Homework 2, due January 24th, 11:59pm

January 17, 2024

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, which are computed as follows. For each random split divide the data at random into 90% for training and 10% for testing, train the models on the training set and compute the training and the test MSE (or R^2) for that split. Repeat this process 20 times obtaining 20 training errors and 20 test errors and report their averages as the training or test MSE or R^2 obtained over the 20 splits.

Report results 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 computed analytically by solving the normal equations, with $\lambda = 0.001$. Report the average training and test R^2 and MSE and their standard deviations. Also report the average value and standard deviation of the logarithm of the determinant of $X^T X + \lambda I_p$ over the 20 splits. (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 and their standard deviations in each case. (3 points)