# CSE512 Spring 2021 - Machine Learning - Homework 5

Name: Sabrina Margetic

Solar ID: 109898930

Net ID Email Address: sabrina.margetic@stonybrook.edu

Names of people whom you discussed the homework with: None (Only the TA)

# 1 K-Mean Implementation

## 1.1 Implementation

## 1.2 Question 1.2

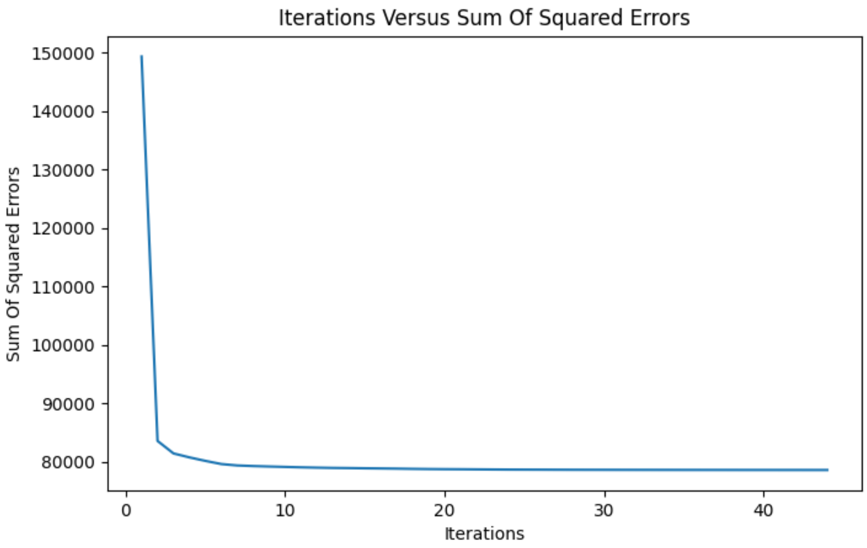

Figure 1

The sum of squared differences at each iteration was: [149322.3606612842, 83531.31979786711, 81391.94676171265, 80714.22636709266, 80115.63241038221, 79574.45699673193, 79331.5584449643, 79229.4546217492, 79151.98737506101, 79082.4933163793, 79013.56183361469, 78957.8152588373, 78909.9146645097, 78882.47315057152, 78846.28359696871, 78818.61155339284, 78788.22314265222, 78747.93431624824, 78715.90378691882, 78694.20310862402, 78680.79500508393, 78661.25800222649, 78641.88128392662, 78625.34971180173, 78617.51667073742, 78611.1775166947, 78601.95610719029, 78594.14965105204, 78589.6635953607, 78586.65442756293, 78582.1896874887, 78578.3947568273, 78576.15834111454, 78573.3174858248, 78570.20968417574, 78568.16672829248, 78566.55870218523, 78563.62524013018, 78561.21263109244, 78560.0362628419, 78558.26981388191, 78557.63953876027, 78555.85676378252, 78554.94388545043]
As you can see, the final sum of squared differences was 78554.94388545043 . This may appear to be a large number, but given that there are 2000 data points, each with 253 dim, and the fact that this is a squared value, the error is relatively small for each data point for each dimension.

## 1.3 Question 1.3

The cluster centers with their corresponding count can be seen in 2. These clusters make sense with some degree of error. We can see that we obtain the numbers 1,2,3,6,8,9 and 0, with numbers 4,5,and 7 being excluded while there are duplicates of 1, 9 and 6. The clusters make sense based on the possible variations that we would see amongst handwritten digits. Since 5 and 6 naturally resemble one another, it is possible that the way individuals write this made this overlap even greater, likewise for 7 and 1, and 9 and 4.
Furthermore, the assigned samples in each cluster are similar to the corresponding centroid. The majority of the centroid values are unique, with the only repetitive values accounting for similar looking numbers. Therefore, a sample is either classified in its unique categorization, or in one to which it bares a striking resemblance.

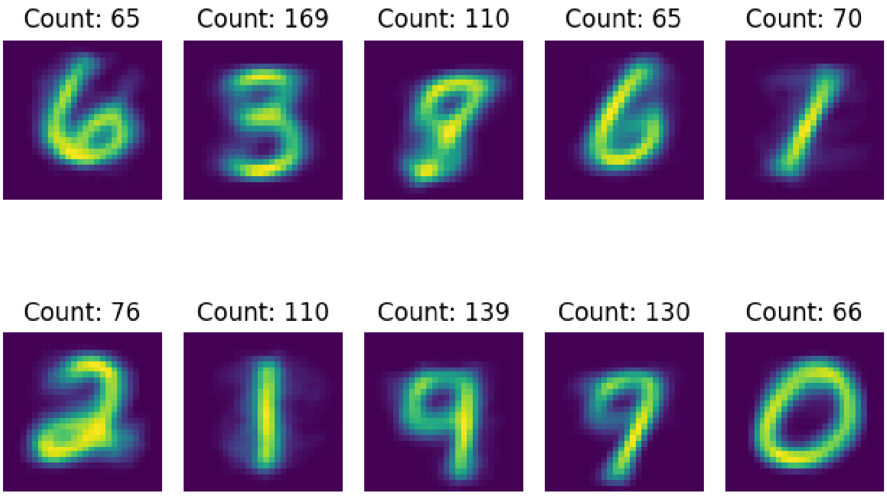

Figure 2

## 1.4 Question 1.4

For 8 clusters: The final value of the sum square error is: 81558.58169718565. The corresponding images for the sum of squared error and the centroid points can be seen in figures 3 and 4, respectively. We see here that firstly, the sum square error is greater than that in which the correct number of clusters were provided. Also, we now observe a unique spike up and down for the sum of squared errors. I'm not quiet sure why this would happen, possibly shifting a point led to a great amount of error.
Additionally, we now see that the centroids less distinct more obscure, we observe a 2 that resembles a 0 as a centroid point. We also observe a 8 that resembles a 9, and 3 that resembles a 8.

For 12 clusters: The corresponding images for the sum of squared error and the centroid points can be

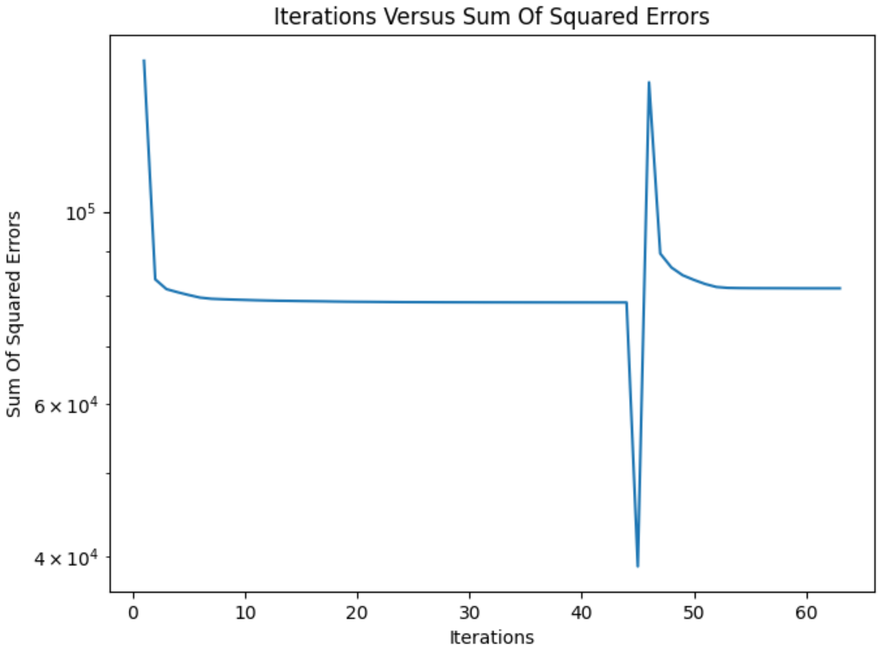

Figure 3

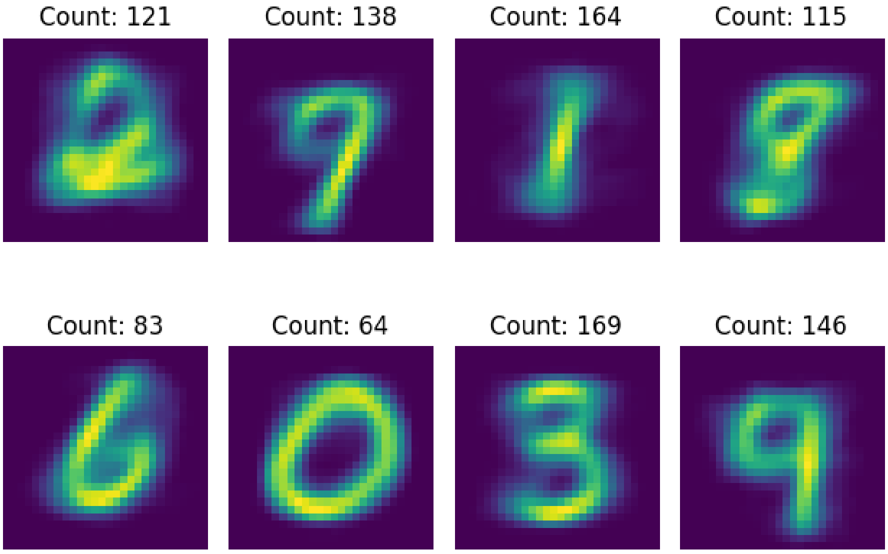

Figure 4

seen in figures 5 and 6, respectively. We now observe two spikes in the iterations versus sum of squared errors graph. Once again, my best guess is that shift a point to another cluster caused unforeseen error. In terms of the centroids, we now observe more distinct and less blurred figures. However, there is a lot of repetition in values. We see now 3 9's, 3 0's, and 2 2's, all while leaving out values like 4 and 7.

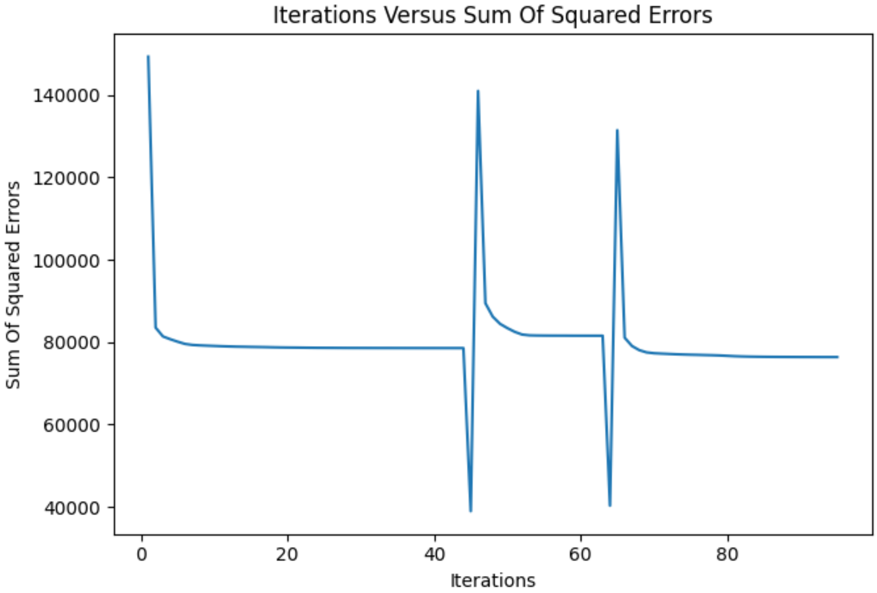

Figure 5

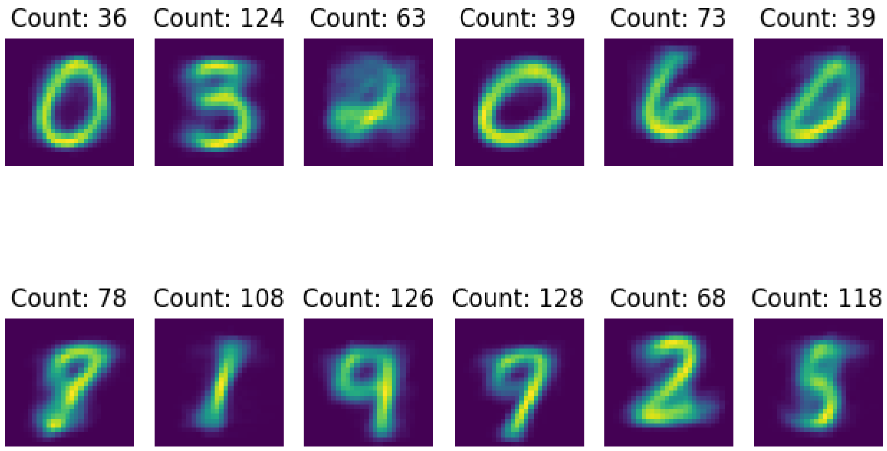

Figure 6

# 2 Question 2

## 2.1 Question 2.1

Sklearns kmeans was implemented for clustering. The feature vector used was the x,y coordinates of the centers of the initial tracklet and the final tracklet.

## 2.2 Question 2.2

Required files are attached. The number of clusters used was 5.