Stephen Hung
Christopher Sofian

Milestone B Option 3 Report

The main technique that we are using in our project is mainly supervised learning, mainly by implementing our own classifier, which in this case is K-nearest neighbors and random forest. The K-nearest classifier uses euclidean distance to compute the K nearest neighbors, and we used the data from these neighbors to predict/classify the test data we have in hand. At this point, our algorithm is working as intended for the most part, except the accuracy is still quite low. There might still be some bugs in the implementation of this K nearest neighbors that causes the accuracy to be this low, I will definitely have to fix this bug before the deadline. At this point, I can confidently say that I am 85% done with our implementation of K nearest neighbors. The random forest classifier splits the data into different groups in order to find the best possible split point using the gini index. Then the forest classifier splits the decision tree using the best possible split point into different decision trees. These trees are each used to train and return their predictions from a test data set. This method is called bagging. At this point, our algorithm works fine and seems to return a fairly high amount of accuracy (> 90%). However, I will still have to check for bugs in the implementation and test for splitting the tree into more than 2 decision trees. I believe I am about 90% done with our implementation of random forest. Ultimately I believe we are close to finishing as we just have to finish our algorithms and then start the comparison of our two classifiers based on several categories.

We decided to switch from Option 2 to Option 3 mainly because while starting Option 2, we were unable to fully understand the complexity and hardness of Option 2 and thus when we got deep into working on it, we realized we were unable to finish in time for milestone B. We read did the preliminary research on Option 3 and found that there were plenty of resources for the random forest and k nearest neighbor classifier and thus decided to switch over.