Homework 2, due September 20th, 11:59pm

September 14, 2023

1. In this problem we use the `abalone` dataset available on Canvas. The dataset is about predicting the age of the abalone from its physical measurements. Use the first 7 variables as predictors and the 8-th as the response.

Report all results as the average of 20 random splits. For each random split divide the data at random into 85% for training and 15% for testing, train the models and compute the training error and the test error (or R^2) for that split. Repeat this process 20 times obtaining 20 different random splits of the data and report the average training or test MSE or R^2 obtained over the 20 splits for the following models:

a) Null model. Report the average train and test MSE of the null model that always predicts training $\bar{y}$ (average training y). (1 point)

b) OLS regression, analytic, by solving the normal equations, with $\lambda = 0.0001$. Report the average training and test R^2 and MSE. (2 points)

c) Regression tree of maximum depth 1, 2, ... up to 7, for a total of 7 regression trees. On the same plot, plot the average training and test R^2 vs the tree depth. On another plot, plot the average training and test MSE vs the tree depth, and show the null model MSE from a) as a horizontal line. (3 points)

d) Random forest regression with 10, 30, 100 and 300 trees. Report the average training and test R^2 and MSE in each case. (3 points)