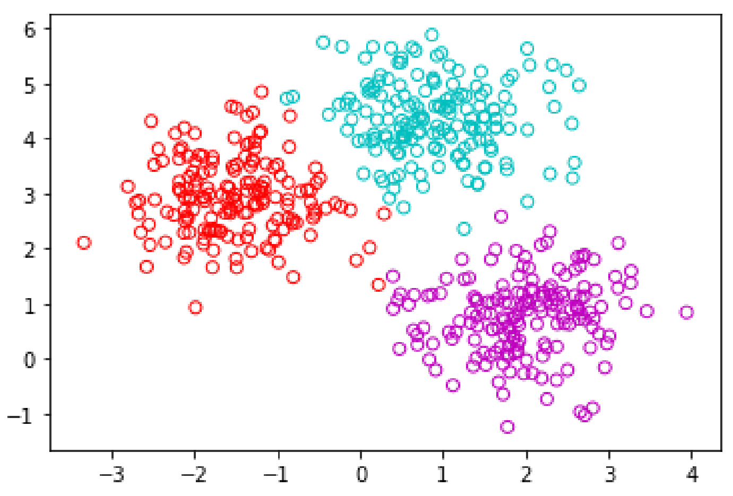
CMPE 462 - Project 3
Implementing K-Means & PCA
Due: June 14, 2020, 23:59

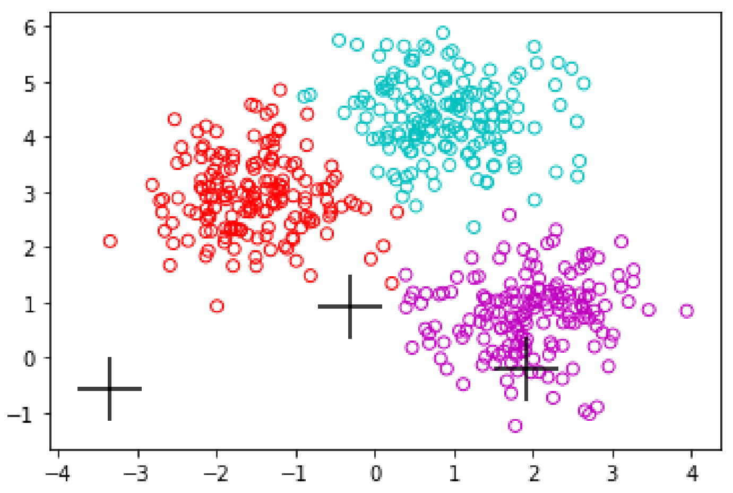
Task 1: K-Means Clustering

We first imported the kmeans_data and plotted it to see its general shape.

Below is the data labeled according to ground truth.

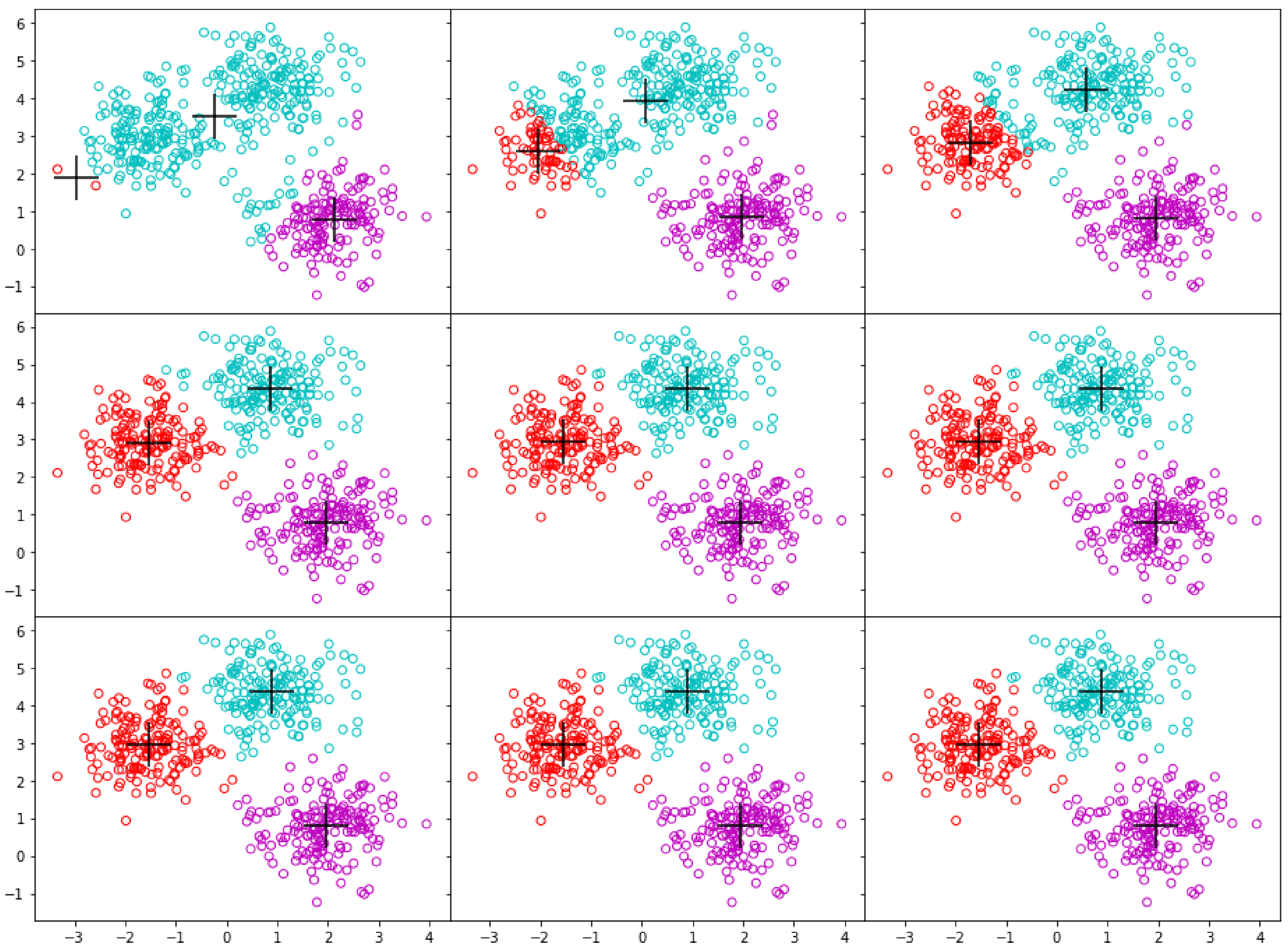


For K-Means implementation, first, we set k to 3, after that we picked 3 random points on the 2d plane using uniform distribution to be the initial centroids.



Then we wrote the loop for assigning the clusters. In the loop, first the distances from all of the points to all of the centroids were calculated, then the points are assigned to the minimally distant centroid's cluster. In each iteration of the loop after memberships are assigned, the new centroids are calculated by finding the centroid of the clusters. The loop runs until it reaches to the given iteration count.

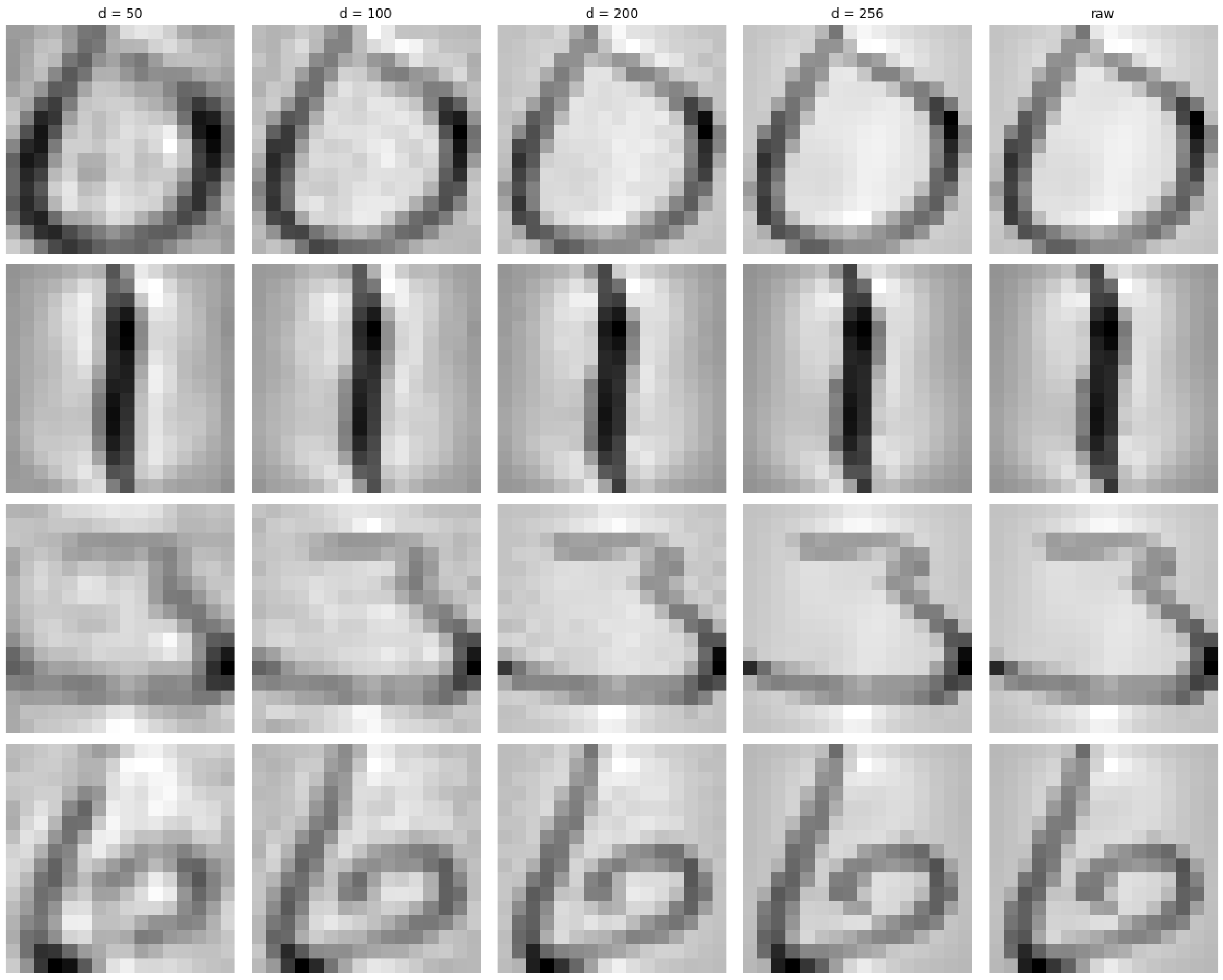
After calulating all these, we plotted them for each iteration for visual investigation. When we look at the results, it seems like the clusters no longer change after fifth iteration.



Task 2: Principal Component Analysis (PCA)

After finishing the K-Means task, we imported the PCA data (USPS handwritten digits). We standardized data by substracting the mean and dividing by standard deviation.

We calculated the covaiance matrix using `numpy.cov`. Then we applied eigendecomposition and sorted the eigenvectors according to their eigenvalues using a simple function. We followed that by forming the transformation matrix for each *d* value (50, 100, 200, 256). Then we applied PCA by doing matrix multiplication of the data and the transformation matrix. After applying PCA we reconstructured the dataset by multiplying again with the transformation matrix. Using that, we reconstructed the images at the indices 0, 500, 1000, 2000.



First, the raw images and the images at d=256 are the same as the data is already 256 dimensional. As d goes down the images seem more noisy but they are still recognizable even at d=50. Some digits are more robust to distortion than others, for example the digit 1 does not get distorted as much as the digit 6.