Cover page for answers.pdf

CSE512 Spring 2021 - Machine Learning - Homework 5

Your Name: Venkata Ravi Teja Takkella

Solar ID: 113219890

NetID email address: venkataravite.takkella@stonybrook.edu

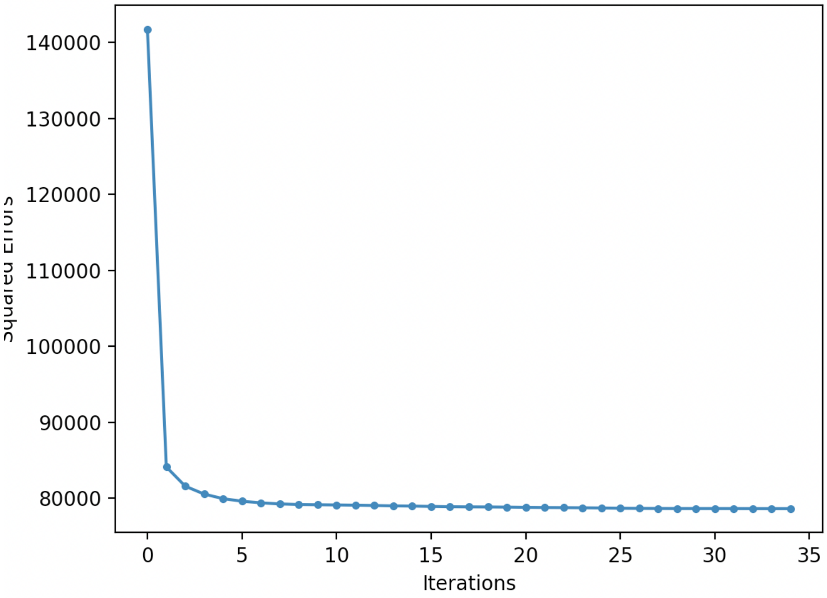
Names of people whom you discussed the homework with: NA

1)

1.2)

Here we were asked to run the k-means algorithm using k=10

a) The plot of squared errors against the number of iterations look as below



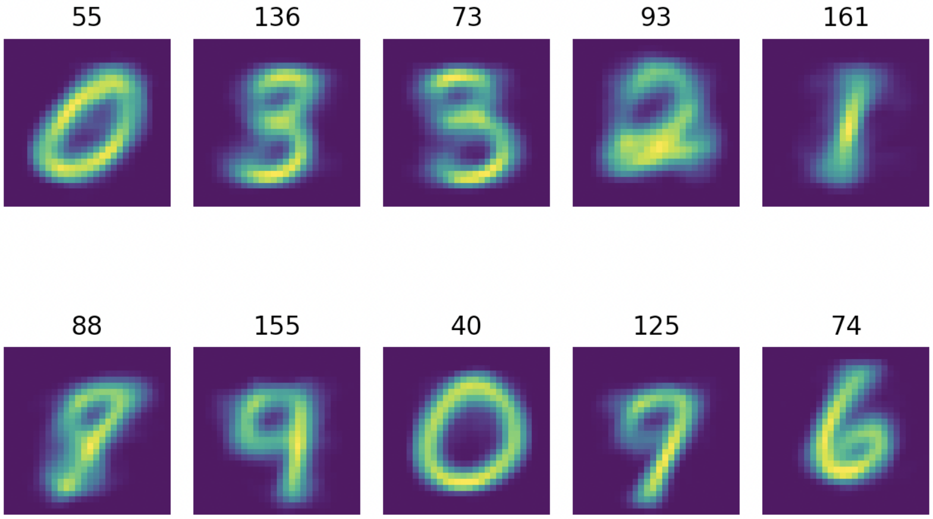
b) The final value of sum of squared errors obtained by the end is

78604.49194794556

1.3)

Here we are asked to run the k-means algorithm built on the test dataset.

a) The plot of all the centroids as images and number of test data points assigned to each centroid/class is shown below.



b) We can see from the images, they are a bit blurry. That’s because we are taking the average of all the points assigned to a centroid to get a new centroid. This point is not actually present in the dataset and is a mixture of few images. So each centroid doesn’t actually look like a specific number but a mixture of few numbers. This mixture might also be a mixture of similar numbers but of different orientation. This can also make the centroid look like a different number when the average is taken.

And we can also see, all the clusters are created with numbers that look similar. For example, the second plot looks like a mixture of 3 and 8 and the third plot looks again a mixture of 5,3,8. This is the array of all the unique classes of the testing data associated with each centroid. So the clustering does make a good sense over here and it also shows the assigned samples are actually looking at a part of the centroids at some specific orientation. They look similar with the assigned centroid.

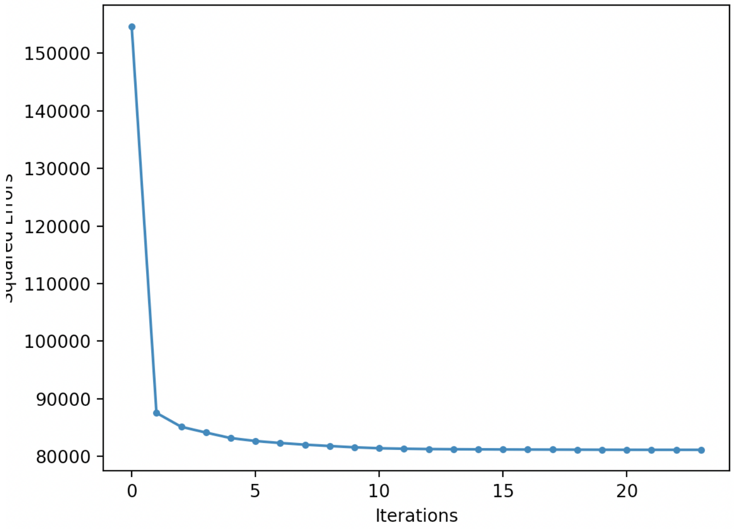
[[5, 0, 9, 8, 2, 3, 6], [3, 8, 5, 2, 6, 9, 0], [5, 0, 8, 3, 2, 7], [6, 2, 0, 4, 9, 7], [1, 2, 7, 8, 3, 5, 4, 9, 6], [5, 7, 4, 8, 0, 2, 9, 3, 1], [4, 9, 7, 8, 5, 2, 0], [0, 3, 6], [4, 9, 7, 5, 0, 8], [6, 4, 9, 0, 2, 5, 8]]

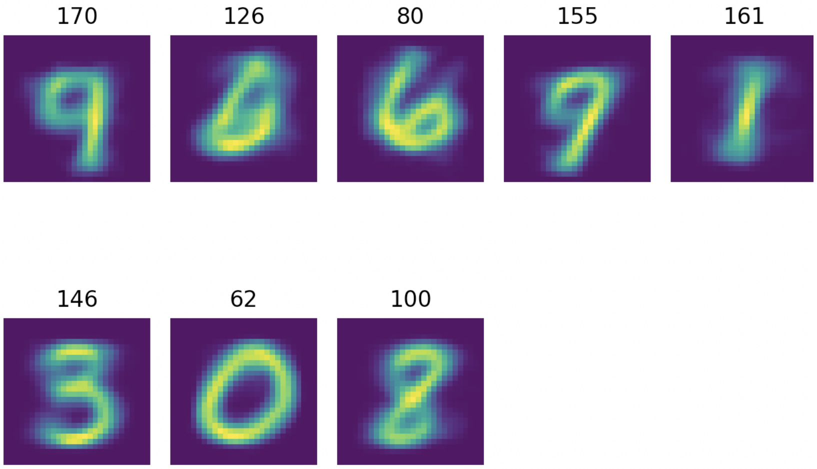
1.4)

k=8

When ran for 8 clusters, the final value of the sum of squared errors is **81449.30806242967** which is higher than what we observed for 10 clusters.

The squared errors against iterations looks like this.





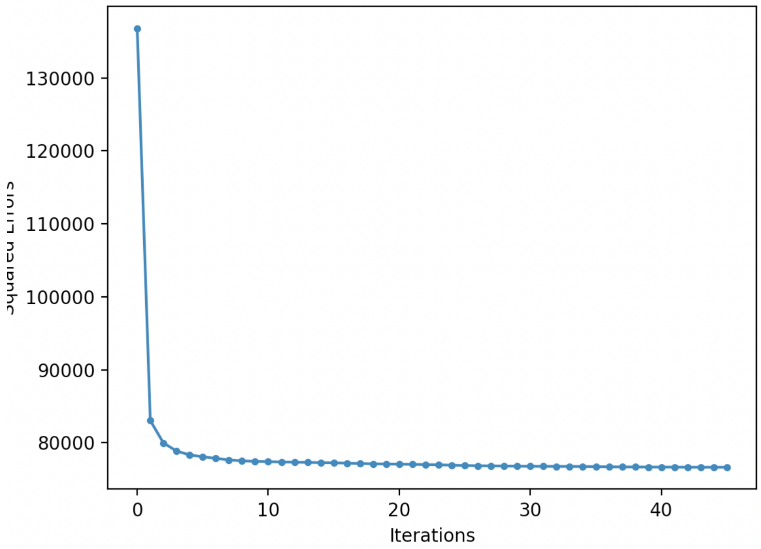
If we also observe, the number of unique centroid images look like decreased . They turned to be more blurry with a heavier mixture of different data points. But the running performance of the algorithm for each iteration is improved because the calculation of squared errors is only 8 centroids and polynomially lesser than when taken 10 centroids. But it also took higher time to converge to the optimum as there are more number of data points assigned to each centroid now. And even here, there are multiple data points labels clustered with different centroid labels too, but again it’s about similarity between the numbers, we can’t get an exact number label for each cluster. It is possible depending on the initialization of the centroid points.

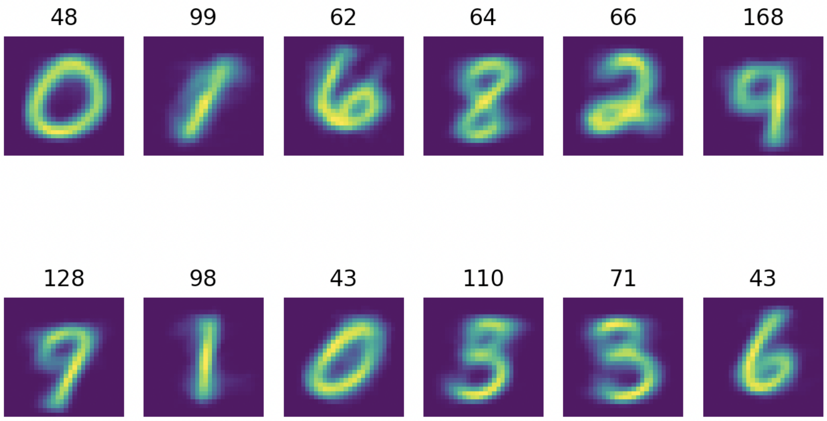
k=12

When ran for 12 clusters, the final value of the sum of squared errors is **76167.69266642367**

which is less than what we observed for 10 clusters.

The squared errors against iterations looks like this.





Here pictures turn out to be more clear compared to the 10 means clustering. And the number of data points assigned with each image also decreased.

The running performance of the algorithm for each iteration is more late because the calculation of squared errors is with 12 centroids and polynomially higher than when taken 10 centroids. And it also took less time to converge to the optimum as there are a lesser number of data points assigned to each centroid now. And even here, there are multiple data points labels clustered with different centroid labels too, but again it's about similarity between the numbers, we can't get an exact number label for each cluster. It is possible depending on the initialization of the centroid points.

2)

2.2) The number of clusters taken for each video is mentioned below

Cam_04_debug : 2

Cam_10_debug : 4

Cam_16_debug : 3

Cam_24_debug : 2

The Json files are submitted as mentioned in the assignment.

2.3) The number of clusters taken for each video is mentioned below

Cam_04_debug : 2 + outlier

Cam_10_debug : 4 + outlier

Cam_16_debug : 3 + outlier

Cam_24_debug : 2 + outlier

The Json files are submitted as mentioned in the assignment.