CSE512 Spring 2021 - Machine Learning - Homework 5

Your Name: Sagar Onkar Toshniwal

Solar ID: 113260061

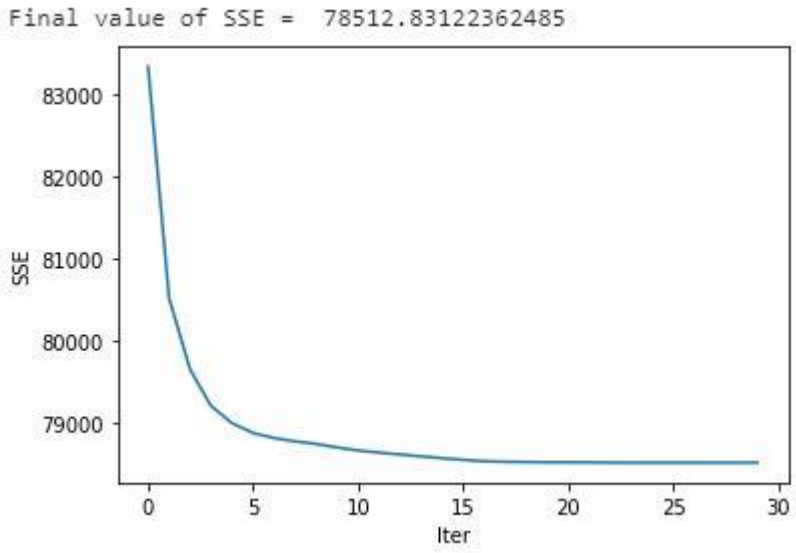
NetID email address: sagaronkar.toshniwal@stonybrook.edu

Names of people whom you discussed the homework with: None but I have taken reference from Stackoverflow, Medium,askpython, towardsdatascience, Google

1.2]

SSE vs Iteration plot for k = 10 (num_of_iterations = 30)

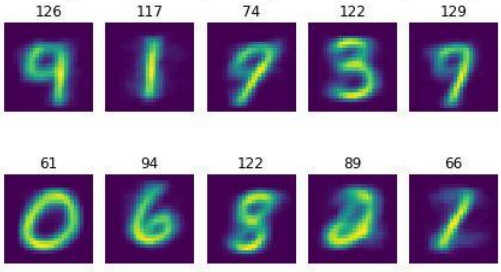
Final SSE = 78512.83



1.3] 10 centroids and their corresponding test points in that cluster

```
[88]: print (acc)
      fig = plt.figure()
      for i in range(10) :
          plt.subplot(2,5,i+1)
          plt.imshow(np.reshape(centroid[i], (28,28)))
          plt.title (acc[i])
          plt.axis('off')
      plt.tight_layout()
      plt.show()
```

```
Counter({4: 129, 0: 126, 3: 122, 7: 122, 1: 117, 6: 94, 8: 89, 2: 74, 9: 66, 5: 61})
```



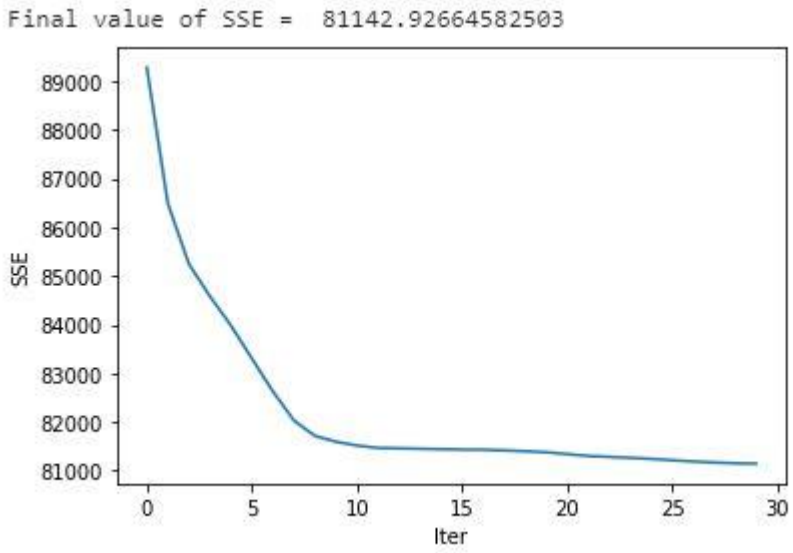
Yes clustering does make sense, we can see most of the digits are assigned to the expected cluster. Still there is some scope of improvement in this as you can observe some of the points are in correct cluster but the cluster centroid is pointing to some other digit

image. This happens mainly during data acquisition and not much could be done in it except increasing the cluster size to reduce error.

1.4]

SSE vs Iteration plot for k = 8 (num_of_iterations = 30)

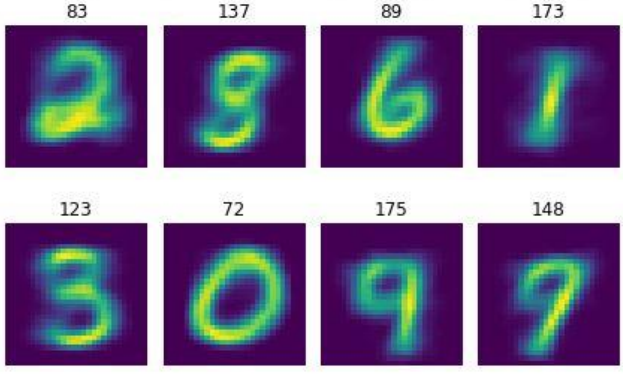
Final SSE = 81142.92



8 centroids and their corresponding test points in that cluster

```
[91]: print (acc)
      fig = plt.figure()
      for i in range(8) :
          plt.subplot(2,4,i+1)
          plt.imshow(np.reshape(centroid[i], (28,28)))
          plt.title (acc[i])
          plt.axis('off')
      plt.tight_layout()
      plt.show()
```

```
Counter({6: 175, 3: 173, 7: 148, 1: 137, 4: 123, 2: 89, 0: 83, 5: 72})
```



Yes clustering does not make sense here as it is underfitting the data and the error rate is huge. we can see most of the digits are not assigned to the expected cluster. So we need to increase the cluster size to assign points to the expected cluster.

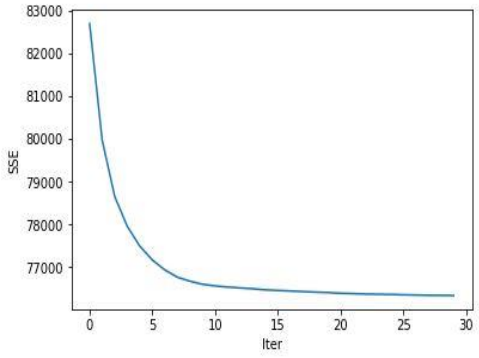
For k = 12

SSE vs Iteration plot for k = 12 (num_of_iterations = 30)

Final SSE = 76334.48

```
[96]: centroid = kmeans1(Xtrain,12)
      final = assignment(Xtest, centroid)
      acc = Counter(final)
```

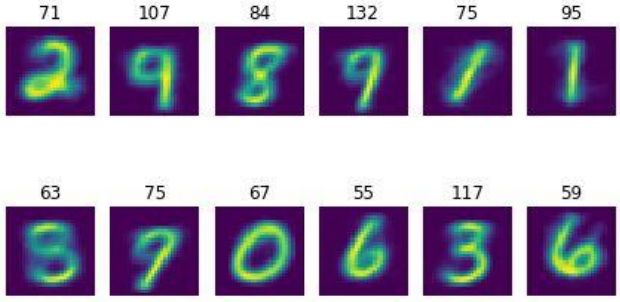
```
Final value of SSE =  76334.4863331424
```



12 centroids and their corresponding test points in that cluster

```
[97]: print (acc)
      fig = plt.figure()
      for i in range(12) :
          plt.subplot(2,6,i+1)
          plt.imshow(np.reshape(centroid[i], (28,28)))
          plt.title (acc[i])
          plt.axis('off')
      plt.tight_layout()
      plt.show()
```

```
Counter({3: 132, 10: 117, 1: 107, 5: 95, 2: 84, 4: 75, 7: 75, 0: 71, 8: 67, 6: 63, 11: 59, 9: 55})
```



Yes clustering does make sense here and from the result we can see that almost every data point is assigned to the expected cluster. The model here fits the test data well as the error is comparatively less. For more better result we can reduce the cluster size to one less.

2.2]

Num_clusters for used out_cam_04_debug.json = 2 (because of number of roads and considering direction. Here both resemble 2)

Num_clusters for used out_cam_10_debug.json = 8 (because of number of roads and their incoming/outgoing direction. So 8)

Num_clusters for used out_cam_16_debug.json = 4 (because of number of roads which is 2 and their incoming/outgoing direction. So 4)

Num_clusters for used out_cam_24_debug.json = 2 (because of number of roads which is 2 and their incoming/outgoing direction is also 2. So 2)