DA Lab 6

Chad Martin, Hayden Edge

K-Means Cluster

Hayden and I originally had built the code without realizing that the return type was specified to be an ndarray. After refactoring the code, we were unable to get the code operational. For the original code we built a cluster class that contained a list of points that were unique to each cluster. We utilized two helper functions; get_closest_centroid(clusters, point), and get_average_point(data, closest_centroids, centroid_id). Our plan was to iterate over all data points and assign each of them to cluster that they are closest to. Then, collect the average of each dimensional value and use a running average to reassign centroids. After which we would repeat the process a large enough number of times that the centroids would come to a resting point.

To refactor we cut the cluster class and instead created a list that would store the id (or index) of the cluster that a point is assigned to. This list is indexed the same as the list of data points. The second list was a list of possible centroids that each point could be assigned to.