Lachlan Sinclair

2/9/2020

Week 5 Assignment 2

Reflections

A. There has been quite a few different definitions of trees and forests flying around, my interpretation of this question is how you can generate multiple different trees from the same data. One of the potential changes to the tree generation process I think you can make is altering the allowed depth of the tree. This should have a large impact on the structure of the tree since the tree will include more decision nodes throughout the branches. However, this will only get a few unique trees before problems might be encountered, the textbook insinuated to deep of trees can cause problems. While looking through the IBM Knowledge centers page on random trees, they give a brief explanation on how random trees are selected. The process entails randomly selecting only a portion of the predictors. Then out of this reduced set it finds the option that reduces the impurity the most (has the greatest information gain) and makes the first split using that predictor, this splitting technique is then continued to fill out the tree. It then continues this process to generate more unique trees. Therefor in a generalized sense, changing the way each tree selects its sequence of predictors to split on will create unique trees. For IBM's random trees there is no limit on the depth, I couldn't find any information why this is the case, but I don't see why it is unreasonable for someone to limit it if they had a valid reason to do so.

B. One-way XGboost can be adapted to categorical targets is by effectively modeling the categorical probabilities as the target, then using those probabilities to make the categorical decision. Since the categorical probabilities are numeric targets, standard gradient boosting formulas such as XGBoost can be used on them. For multiclass categorical targets rather than being able to simply use the probability of one option over the other, a model must be generated for every categorical option, these use the probability of that category verse not that category. Each of these models will be a GBT model, and the result of said models is used in the softmax function referenced in the lectures. This in turn is done for every weak learner which results in quite a large process. Ultimately the class with the highest probability is selected at the end. In the XGBoost part of the assignment while looking at the preview table you can see the $XGTC-LEAVE field which is the confidence level of the predicted category. This confidence level is the probability value that served as the numerical target value that was used by the gradient boosting. After the gradient boosting process was finished that probability was then used to assign the categorical value as expected.