DA Lab 06 Write Up

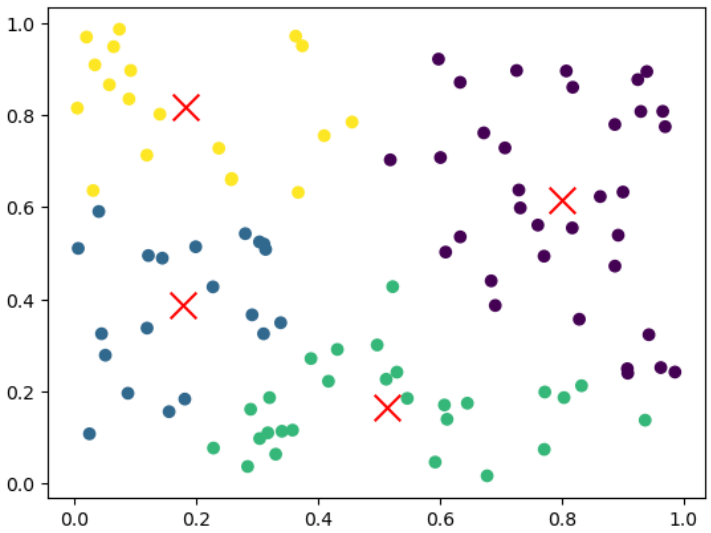
Justin Rickey, Quynh Thai, Dani Monroe

Introduction

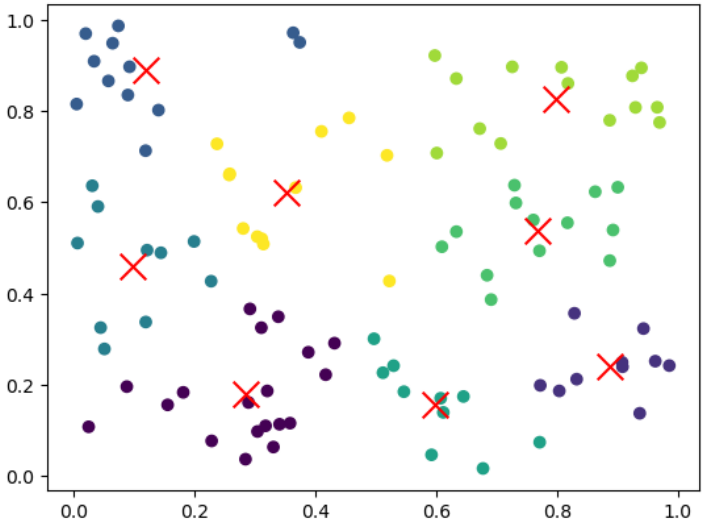
For this assignment, we were tasked with creating a K-Means Clustering algorithm on a seeded randomly generated data set that we created. Oue goal was to sort 100 points of data into related clusters by position. We then evaluate the error deviation of each point to their centroid.

Coding

We initialized our program by generating a random set of 100 points between the values of 0 and 1. Our first method uses randomly selected points from the generated data to initialize the first set of centroids. We then take the distance of each point to each centroid, then assign them to a cluster based on the shortest Euclidean distance from point to centroid. We also assign new centroid locations by averaging the values of each point in a cluster's x, y coordinates. If there is no change in centroid locations between iterations, the program has successfully completed clustering the data to optimal. After all clusters are in their final state, we used a new method to calculate the error via mean squared error method from points to centroid in each cluster. We decided to include the scatter plot of each run to provide a visual representation of each point in the data set, their assigned cluster, and where the cluster's centroid is located.



After the initial results, we decided to vary our runs by changing the number of clusters from the original 4 to 8 on the same data set.



Conclusions

Our initial run on the K-means with 4 clusters resulted in an average error of .0439 among all 4 clusters. After we updated our program to instead run with 8 clusters, we now had an average error of .00173, significantly reducing the number of errors in our program. We concluded that as the number of clusters increased, the error would decrease as each point can be more closely associated with their centroid which reduces the effects of outliers on centroid positioning. If the data is more spread out, error will increase as the distances grow from each point to their respective centroids.