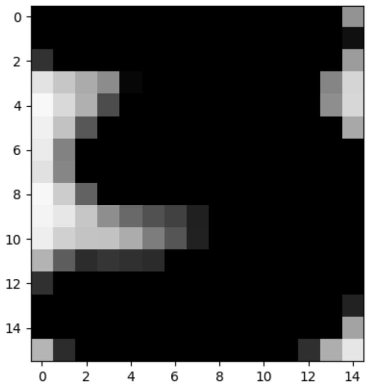
Assignment 4

Lorik Mucolli
Aavash Shrestha

We have selected the digit '3' to do the clustering. Below we describe how the different codebook vectors appear when different cases of K are chosen.

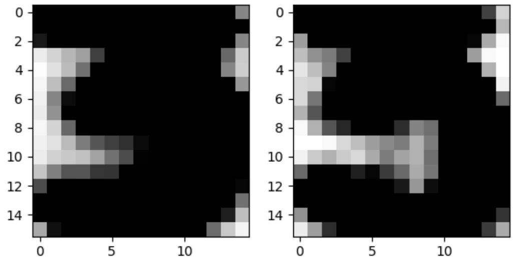
K = 1



Codebook Vector for k=1 case

As we can see, from the image of the codebook vector, the digit '3' is dense and captures a large area of the image. This is expected for this case as when K=1, there is only a single cluster, and the codebook vector is simply the mean of all the 200 points. This means that the image here is like a combined picture of all the 200 instances. Since each '3' (although retaining the shape of three) could have very different values for the 240 pixels, this dense and thick representation provides a typical view of what the digit three could and should look like. Most of the 3s in the 200 instances lie in the darker region and the blurry edges come from a few other instances with slightly different orientations.

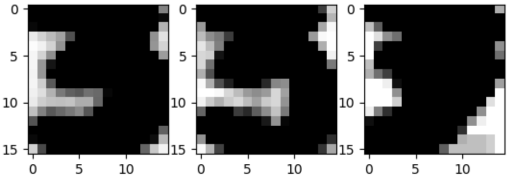
K = 2



`Codebook Vectors for k=2 case`

For K=2, we can see that the two images are slightly different from K=1 case. In particular, the codebook vector for the second cluster seems to be less general than the first cluster. It is likely that the second cluster has less instances of the '3' digit vectors than the first cluster.

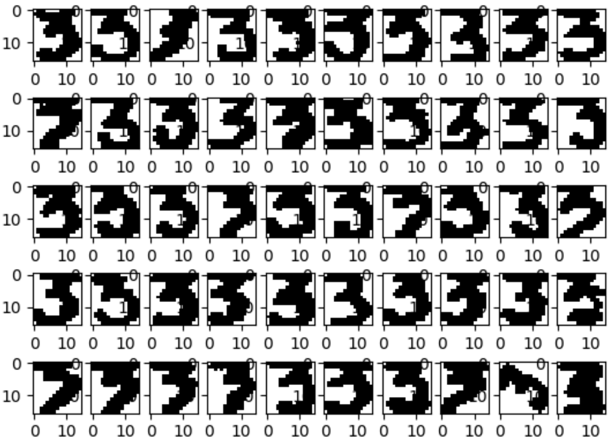
K = 3



`Codebook Vectors for k=3 case`

Similarly for K=3, the images now look more refined as the 200 instances are now distributed in 3 clusters. Here too, the second cluster seems to have fewer data points as the image is sharper, while the third cluster seems to have more oddly oriented '3's together in the same cluster(images with middle part of the '3' being close to the upper part). We can also observe that the blurriness at the edges is decreasing, which is obvious.

K = 200



Codebook Vectors for k=200 case, only 50 shown

In this case, all the 200 instances are divided into 200 clusters. Each cluster has only one image vector. Thus for every cluster, the original data point is itself the codebook vector for that cluster, being the only member of that cluster. Hence, when visualizing the codebook vectors(essentially the original data vectors themselves), we get to see how the data itself looks like(with or without k-clustering).