Homework 5

Due Thursday 3/24

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

Haojia Tu, Xuyue Wan and Qinyuan Xing will lead the HW discussion.

1. (60 points) *Modeling High COVID-19 Community Levels for Indoor Masking Recommendation.*

   This question is a continuation of HW 4 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 a classification tree [using the Gini index as the splitting evaluation criteria]. Use cost-complexity pruning to obtain a final tree based on optimal α chosen by cross-validation. Include a plot of the test error as a function of the size of the tree.

   (b) Obtain an estimate of the test error for your pruned classification tree using cross validation with 0-1 loss.

   (c) Display the tree diagram and print the tree object. Translate the output to the set of rules that identify counties for which the tree predicts masks should be worn indoors.

   (d) Each terminal node m of a classification tree is associated with an estimate $\hat{p}_{m1} = \hat{P}(Y = 1|\mathbf{X} \in R_m)$. Define class predictions based on this predicted probability using the cutoff $c = \frac{1}{N}\sum_{i=1}^{N} y_i$ and obtain the estimate of test error. How does this test error estimate compare to those obtained using the default cutoff $c = 0.5$?

   (e) Plot the ROC curve for the classification tree. What do you notice about the shape? Discuss what the ROC curve tells you about the predictive power of the classification tree for predicting high community levels for indoor mask recommendations.

   (f) Compare the test error from your chosen classification tree to that from your chosen "best" classifier in HW 4. Which would you recommend and why?

2. (40 points) *Program Your Own Regression Tree.*

   This question uses the *Hitters* data from Chapter 8 of *An Introduction to Statistical Learning*: `https://hastie.su.domains/ISLR2/ISLRv2_website.pdf`. You will be predicting $\log(Salary)$ with $Years$ and $Hits$. The data is the *Hitters* object in the `ISLR` package in R. Note that you are only using two of the many predictors and you will need to drop observations with missing *Salary* values.

   Implement the basic regression tree CART algorithm from scratch* using squared error as the split criterion. Do not prune the tree, but rather use the following stopping rule: do not attempt to split the node if the number of observations in the node is less than 100. This stopping rule is equivalent to setting $minsplit = 100$ as a control parameter in the `rpart` function.

   * You may not use pre-packaged functions in R. You are programming the algorithm from scratch. You may, however, use `rpart` to check your final tree .