They-Means/They-Nearest neighbors

Team Members and Division of Labor:

Nikolas Varga will design the class and program the K-means section of the class with all relevant methods and attributes. Wendy Quizhpi will program the K-nearest neighbors section. There will be overlap in the shared methods such as distance calculation, plotting, and printing relevant information to the console.

Application Description:

The main features to be developed in this project will be unsupervised K-means and supervised K-nearest neighbors clustering algorithms. The K-means algorithm will be able to separate individual data points into K clusters. Points will initially be separated based on their distance from a random cluster head and points will be clustered to the nearest cluster head. The cluster head will then move to the mean of that cluster, and points will be regrouped. This algorithm will continue to run until the mean of the data points belonging to each cluster is at its minimum and the value of K is no longer changing. The KNN (K nearest neighbor) algorithm is able to separate an added point into a group using the closest members of the groups. The K value is the number of neighbors the algorithm will use to sort the added point into a group. This means that given K, the algorithm will find the K neighbors in closest proximity to the added point.The point is assigned to whichever group has the greater number of neighbors. For example data point x is being sorted with a k value of 5. The surrounding 5 points around the data set are 3 data points from category 1 and 2 data points from category 2. Since the category with the most data points closest to point x is category 1, point x will be sorted into category 1.

Application Design:

Our implementation of the K-means and K-nearest neighbors algorithms will take place in a single overarching class. This class will take inputs for the number of cluster heads or nearest neighbors as K and number of points as n. The class will have attributes for points, clusterheads, and the current classification of each point with distance included. When called, the class will require an input for K and n. Clusters in KNN will be preselected by the programmers. Methods of the class will include executing K-means or K-nearest neighbors, printing the values of points and cluster heads, updating the location of the cluster heads, and forming clusters. There will be protected methods that find the distance between points, generate the points, generate the first cluster heads or nearest neighbor, and serve other internal functions not necessary for the user.

Results:

Ideally, the final project will contain the K-means and K-nearest neighbor algorithms implemented with full documentation by Friday, December 11. If the project is not finished on time, the risk is dropping

at least one letter grade in the class. If this happens, we will have to finish the project outside of the given deadline. We will most likely ask for an extension of the deadline if we are struggling for time, or work heavily on it on Thursday and Friday of finals week. Going forward, Wendy and I will work on the project separately on Monday and Tuesday and together on Wednesday, Thursday, and Friday.