Homework 6

Due Thursday 4/7

Solutions should include R code and output – try to use `R markdown`. Homework must be submitted on Blackboard as one file.

Zhaoyu Yang, Ruoyu Zhang, Shuhan Zhang and Zhang Zhang will lead the HW discussion.

1. (90 points) *Modeling High COVID-19 Community Levels for Indoor Masking Recommendation.*

   This question is a continuation of HW 4 Q1 & HW 5 Q1 and usees the same data. We are interested in models for classifying geographies into COVID-19 community levels reflecting whether or not masks should be worn indoors, based on 17 geographic and demographic characteristics.

   (a) Fit the following models
   - bagging with classification tree base learner,
   - random forest,
   - adaBoost with classification tree base learner, and
   - gradient tree boosting,

   and obtain an estimate of test error using 0-1 loss.

   For bagging, use cross-validation or OOB error to select the number of trees B. You may use the default tree size in `bagging` and `randomForest` which is grow a full tree (minimum node size of 1 or equivalently $\alpha = cp = 0$). For random forest, use cross-validation or OOB error to select *mtry* and *ntree* – make sure to choose the optimal values based on a joint grid search of the two tuning parameters. For the two boosting algorithms, use cross-validation to select *n.trees* (number of trees) and *shrinkage* (shrinkage parameter for learning rate) – make sure to choose the optimal values based on a joint grid search of these two tuning parameters. You may assume the default *interaction.depth* = 1 (i.e. boosting based on stump trees).

   Note: If estimating test error by cross validation is too time consuming computationally, then it is ok to just split the data into a 50% training sample and a 50% test sample to obtain an estimate of test error based on fitting with the training data and evaluating accuracy on the test data.

   (b) For each of the methods, plot the test error curve/surface plot as a function of the tuning parameter(s). Explain the optimal tuning parameter value you chose for each of the methods.

   (c) Discuss how the test error estimates compare across the methods.

   (d) Obtain variable importance measures for the bagging and random forest models and relative influence measures for the boosting models. Compare the rankings of predictors. Also compare to the variables selected in your decision tree from HW 5. Discuss your conclusions about the important predictors of the indoor masking recommendation.

(e) Compare the test error for all these classifiers and the classifiers fit in HW 4 and HW 5. Which approach would you recommend?

2. (10 points) *Boosting as an Additive Model*. Do problem 2 in Chapter 8 (pg. 361) of ISLR: `https://hastie.su.domains/ISLR2/ISLRv2_website.pdf`.