Jack Asaad - jfa180002

Isabelle Villegas - iav180000

Kernel and Ensemble Methods

a. SVM works by having data in a data set and separating many different points based on different classes in the data. For example, in a data set with many different data points and you are trying to separate them into two different classes based on an attribute, then SVM will try and find an ideal line to separate those two classes. Then it will give us the best approximation of that line separating those two classes and from there we can record the support vector, which is the points on either class closest to the optimal line. From there, SVM tries to create a line that creates a wide boundary between both support vectors. This can all apply to linear, polynomial, and radial kernels, in which a better fit can be made given the type of kernel.

SWM works extremely well when given a data set with two well defined classes that can be easily separated. Given a large data set, SVM does not work the best especially when the classes overlap with each other. SVM works well when the data set is small though and is relatively memory efficient.

Overall, I think SVM works well, however, since we needed to use relatively large data sets compared to pretty small ones, it was definitely a lot harder to create good approximation lines.

b. How Random Forest works, how the other 2 algorithms you used work compared to the simple decision tree, your impression of the strengths and weaknesses of these ensemble techniques

b. Random Forest takes a lot from the standard decision tree algorithm and tailors it towards accuracy. It takes multiple decision trees and operates as a group, where each tree in the forest will make a decision and when doing classification we take the majority of tree results and that becomes the prediction for the data set. This gets us a power in numbers effect with the data, if multiple trees agree on a classification the more inclined I am to believe there is some relationship there. One massive weakness is the computational complexity of random forest which was borderline unbearable to my old laptop but we somehow got something in the end.

Adaboost is another ensemble learning method that gradually improves classification by strengthening other weak predictors and making them strong ones. Adaboost also is a bit of a resource strain based on how long it took to calculate on my machine but also produces strong results compared to the decision tree. Additionally xgboost is another boosting machine learning algorithm that does some improvisation via gradient boosting and is far more efficient than most of the algorithms explored previously.