MATH 80619 Advanced statistical learning

Assignment #3 (part 2) – chapter 4

Due March 8th, 2021

Specific instructions

Same as in part 1. Please submit a single document for the assignment (not one document per part).

Question #2

In this question, you will practice tree-based boosting (adaboost and L2-gradient boosting) on a dataset with a binary outcome. We'll use the ionosphere dataset, taken from the UCI Machine Learning Repository. The original data set has 351 observations with 34 predictors, but the first two features are eliminated. So, the total number of predictors is 32. The class '-1' is considered as outliers' class (bad class) and the class '1' as inliers (good class). A more complete description of the dataset can be found here: http://archive.ics.uci.edu/ml/datasets/Ionosphere
The train and test datasets are given in the files `ionosphere_testdata.csv` and `ionosphere_traindata.csv`.

Use the following methods and compare the misclassification rate:

1) Single tree (pruned)
2) Random forest with 100, 200, 300, 400, 500 trees. Select the optimal number of trees.
3) Adaboost based on a stump with 100 trees
4) Adaboost based on trees with maxdepth=5 with 100 trees
5) L2-Gradient boosting (equivalent to L2 boosting) based on a stump with 100 trees using a shrinkage parameter of (0.01, 0.05, 0.1, 0.2). Select the optimal shrinkage.
6) L2-Gradient boosting (equivalent to L2 boosting) on trees with maxdepth=5 with 100 trees using a shrinkage parameter of (0.01, 0.05, 0.1, 0.2). Select the optimal shrinkage.

Comment your results.