Hw 5 Report by Akanksha

Experiment 1

Description:

In this experiment we have done k means clustering for optdigits data set with number of clusters centers (k=10). We have done this for 5 runs on the training data and out of it whichever is having minimum error we are taking that as best run and using it for testing purposes.

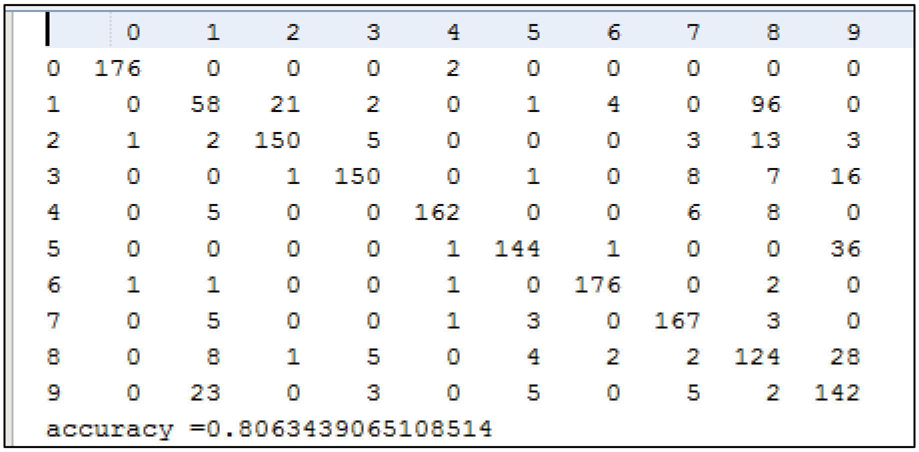
Average Mean square error-

amse = 649.483366

Mean square separation-

MSS = 1258.134747

Accuracy and Confusion Matrix-



	0	1	2	3	4	5	6	7	8	9
0	176	0	0	0	2	0	0	0	0	0
1	0	58	21	2	0	1	4	0	96	0
2	1	2	150	5	0	0	0	3	13	3
3	0	0	1	150	0	1	0	8	7	16
4	0	5	0	0	162	0	0	6	8	0
5	0	0	0	0	1	144	1	0	0	36
6	1	1	0	0	1	0	176	0	2	0
7	0	5	0	0	1	3	0	167	3	0
8	0	8	1	5	0	4	2	2	124	28
9	0	23	0	3	0	5	0	5	2	142

```
accuracy =0.8063439065108514
```

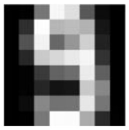
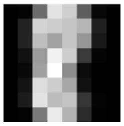
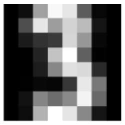
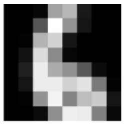
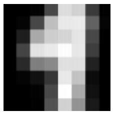
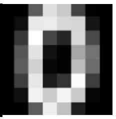
Visualization Results-

Files in the attached k_10 folder.

The pictures are-

Cluster 0-   cluster 1- cluster 2- cluster 3- cluster 4- cluster 5- cluster 6- cluster 7- cluster 8- cluster 9-

label 9-   label 8-   label 2   label 3-   label 7-   label 6-   label 1-   label 5-   label 0   label 4

Summarize results-

The kmeans clustering algorithm has performed average with the less number of clusters as we can see from the results there is a huge difference between the two accuracies around 10-12 %. Also the error is more in this case. For few of the visualization we can see resemblance with the label for which it is written into a grid.

Do the visualized cluster centers look like their associated digits?

Few of the visualizations resembles the associated digits but the grid is not very symmetric with the numbers at each position so it is very difficult to judge the visualization sometimes but mostly u can figure out the digits.

<u>Experiment 2</u>

Description:

In this experiment we have done k means clustering for optdigits data set with number of clusters centers (k=30).We have done this for 5 runs on the training data and out of it whichever is having minimum error we are taking that as best run and using it for testing purposes.

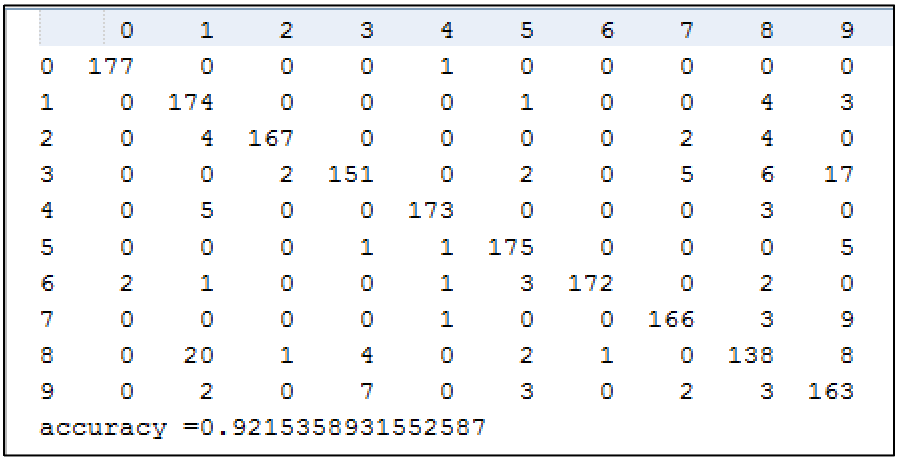
Average Mean square error-

mse = 475.328805

Mean square separation-

MSS = 1510.720289

Accuracy and Confusion Matrix-



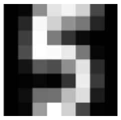
```
       0    1    2    3    4    5    6    7    8    9
0    177    0    0    0    1    0    0    0    0    0
1      0  174    0    0    0    1    0    0    4    3
2      0    4  167    0    0    0    0    2    4    0
3      0    0    2  151    0    2    0    5    6   17
4      0    5    0    0  173    0    0    0    3    0
5      0    0    0    1    1  175    0    0    0    5
6      2    1    0    0    1    3  172    0    2    0
7      0    0    0    0    1    0    0  166    3    9
8      0   20    1    4    0    2    1    0  138    8
9      0    2    0    7    0    3    0    2    3  163
accuracy =0.9215358931552587
```

Visualization Results-

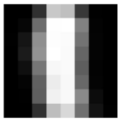
Files in the attached k_30 folder.

The Pictures-

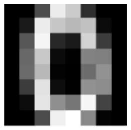
Cluster 0(Label 5)



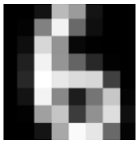
Cluster 1(Label 1)



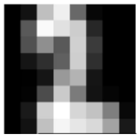
Cluster 2(Label 0)



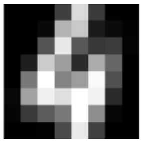
Cluster 3(Label 6)



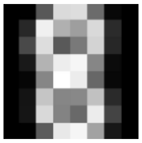
Cluster 4(Label 2)



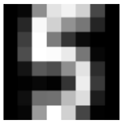
Cluster 5(Label 4)



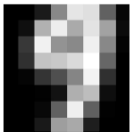
Cluster 6(Label 8)



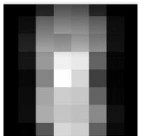
Cluster 7(Label 5)



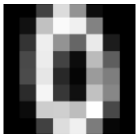
Cluster 8(Label 9)



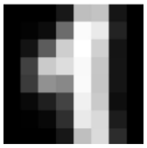
Cluster 9(Label 1)



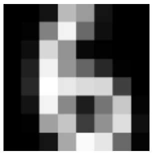
Cluster 10(Label 0)



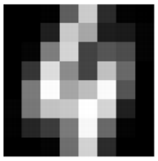
Cluster 11(Label 1)



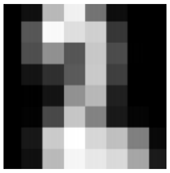
Cluster 12(Label 6)



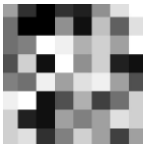
Cluster 13(Label 4)



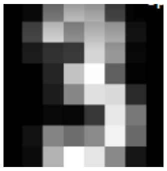
Cluster 14(Label 2)



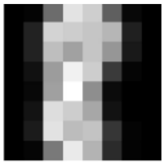
Cluster 15(Label 9)



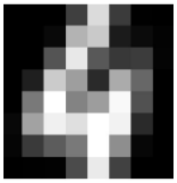
Cluster 14(Label 3)



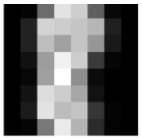
Cluster 15(Label 8)



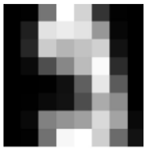
Cluster 16(Label 4)



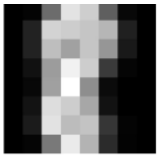
Cluster 17(Label 0)



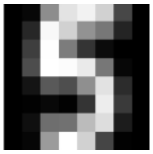
Cluster 18(Label 4)



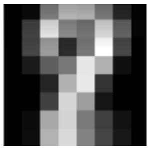
Cluster 19(Label 0)



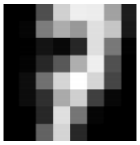
Cluster 20(Label 5)



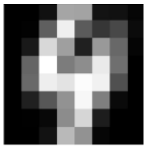
Cluster 21(Label 2)



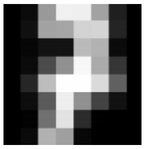
Cluster 22(Label 3)



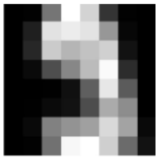
Cluster 23(Label 4)



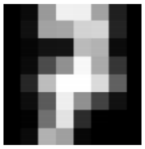
Cluster 24(Label 7)



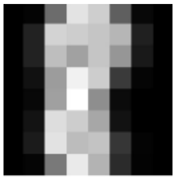
Cluster 25(Label 9)



Cluster 26(Label 7)

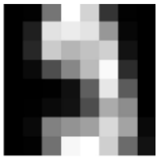


Cluster 27(Label 1)



Cluster 28(Label 4)

Cluster 29(Label 3)



Summarize results-

The kmeans clustering algorithm has performed better with the more number of clusters as we can see from the results there is a huge difference between the two accuracies around 10-12 %. Also the error has decreased with the increase in number of clusters. I still have some doubt on how to compare visualization with the actual digit but for few of them we can see resemblance with the label for which it is written into a grid.

Compare the results of Experiments 1 and 2.

As the number of the clusters increases the accuracy of the digit classification increased also the average mean square error has reduced where as the distance between two centroids has increased. This can be seen from the obtained results, the accuracy for k=10 is around 80 % and with k=30 is 92% which is very much high comparatively, similarly the average mean square for k=10 is around 650 and for k=30 it is around 450.So we can conclude from these results that this algorithm has performed better with more number of clusters.

Do the visualized cluster centers look like their associated digits?

Few of the visualizations resembles the associated digits but the grid is not very symmetric with the numbers at each position so it is very difficult to judge the visualization sometimes but for mostly u can figure out the digits.