Intro2MachineLearningHW07 – Gökalp Ünsal

In this homework, I went online and found that for binary classification, logistic regression is one of the best. Also, for large datasets, random forest algorithm is very useful and simplistic. How logistic regression algorithm works is that instead of drawing a linear line to distinguish the classes, it uses sigmoid function to create an s-shaped border between classes. What Random Forest does is that, it creates numerous decision trees and labels the data by taking votes of all the trees in the forest, then decides the label with the most votes. I tried them both and got a better result with random forest which was predictable with its popularity. For what I found online was that overfitting with this algorithm was highly unlikely. So that I could increase the number of trees. And I also it was a bad idea to test our data with the trained data, so I split the train data and the validation data with the ratio of 0.75. The validation part was selected as random indices from the data because the value of the first feature was same as one index's adjacent ones. If I had to find the best possible split in my Decision Tree, I would have to create a cross validation algorithm and test it with different splits. However, the purpose of Random Forest algorithm is to create a bunch decision trees with different splits and take the majority voting to decide the class of the data as the best possible way. The score I got was AUROC = 0.7600026 with 200 trees which is the best I could reach with using validation of course.