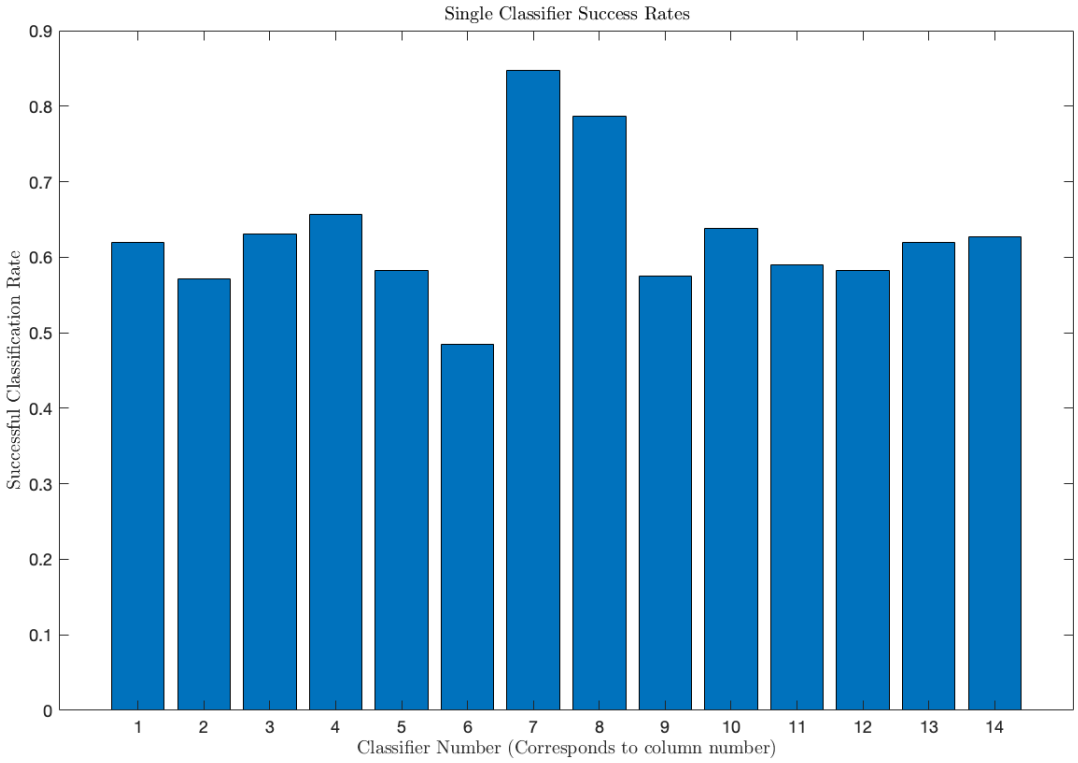
Homework 5

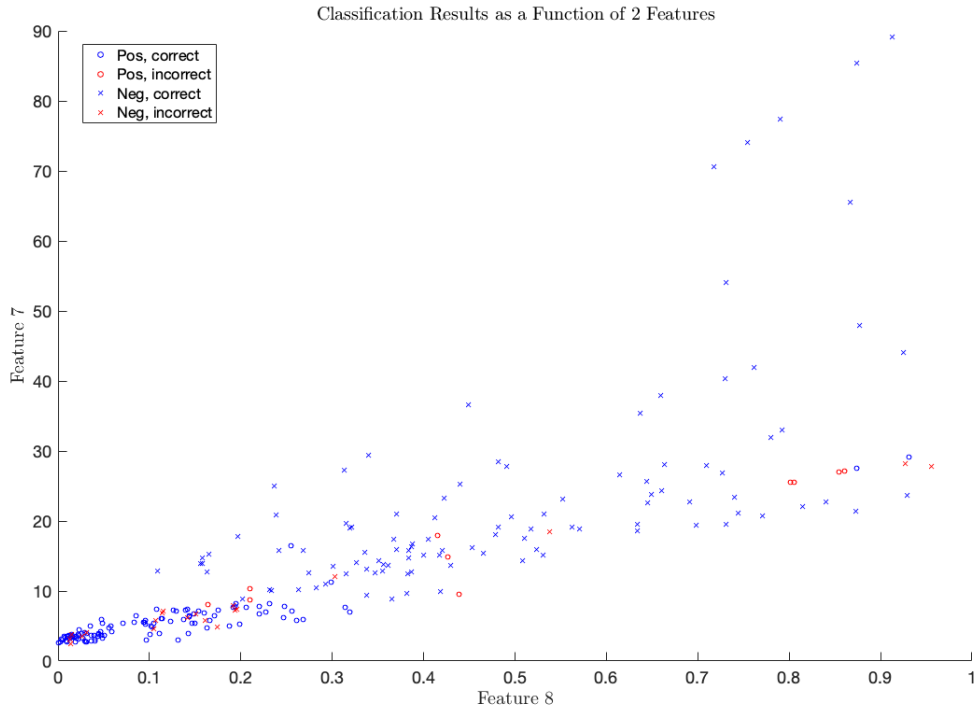
Ryan Nguyen

December 2018

As a baseline, using all classifiers leads to a 95.9% correct classification rate. I made a graph mapping the successful classification rates of each classifier alone below. We see that features 7 and 8 have the highest successful classification percentage. The script for this can be found in main.m

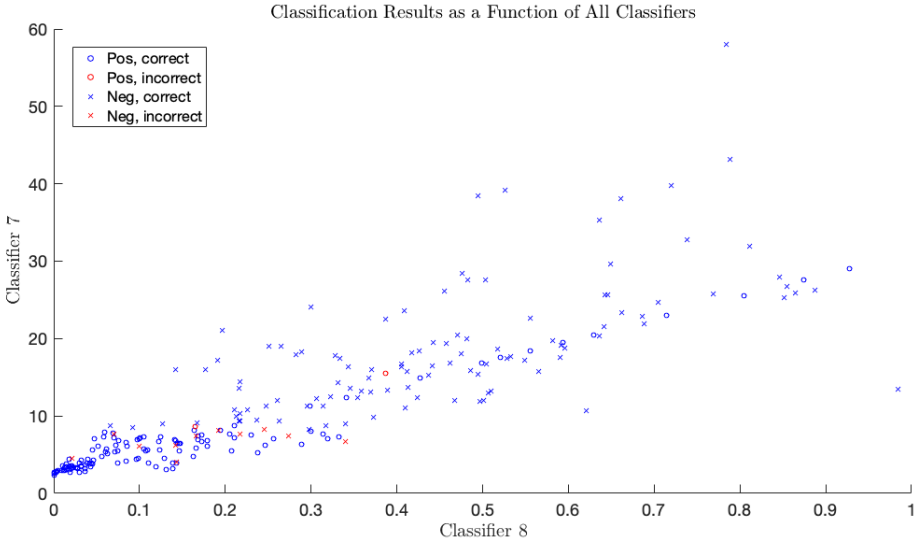


I then built a classifier based on only features 7 and 8. This leads to a 91.8% correct classification rate. Shown below is a graph depicting these classifications. Each axis represents units of features 7 and 8, respectively. Blue circles represent SNV's classified correctly as positive, and red circles represent SNV's classified correctly as negative; Blue x's represent SNV's incorrectly classified as positive and red X's represent SNV's incorrectly classified as negative. The script for this can be found in main2.m



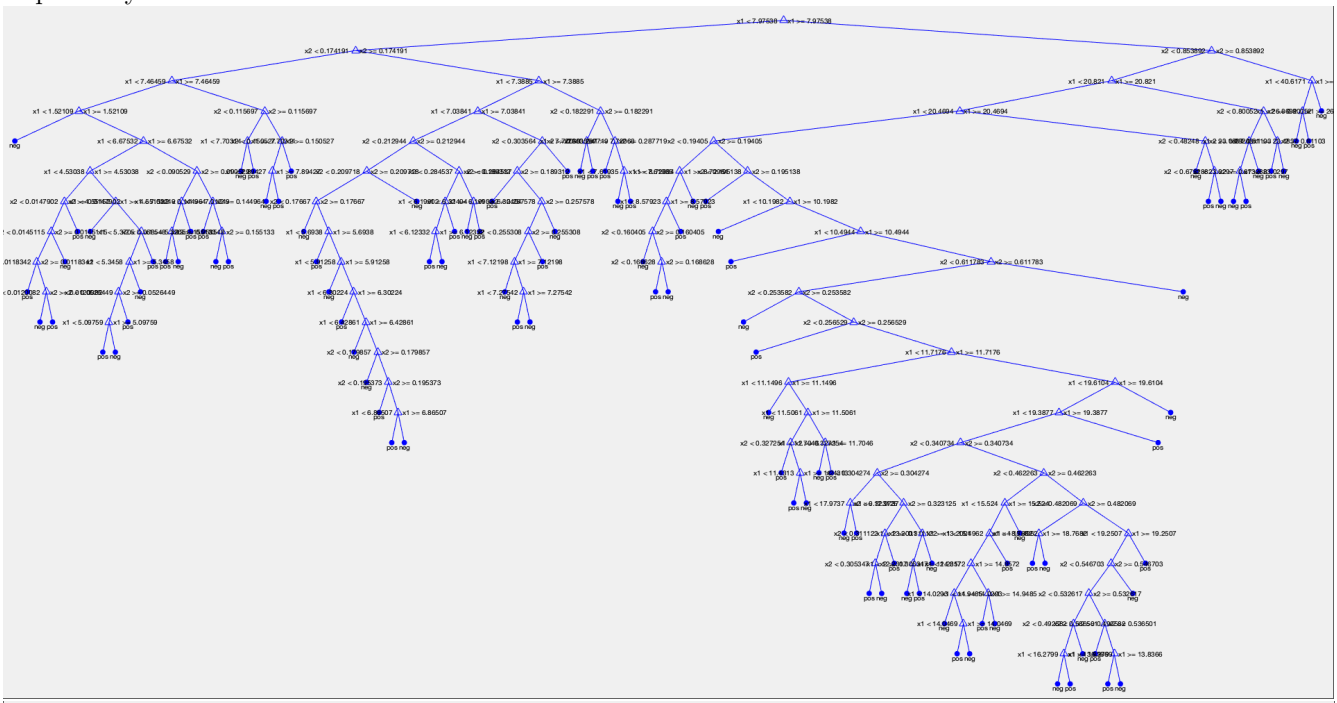
I notice that many of the incorrect classifications tend to localize at lower values of feature 7 and feature 8. In middle to low values (0.1-0.6) of feature 8 and low values of feature 7, some misclassification is noted. At higher values of feature 8, this classifier is actually quite accurate but some mistakes are made. Many of the incorrect classification tend to be red x's, indicating that the classification system tends to call some SNV's negative when they are supposed to be positive.

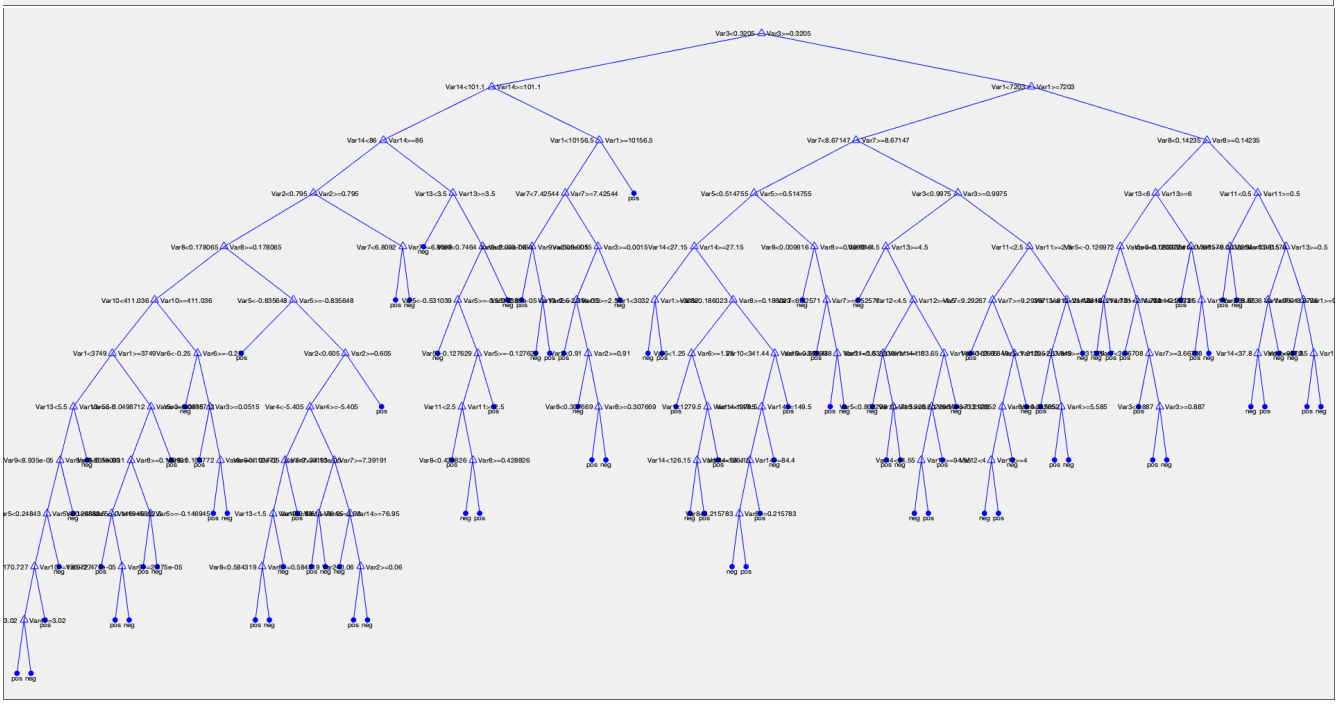
This same analysis was done using all classifier and can be seen in the graph below. In general, you see a visible improvement in classification, especially in the mid to high values of feature 8. In values of feature 8 less than 0.1, we see better classification as opposed to that of the two classifier system. However, 0.1-0.3 is also a trouble area in classification. Many of the incorrect classification also tend to be red x's, indicating that the classification system calls some SNV's negative when they are supposed to be positive. However, this occurence is less frequent than that of the two classifier system. The script for this can be found in RandomForestIntro.m.



These two systems match up by saying that low values of feature 8 tends to mean more positive SNV's while higher values result in more negative SNV's. Values of feature 7 tend to be greater than 10 if one has a negative SNV. In general, one can also see positive correlation between higher values of feature 8 and negative SNV's. Biological interpretations of this are limited because the only descriptors I have for feature 7 and 8 are normalized distance and allele frequency, respectively. Like seriously, I don't even know the units for feature 7.

I then plotted the tree of both the two classifer system and the all classfiers system which is shown below, respectively.





The two classifier one has an error rate of 34% while the all classifier one has an error rate of 32%. Though more classifiers are being used in the all classifier system, the maximum amount of branches to travel for any given data point is 10 branches. In contrast, the two classifier system has a maximum amount of branches to travel as 16. In general the two classifier system has a more complex decision tree. Both trees tend to label features 7 and 8 as as important early on. This is obvious in the case of the two classifier system and can be seen on the right uppermost node on the all classifier system.

This picture concludes this assignment and semester.