K-Means Clustering

Question 2.1

K-means clustering algorithm represents the number of clusters by k-means. This algorithm works well in classification in the task of hand-written digit recognition. This is because k-means algorithm allows to assign data points to a cluster based on the sum of the square distance between each point within the clusters. The k-means works by:

1. Loading the data
2. Selecting number of clusters in a given dataset or choosing k from the elbow method.
3. Initializing centroids by randomly selecting distinct data points for each cluster.
4. Keep iterating following steps shown below until there is no change to the centroids. For example, assigning data points to clusters isn't changing. This will be the optimal centroid when:
    a. Calculating the distance between each data point from each of the randomly assigned k points.
    b. Categorising each data points to its closest k points.
    c. Compute centroid for the cluster by taking the average of data points that are categorised in that k points.

Question 2.2

Question 2.2.1

The k-means clustering for handwritten digits program is located in a source code file called k_means.py.

Question 2.2.2

The k-means clustering for handwritten digits program is located in a source code file called k_means.py. Also, I implemented my own version of k-means for this problem.

Question 2.2.3

The dataset contains 1,797 samples and 64 features. Each digit is stored in an image with 8x8 pixels, and the 64 features represent the level of each pixel brightness in an image.

The Elbow method gives me the best optimal k value to run k-means clustering on the dataset as the k values are represented in a scatter graph in Figure 1. The Elbow method set in a range of k values from 1 to 20 and each k value is calculated by the sum of squared errors (SSE). Then, I plot a line chart of the SSE for each k value. If the line chart looks like an arm, then the "elbow" on the arm is the best optimal k value. Looking at Figure 1 shows the best k value rely on the range between the minimum k = 4 and maximum k = 10.

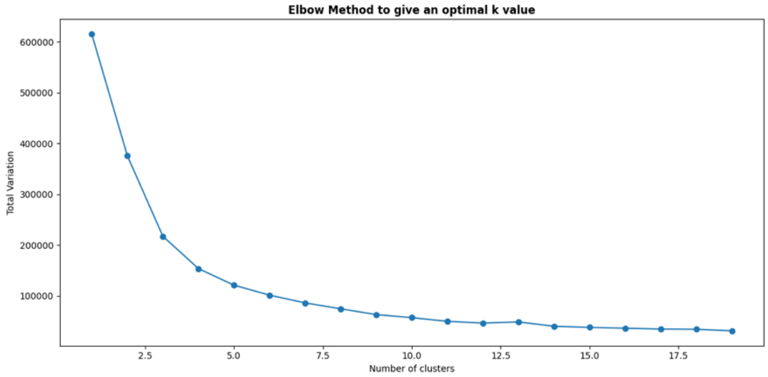


Figure 1: Elbow Method

I set k = 4 to run the k-means clustering and plotting in the PCA graph. The graphs shows that it works well as all the data points are very clustered (it as shown in Figure 2.1). However, the elbow method doesn't always work well especially, when representing the clustered data on the TSNE graph method. This is because the data is not clearly clustered as shown in Figure 2.2. This means k = 4 is the best optimal k value for the PCA graph method.

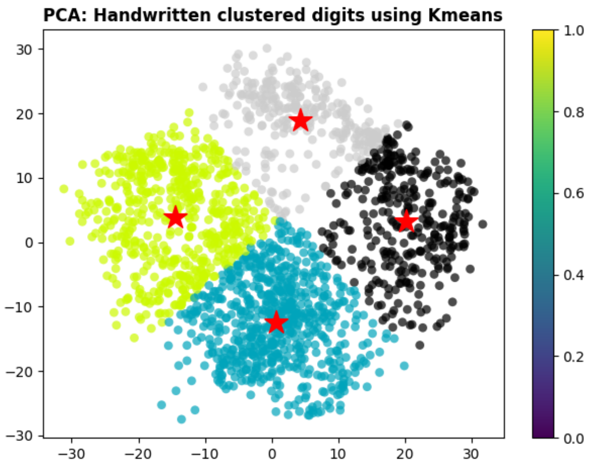


Figure 2.1: PCA graph when k = 4

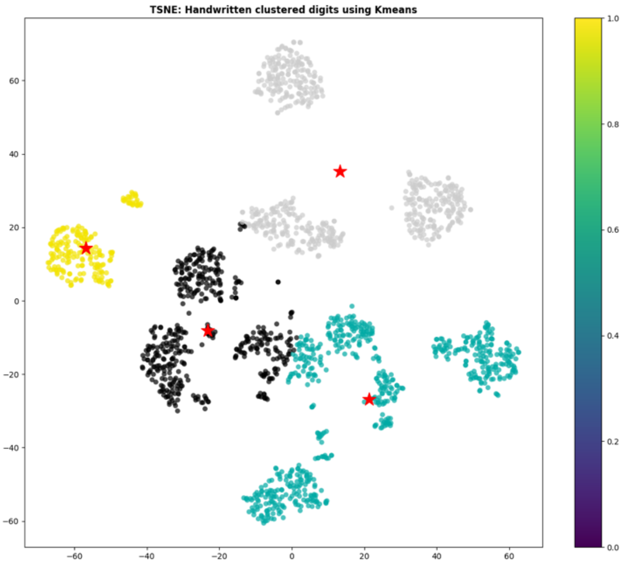


Figure 2.2: TSNE graph when k = 4

To improve the clustering, I set the k = 10 and plotting in the PCA graph method and TSNE graph method as shown in Figure 3.1 and Figure 3.2. This shows that TSNE works very well as all the data points are very clustered and it doesn’t clash with another cluster. Therefore, k = 10 is the best optimal k value for the TSNE graph method.

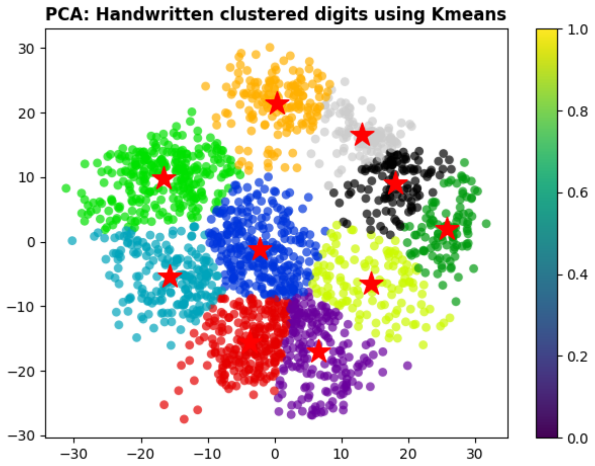


Figure 3.1: PCA graph when k = 10

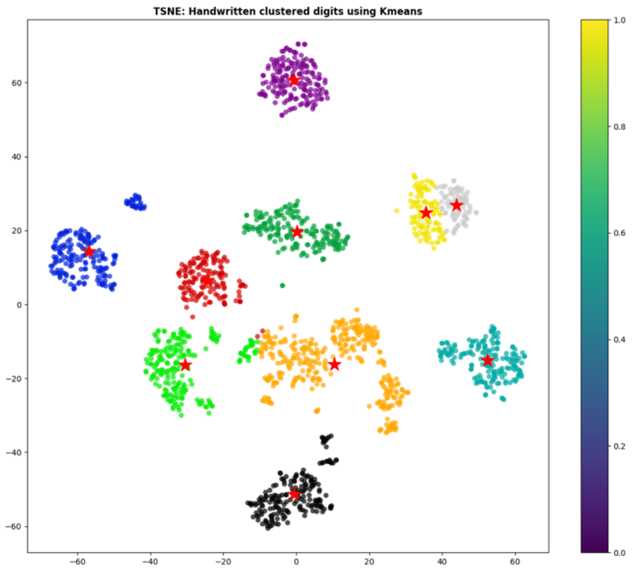
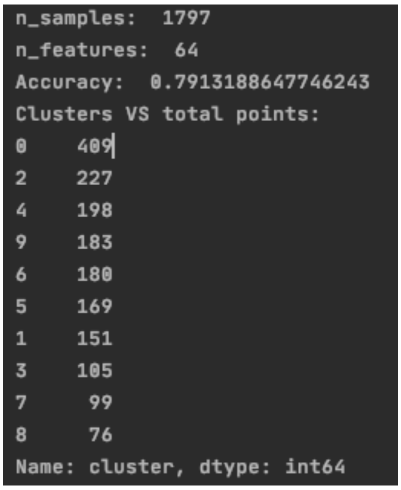


Figure 3.2: TSNE graph when k = 10

The reason for presenting two graphs is that with TSNE graph method makes the data in the graph looks better since PCA has a linear constraint and doesn’t work well with large dataset as an example shown in the Figure 3.2. However, PCA graph method shows dimensionality reduction and projecting data onto its orthogonal feature subspace as an example shown in the Figure 2.1.

After creating k-means algorithm to cluster the data, I analyse the data by using Elbow method to give an optimal k value, using PCA graph method and TSNE graph method to show the clustered data in the scatter graph and finding the accuracy in the similarity of the digits within the data.

Now to check how accurate my unsupervised clustering is, when finding similarity of the digits within the data.



```
n_samples:  1797
n_features:  64
Accuracy:  0.7913188647746243
Clusters VS total points:
0    409
2    227
4    198
9    183
6    180
5    169
1    151
3    105
7     99
8     76
Name: cluster, dtype: int64
```

The Figure 2.2.6 shows around 80% accuracy in grouping data of the input data. This means the input digits match the learned cluster labels with the true labels found within them.

Figure 4: Shows the output of the number of samples, number of features, accuracy and a table of total point in each cluster.

Finally, I set the k = 10, which are images of 8x8 pixels with the colour intensity for each pixel. The Figure 5 shows their appearance of the digit images by clusters centroids learned by k-means. This shows that even without the labels, k-means is able to find clusters and recognise digits from 0 to 9. This process done by matching each learned cluster label with the true label found in them.

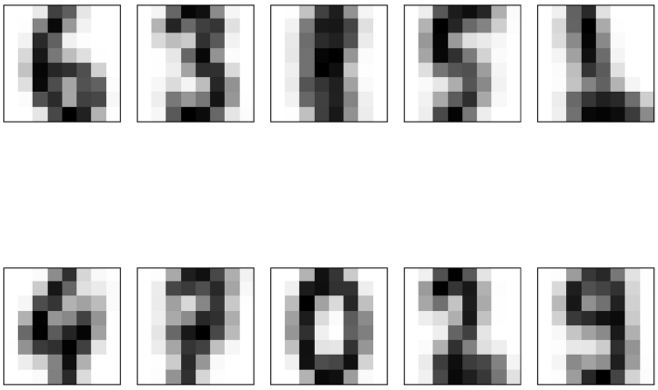


Figure 5: 10 clusters centroids images with 8 x 8 pixels.

Overall, this shows unsupervised learning is more accuracy, when extract information from the dataset.

Question 2.3

The limitations of k-means clustering are difficult to determine the number of clusters (k values), especially when we use the Elbow method in selecting number of clusters. Below, I have shown an example of the elbow method. The Elbow method set in a range of k values from 1 to 20 and each k value is calculated by the sum of squared errors (SSE). Then, plotting a line chart of the SSE for each k value. If the line chart looks like an arm, then the "elbow(curve)" on the arm is the best optimal k value.

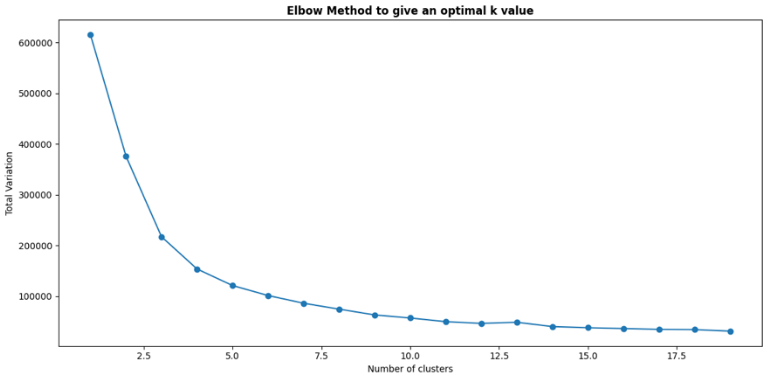


Figure 1: Elbow Method

Looking at Figure 1 shows the best k value rely on the range between the minimum k = 4 and maximum k = 10. According to these observations, it's possible to define k as 10 as the optimal value because the level of grouping data accuracy is higher than having k as 4 as it shown in Figure 2.1 and Figure 3.1.

This means the elbow method doesn't usually work because it measures a global clustering characteristic only to estimate the optimal number of clusters.

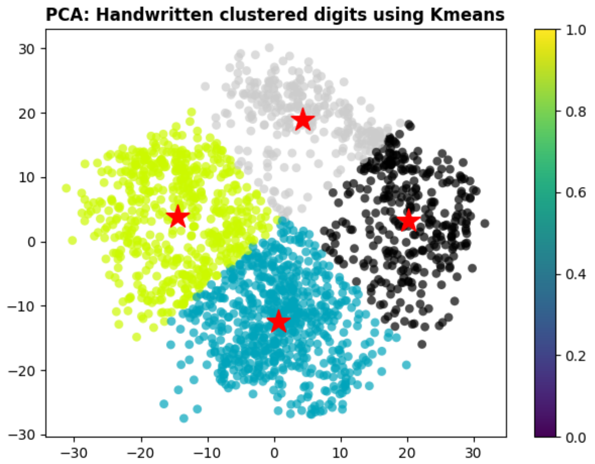


Figure 2.1: PCA graph when k = 4

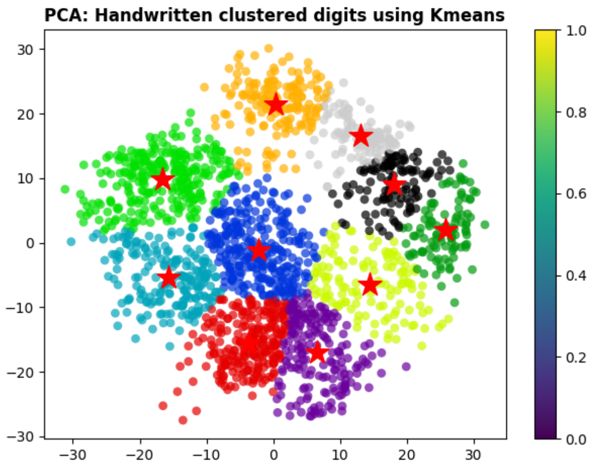


Figure 3.1: PCA graph when k = 10

Another one is assuming that all clusters are equally sized and have the same variances as this doesn’t work well all the time. This is because clusters will differ in their size, density and variance as it’s hard to make an to use k-means. Below, I have shown an example dataset clustering with no clustering and with clustering.

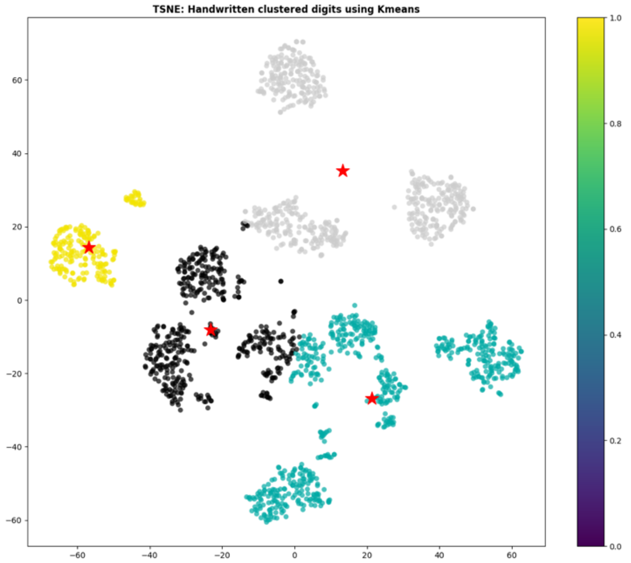


Figure 2.2: TSNE graph when k = 4

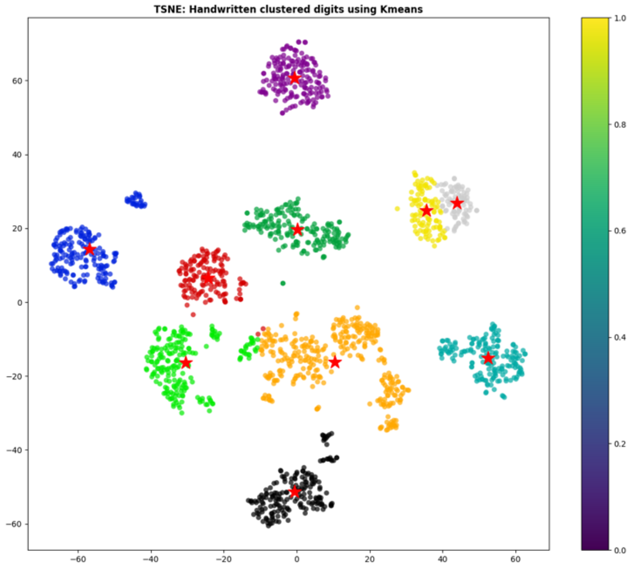


Figure 3.2: TSNE graph when k = 10

The Figure 2.2 and Figure 3.2 shows an exemplary clustering of the digit data set using k-means and no clustering. We can clearly see the downside of k-means clustering, where if the clusters are defined by the k-means like k = 4 is not able represent in a graph very well as the clusters differ in their size, density and variance. For example, the Figure 3.2 shows the overlapping between clusters to determine, which cluster to assign each data point. This figure shows k-means doesn’t have an intrinsic measure for uncertainty and splits the data incorrectly.