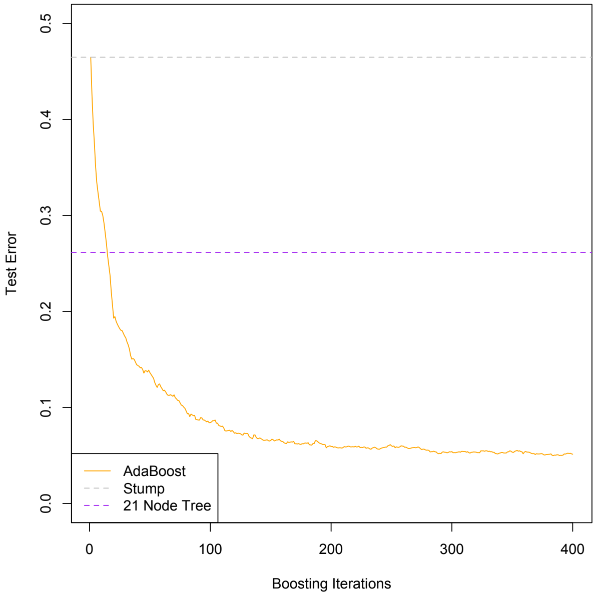
NOTES. **NO** late submission will be accepted. Computer generated output without detailed explanations and remarks will not receive any credit. You may type out your answers, but make sure to use different fonts to distinguish your own words from computer output. Scan or take photos of your homework solutions and submit through CANVAS. For the simulation and data analysis problems, keep the code you develop as you may be asked to present your work in a in-class workshop.

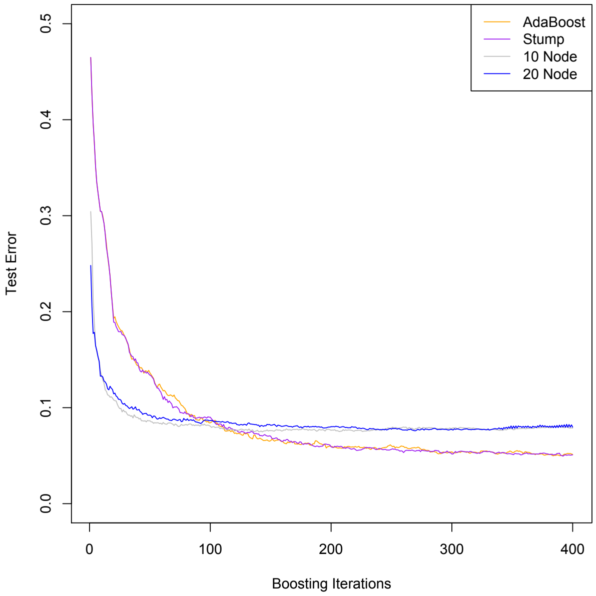
1. Write a program to implement AdaBoost with trees (Algorithm 10.1). [Hint. The `rpart()` function has a argument `weights`, which you need to supply for Step 2(a) of the algorithm. Also, use the `control=rpart.control(maxdepth=1)` so that a stump is added in each step.] Do the following using your program.

(a) Reproduce the following figure. [Note. Use your program, don't use the `gbm()` function.]



(b) Investigate the number of iterations needed to make the test error start to rise in the figure above.

(c) Reproduce the following figure. [Note. For `AdaBoost`, use your program, don't use the `gbm()` function. The three other curves can be directly reproduced using the code given in the slides.]



2. This problem uses the `spam` data, which have been uploaded to Canvas.

(a) Try to play with the parameters `interaction.depth` (J), `shrinkage` (ν), `bag.fraction` (η) and the number of iterations M. Report the best test errors you are able to obtain, together with the corresponding J, ν, η and M.

(b) Try logistic regression, SVM, and compare the results.

3. This problem uses the `Caravan` data set that is available in the package `ISLR`. See Sectoin 4.6.6 for an introduction of this data set. Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.

(a) Fit a boosting model to the training set with `Purchase` as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?

(b) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20%. [Hint. When you use the `predict()` function, set the parameter `type` as `type="response"`. This gives the predicted probability of making a purchase.] Form a confusion matrix (you may want to check Page 145 of ISLR for the definition of the confusion matrix). What fraction of the people predicted to make a purchase do in fact make one?

(c) How does this compare with the results obtained from applying logistic regression to this data set?

4. Derive expression (10.12) of ESL for the update of β in AdaBoost.