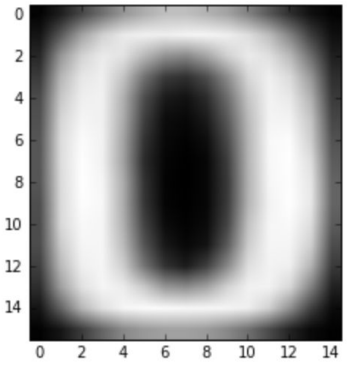
Machine Learning Exercise 4

Shun-Lung Chang, Anna Ruby

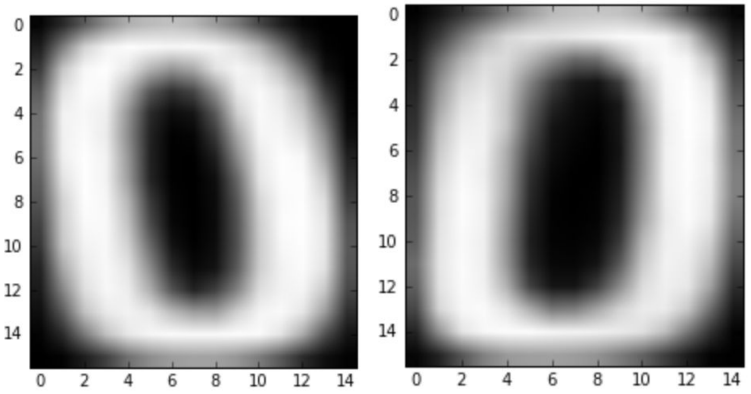
For our K-means clustering sample we have decided to work with the images of the digit 0. After creating our K-means cluster algorithm, we applied it to the 200 image vectors 4 times, resulting in codebook vectors for the case of 1 cluster, 2 clusters, 3 clusters, and 200 clusters. These codebook vectors produce codebook images which show the center, or average, of the cluster as a whole.

With 1 Cluster:



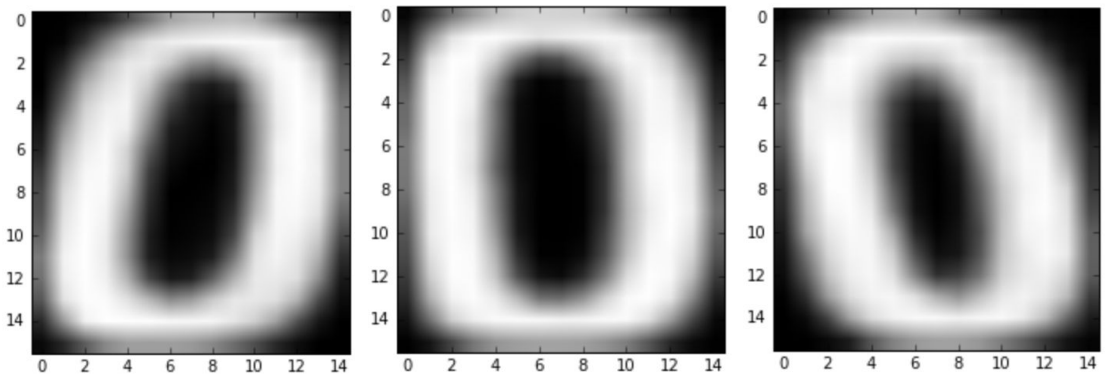
Here we have only one codebook image, as we have only one cluster. The above image is the average of all 200 image vectors in the sample set, so each component of the final codebook image vector is an average of each respective component from the original 200 image vectors. The 0 is notably lacking sharpness, which is an obvious result of the averaging, but has a very accurate 0 shape.

With 2 Clusters:



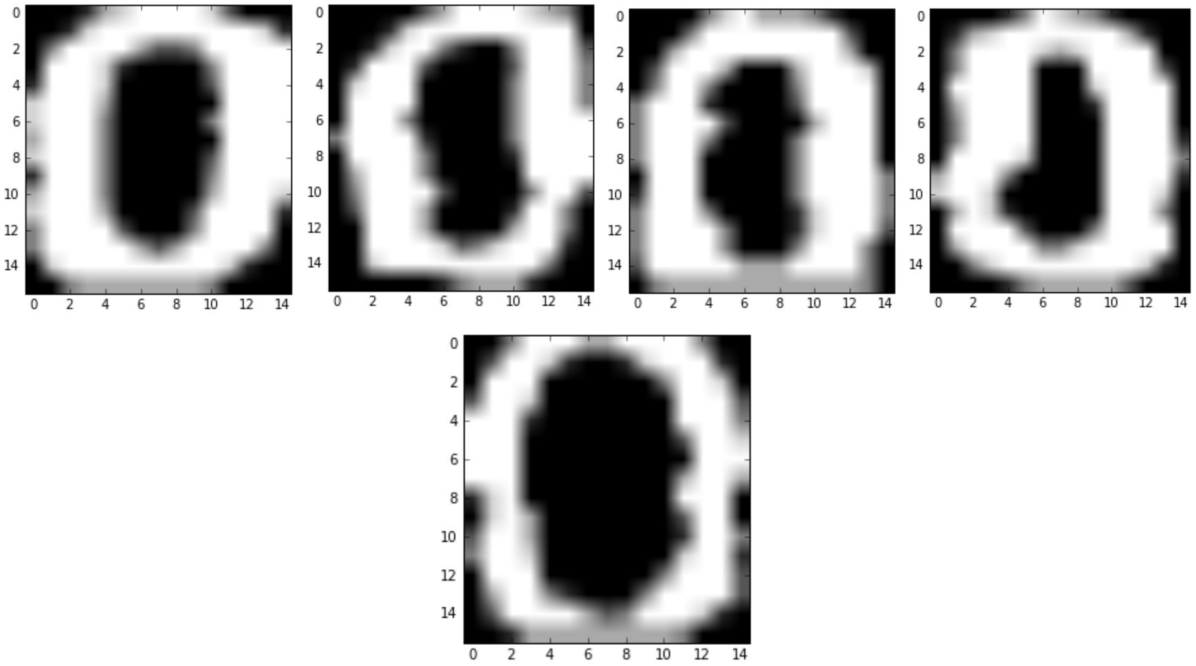
Here we now have two codebook images, the first showing the average image of the image vectors in the first cluster, and the second being the average of the second cluster. It appears that our two distinct clusters are being created based on the orientation of the 0; the first image shows a slightly left-leaning 0 while the second shows a very slightly right-leaning 0. Both images gain greater sharpness over the codebook image generated with only one cluster, and retain good 0 shape.

With 3 Clusters:



With three codebook images we now have the average image found in each of our three clusters. Again, we notice an increase in sharpness as our clusters become more uniform within themselves, and again we still have very clear 0 shapes. Additionally, we now have three distinct "leans" of the 0: leaning right, no lean, and leaning left, respectively. These three clusters, with their three different leans, could be showing the difference in right and left handedness, along with those 0's which cannot be classified distinctly as being written with a left or right hand, and so if we were to choose a cluster amount to use in analysis and future training, we would most likely select 3 clusters.

With 200 Clusters (5 examples):



Here we have codebook images that are notably less blurred than those images above, but they also have lost the look of a "well drawn" 0. These images are the most individual because, with as many clusters as there are image vectors, each image vector is itself the center of its own "one man" cluster, so instead of averaging, the image vector is mapped directly to the codebook image vector. Simply, the codebook image vector is the same as the

original image vector. As such, we have none of the smoothness to the digits that the clusters above produced.