Description:

This experiment is about implementing the support vector machine K-mean clustering algorithm. In this algorithm, training and testing are done on 'Optdigits' dataset. The optdigits dataset contains 3823 training example and 1797 testing examples digits. Each sample has 64 features and one class label. The K-means algorithm starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs repetitive calculations to optimize the positions of the centroids in training set. After it optimizes correctly, we calculate the average mean square error, mean square separation and mean entropy for the training set. For the minimum mean square error, we use the centroid to test the test set. The predicted clustering is shown using the visualization. The results are shown for each experiment below.

Experiment [1]:

In this experiment, after reading the dataset from the file, the data is divided into attributes and labels. First, we select the k=10 random centroid from the training set and run the K-mean clustering algorithm to optimize the clustering. We run these steps for the 5 times and calculate the average mean square error each time and then select the centroid with the minimum average mean square error to evaluate the testing set. For that, we associate each cluster center with the most frequent class it contains in the training data and then assigns each test instance the class of the closest cluster center. At the end, we get the 10 clusters. Using the maploatlib library, we visualize the cluster center on the 8x8 grid.

Results:

- Training data:
  Average mean square error : 608.6003885226667
  Mean square sepration : 1292.5556113504422
  Mean entropy : 0.1824759762177

- Testing data:

  Accuracy : 0.731669449081803
  Confusion Matrix:

```
Confusion matrix :
[[ 99   0   0   0   0   0   0   0   0   2]
 [  0 165   0   0   0   1   1  60  21   0]
 [  5   5  84   0   1   7   7   2   0   0]
 [  0   9   3   0   0   0  10   0   0   0]
 [  6   5   2   0 176  0   5   3   0 160]
 [  1   0   0   0   0 161  79   1   0   1]
 [  3   0   1   0   1   0   1   0   0   0]
 [  1   0   8   0   0   0   4 175   0   1]
 [ 28   1  10   0   0  18 106  10 150  0]
 [  0   2  34   0   0 130   5   9   0   0]]
```

Visualization:
Cluster_class: [1.0, 0.0, 6.0, 3.0, 8.0, 7.0, 5.0, 7.0, 2.0, 9.0]
Predicted cluster class:

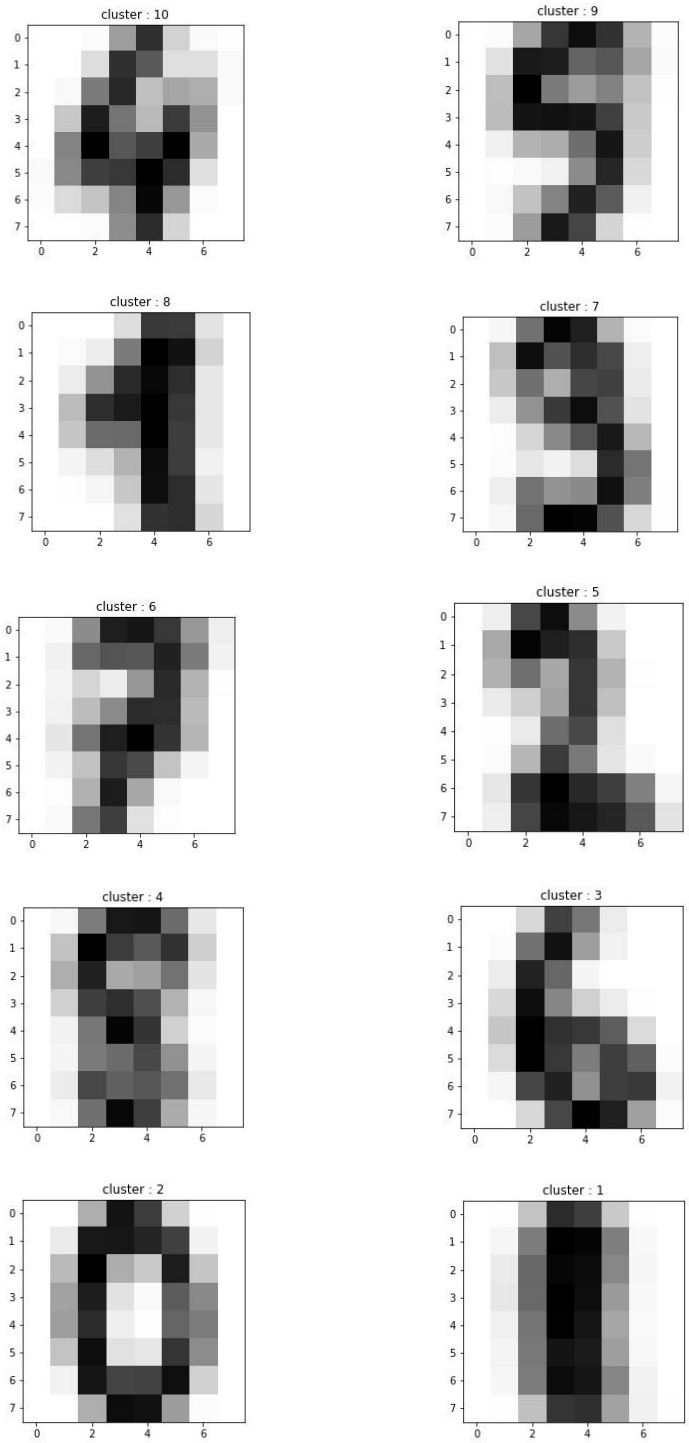


Figure 1: predicted cluster class

From the confusion matrix, we can observe that digit 9 is more confused with the digit 4 and more predicted as digit 4. And digit 0 and digit 1 is more accurately predicted by the algorithm. We get the 73.16% accuracy using the centroid with minimum mean square error (608.6 for this experiment). From visualization, we can observe that digit 1,0, 6 and 7 looks like their associated

digit. Where digit 9 is associated as digit 4 and digit 5 is associated as digit 3. From figure 1, we can observe that digit 8 is not visualized.

Experiment [2]:

In this experiment, we select the k=30 random centroid first and run the K-mean clustering algorithm to optimize the clustering. We run these steps for the 5 times and calculate the average mean square error each time and then select the centroid with the minimum average mean square error to evaluate the testing set. The results are shown below.

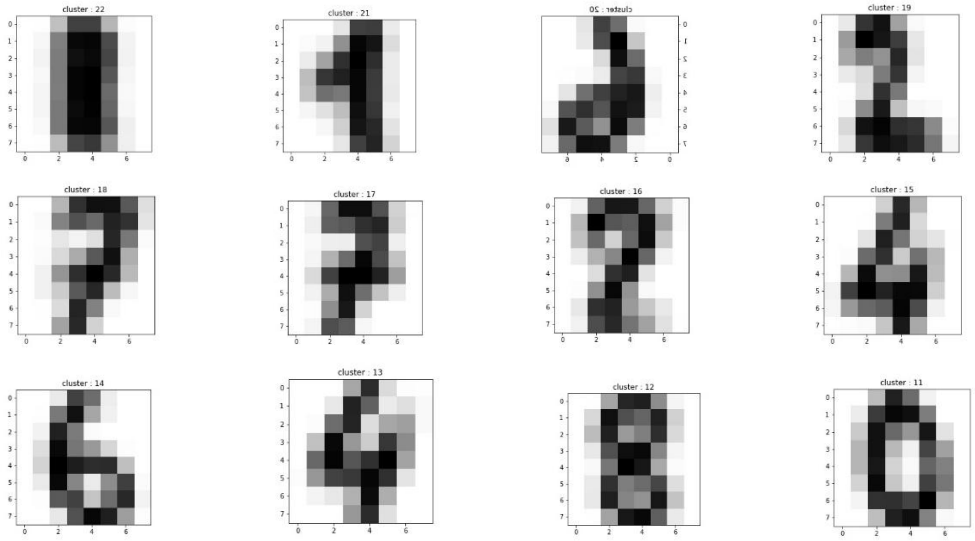
Results:

- Training data:
  Average mean square error : 455.9687530888267
  Mean square sepration : 1515.837901132602
  Mean entropy : 0.493611053732106
- Testing data:

  Accuracy : 0.89453121
  Confusion Matrix:

```
Confusion matrix :
[[151   0   0   0   0   0   0   0   0 178]
 [ 70 121   0   0   3   0  26   0   0   0]
 [  1   2  64   0   0   1   4   0  19  87]
 [  2   1   1  56   0   4  23  27   5   1]
 [ 30   3  49   0  98   7   0   0   3  95]
 [  0   0   1   0  18   0   7   1   1   3]
 [ 35   0  72   0  71  55  92   1   0   2]
 [  0   7   1   0   1   0   3 111  98   0]
 [  6   8   1   2  41   1  13   0  75   8]
 [  1   0   0   1   3   0  34   2   0  37]]
```

Visulization:
Cluster_class: [8.0, 0.0, 3.0, 4.0, 9.0, 4.0, 2.0, 9.0, 1.0, 0.0, 9.0, 7.0, 9.0, 2.0, 0.0, 8.0, 8.0, 6.0, 9.0, 4.0, 1.0, 6.0, 6.0, 1.0, 2.0, 7.0, 0.0, 4.0, 9.0, 0.0]
Predicted cluster class:



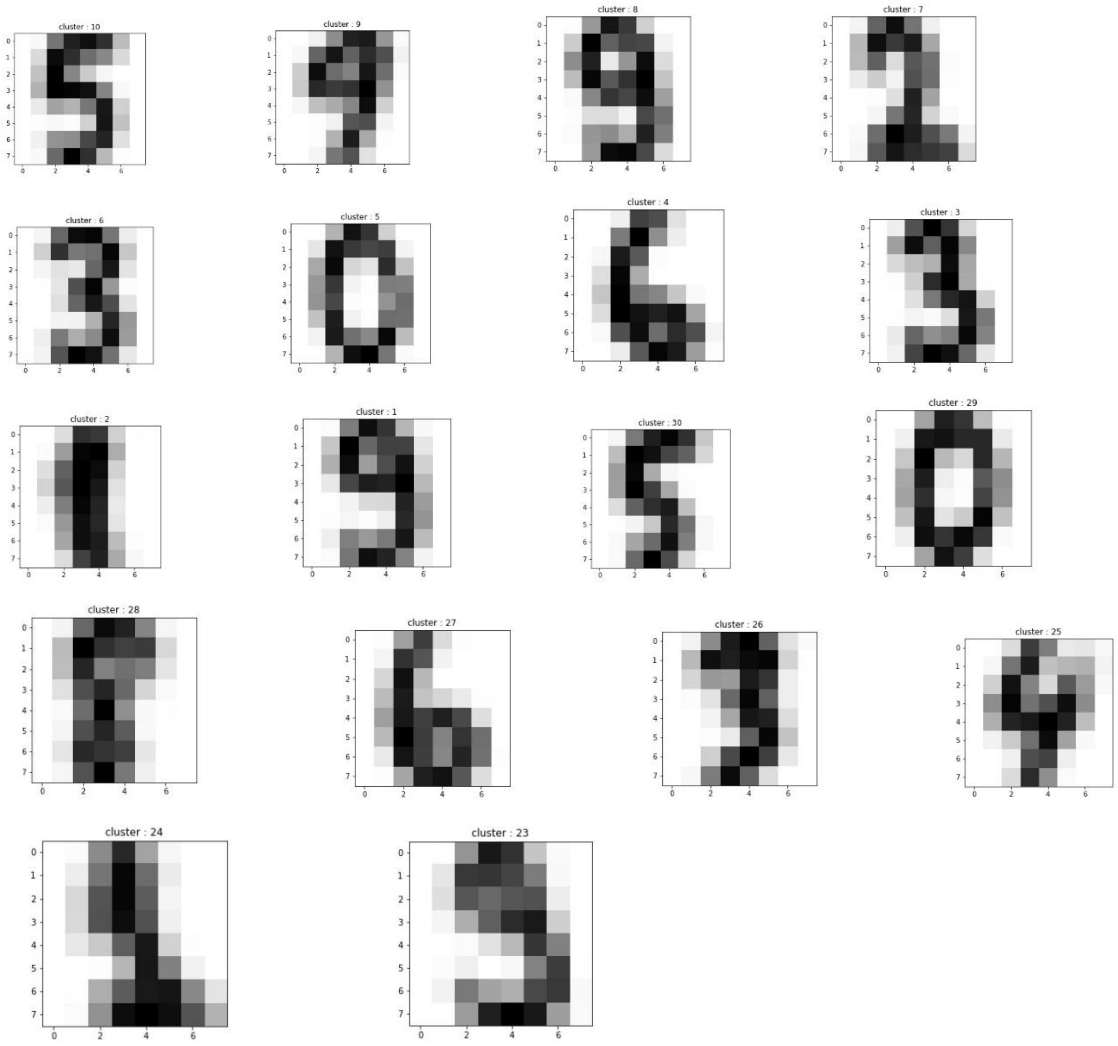


Figure 2: predicted cluster class

From the confusion matrix, we can observe that digit 9 is more confused with the digit 4 and more predicted as digit 4. And digit 0, 1 and digit 7 is more accurately predicted by the algorithm. We get the 89.46% accuracy using the centroid with minimum mean square error (455.6 for this experiment) which better than the k=10 centroid. From visualization, we can observe that digit 1,0, 6 and 7 looks like their associated digit. Where digit 9 is associated as digit 4 and digit 5 is associated as digit 3. In this experiment, the average mean square error is less compare to the previous experiment and the mean entropy is 0.4. Here the 10 digits are divided into the 30 clusters because that it is predicting better than the k= 10 centroids. For K=30, compare to the previous experiment all the digits are predicted and visualized, wherein a previous experiment some of the digits like 8 was not visualized. From figure 2, we can observe that it is visualizing all the digits.