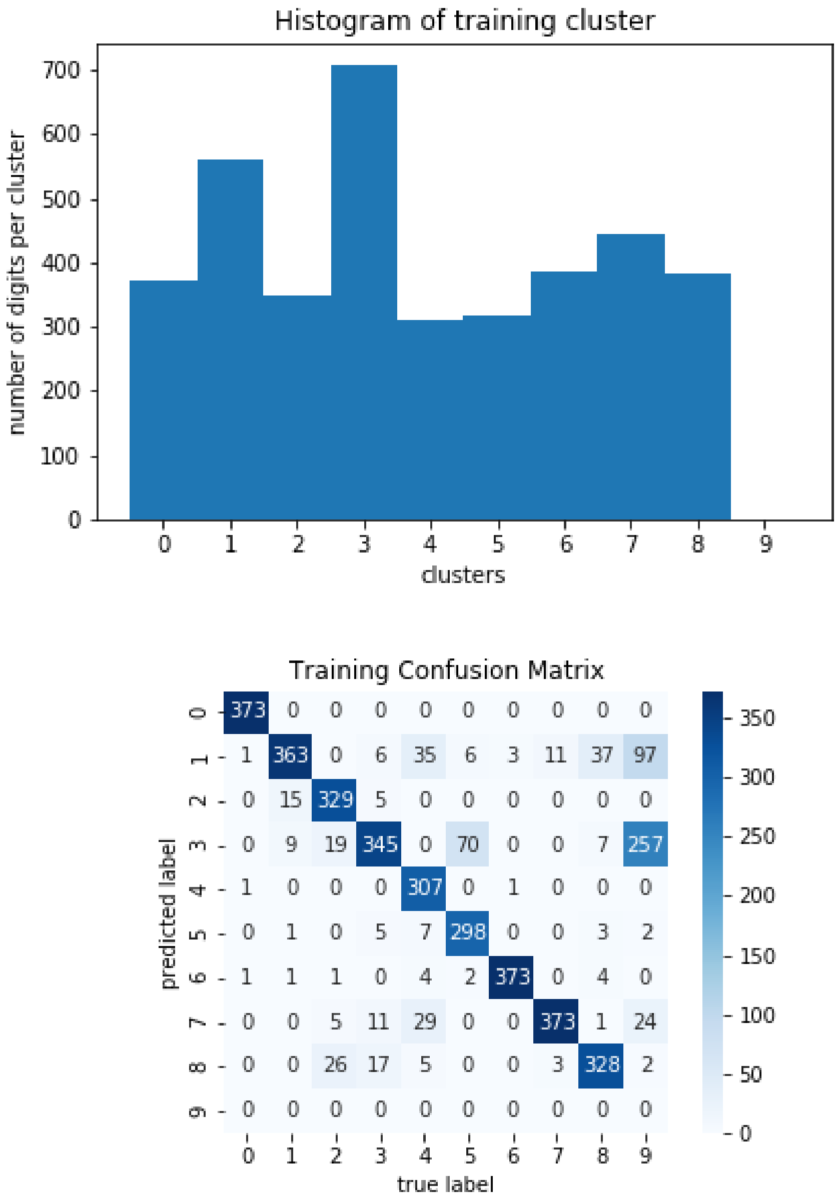


Figure 2: Histogram of Test Data Labels

2. After fitting the model with initialization set to ‘k-means++’, the result from predicting with the training label gave an accuracy of 80.8004 % as shown below. From the number of digits predicted per cluster, we found out that digit 9 was not predicted according to the true label, in fact, as shown in the confusion matrix, cluster 9 has no digits in it. 257 of the true label 9 was classified to be a 3, and 97 was classified to be a 1. Otherwise, most of the other digits were classified correctly. 3089 digits were classified correctly.

```
Accuracy: 80.8004 %
```



predicted label \ true label	0	1	2	3	4	5	6	7	8	9
0	373	0	0	0	0	0	0	0	0	0
1	1	363	0	6	35	6	3	11	37	97
2	0	15	329	5	0	0	0	0	0	0
3	0	9	19	345	0	70	0	0	7	257
4	1	0	0	0	307	0	1	0	0	0
5	0	1	0	5	7	298	0	0	3	2
6	1	1	1	0	4	2	373	0	4	0
7	0	0	5	11	29	0	0	373	1	24
8	0	0	26	17	5	0	0	3	328	2
9	0	0	0	0	0	0	0	0	0	0

Figure 3: Performance of K-means on training when init=’k-means++’

3. After fitting the model with initialization set to derived cluster center variable, the result from predicting with the training label gave an accuracy of 87.8368 % as shown below. From the number of digits predicted per cluster, we found out that 323 digit 9 were now classified correctly according to the true label as shown in the confusion matrix in figure 4. The worst prediction this time was for digit 5, with only 265 classified correctly and 105 classified as a 9. Otherwise, most of the other digits were classified correctly. 3358 digits were classified correctly.

```
Accuracy: 87.8368 %
```

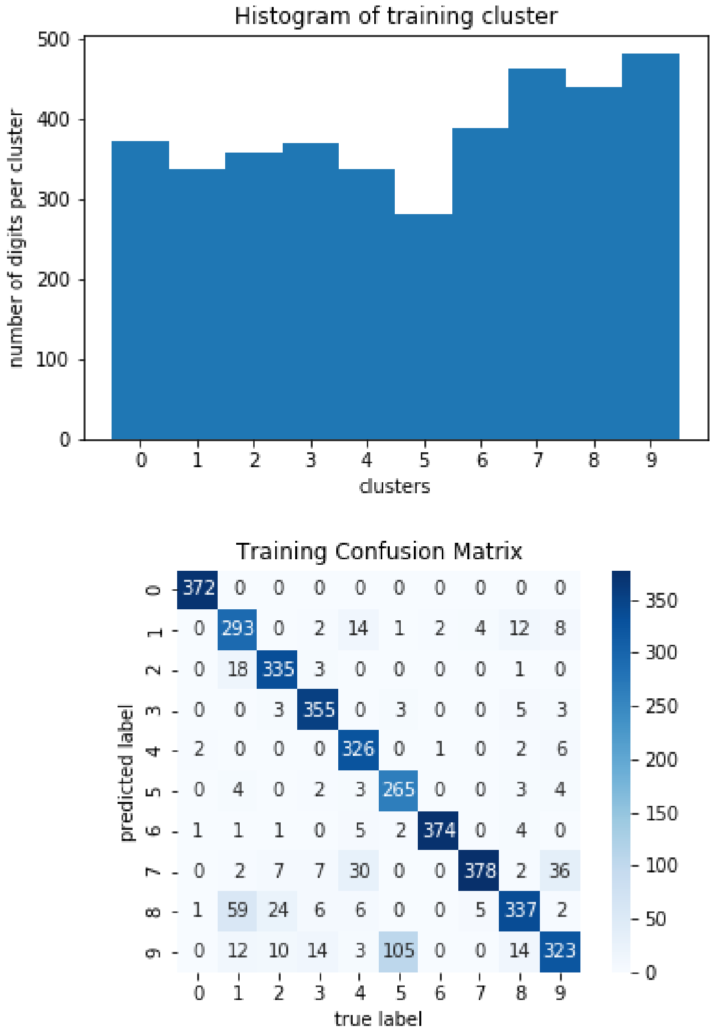


Figure 4: Performance of K-means on training when init = derived cluster center

4. The plot of the silhouette for different values of number of cluster from 10 to 25, the plot is as shown below in figure 5. It can be observed that the best silhouette score was for K = 10. Which means 10 is best number of clusters, which is in agreement with our intuition of the fact that the training data contain 10 digits 0-9. The silhouette score for 10 is 0.1915

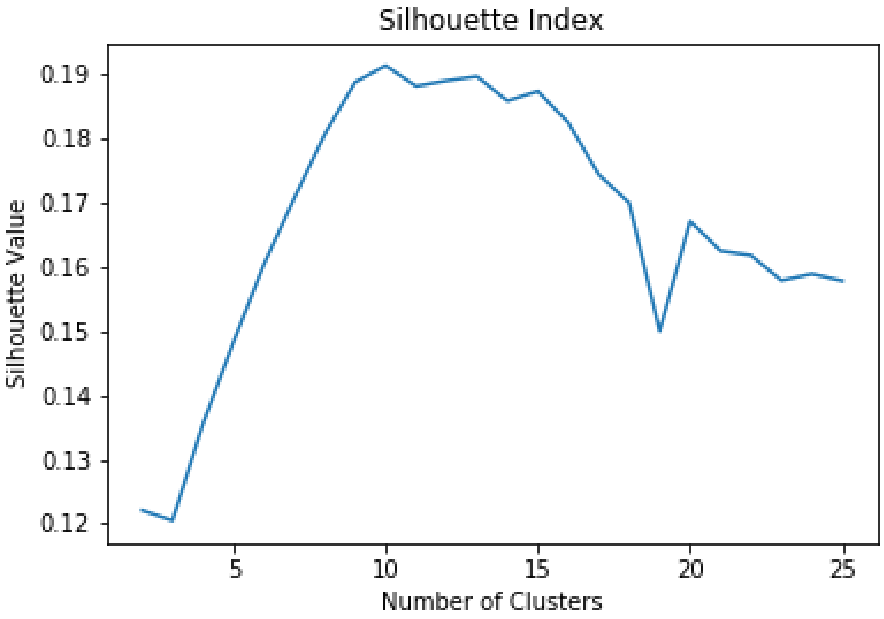
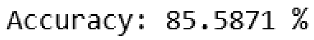


Figure 5: Silhouette score for different Number of Clusters

5. The test was carried out by carrying out prediction on the test data and an accuracy of 85.5871% was realized. The confusion matrix is shown below in figure 6, It can be seen that most of the digits were well classified, except for 5, which has 42 of its samples classified as 9 and 1 had 33 of its digit classified as 8 while 12 8's were classified as 1. Overall, 1538 out of 1797 test digits were classified correctly. From the histogram, most digits were wrongly classified as 9 and digit 1 was the most wrongly classified



```
Accuracy: 85.5871 %
```

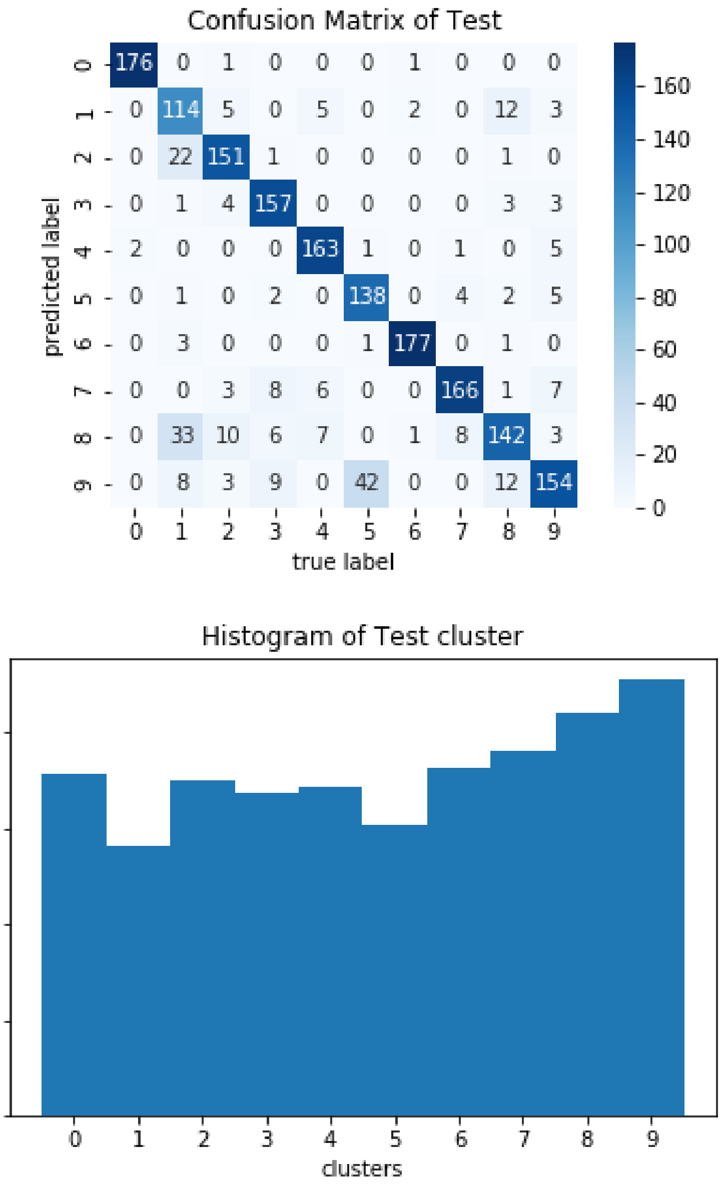
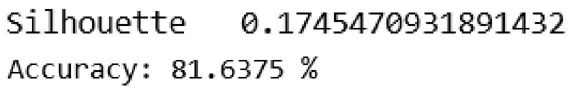


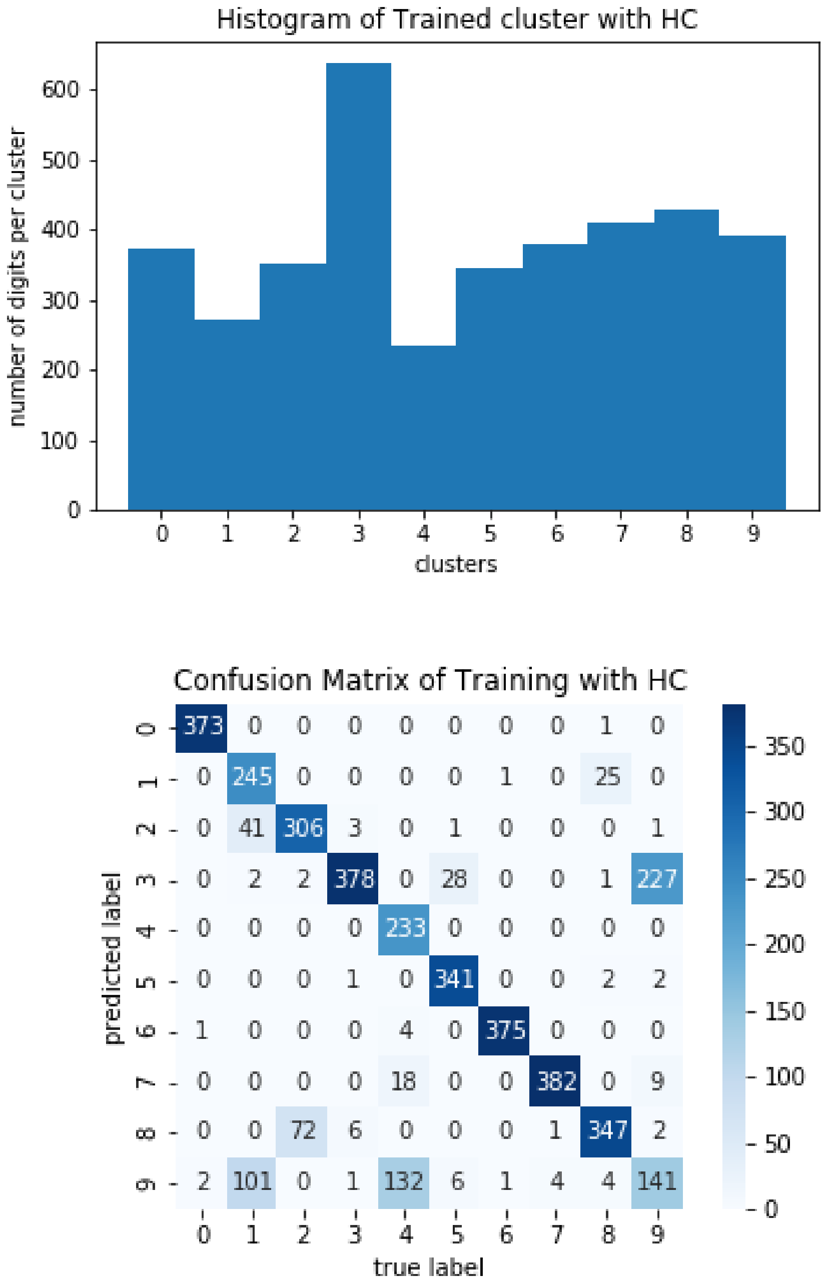
Figure 6: Performance of Kmeans Model on Test Data

6. The silhouette score gotten for this hierarchical clustering training process at number of clusters K = 10 is 0.1745 which is lower than that of K-means for the same number of clusters. This means that clustering at K = 10 is more suitable with K-means than with Hierarchical clustering. The accuracy on training computed is 81.6375%, lesser



```
Silhouette   0.1745470931891432
Accuracy: 81.6375 %
```

than the best case with K-means which was 87.8% at best. From the confusion histogram in figure 8, it can be seen that all the digits were represented in the cluster, meanwhile so many digits were wrongly predicted as 3. 227 samples of 9 was confused for 3 as shown in the confusion matrix and 132 4s and 101 1s were confused as 9. The total number of digits samples predicted correctly was 3121 digits.



	0	1	2	3	4	5	6	7	8	9
0	373	0	0	0	0	0	0	0	1	0
1	0	245	0	0	0	0	1	0	25	0
2	0	41	306	3	0	1	0	0	0	1
3	0	2	2	378	0	28	0	0	1	227
4	0	0	0	0	233	0	0	0	0	0
5	0	0	0	1	0	341	0	0	2	2
6	1	0	0	0	4	0	375	0	0	0
7	0	0	0	0	18	0	0	382	0	9
8	0	0	72	6	0	0	0	1	347	2
9	2	101	0	1	132	6	1	4	4	141

Figure 8: Performance of Hierarchical Clustering on Training

7. The Dendrogram at k =10 is shown below in figure 7

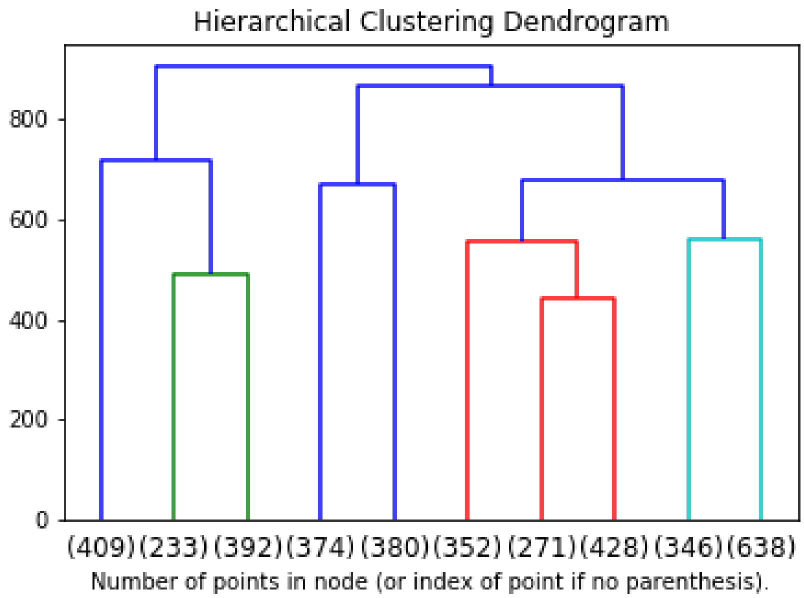


Figure 9: Dendrogram of Hierarchical Clustering

8. After the model was use to carry out prediction base on the test data, it performs quite bad with an accuracy of 68.3918%. Only 1229 digits out of 1797 digits were classified correctly. From the confusion matrix in figure 10, it is obviously that the algorithm

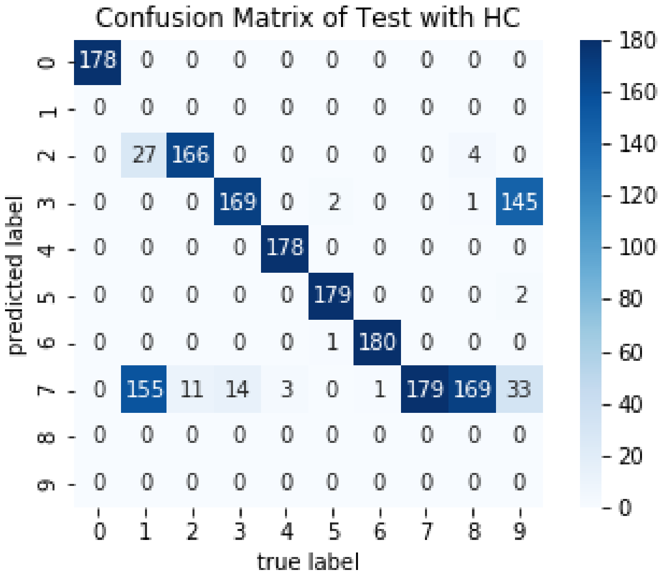


Figure 10: Confusion Matrix of Hierarchical Clustering on Test Data

could not differentiate digit 8 from 7, it predicted 169 8s as 7 and it also had issue differentiating 9 and 3, it predicted 145 9s as 3.

9. To confirm that the best number of cluster is at K = 10, the silhouette below in figure 11 provides the proof. It can be seen that the peak of the plot was when Number of Clusters is 10.

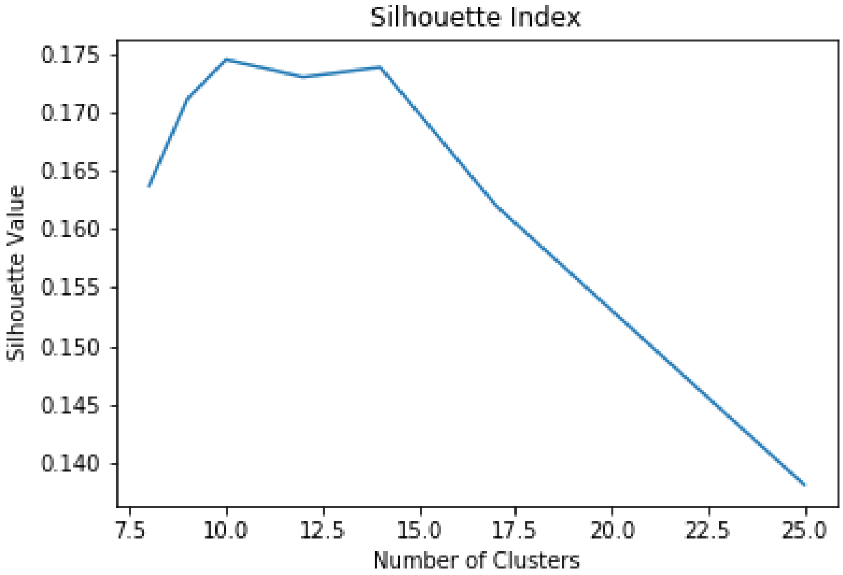


Figure 10: Silhouette score for different Number of Clusters for HC

1.6 Conclusion and Comparison of K-means and HC

From the practical procedure, we observed that for hierarchical clustering we do not need to set an initialization, but for K-means we needed to set an initialization. The Sklearn API allowed us to automatically set the initialization using value 'kmeans++' which varies the initialization centroid over set number of iterations = 10 and pick the best in terms of lowest inertia (quantization error J). We could also set init='random', which randomly pick the initialization centroid, using this we obtain the best performance of 87% training accuracy. The initialization centroid at this point was saved in a csv file and was used to train a more accurate model. Using the silhouette score for varying number of clusters; k, we obtained the highest silhouette score for K = 10, both when we use hierarchical clustering and K-means clustering techniques. Although, the value for silhouette score for K-means clustering was larger 0.1915 for number of cluster = 10 than that of hierarchical clustering which was 0.1745. This could mean that the K-means clustering would have a better performance on the data set. In fact, this was confirmed from the accuracy obtained on the test data for both algorithm. For K-means we got an accuracy of 85.58% compare to 68.3918% for Hierarchical clustering. The developed K-means model was able to correctly predict 1538 out of 1797 test digits, while the hierarchical clustering model was only able to predict 1229 digits out of 1797 test digits.

From our observation, we can conclude that K-means performed better on this data set than hierarchical clustering.

From the practical, we were able to get a better understanding of unsupervised learning using k-means and hierarchical clustering, and we were able to appreciate their performance.