OLS, Regression Trees and Random Forest Regression

By

Charles Silkin

Contents

PROJECT INSTRUCTIONS 3
PART A 4
PART B 5
PART C 6
PART D 7
REFERENCES 8

PROJECT INSTRUCTIONS

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

Report all results as the average of 15 random splits. For each random split divide the data at random into 90% for training and 10% for testing, train the models and compute the training error and the test error for that split. Repeat this process 15 times obtaining 15 different random splits of the data and report the average training error obtained over the 15 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).

b) OLS regression, analytic, by solving the normal equations, with $\lambda = 0.001$. Report the average training and test R^2 and MSE.

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.

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

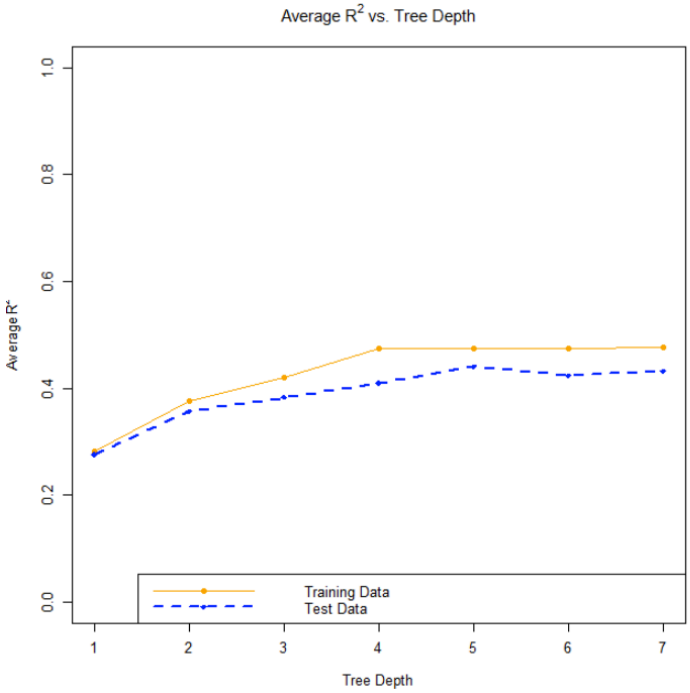
PART A

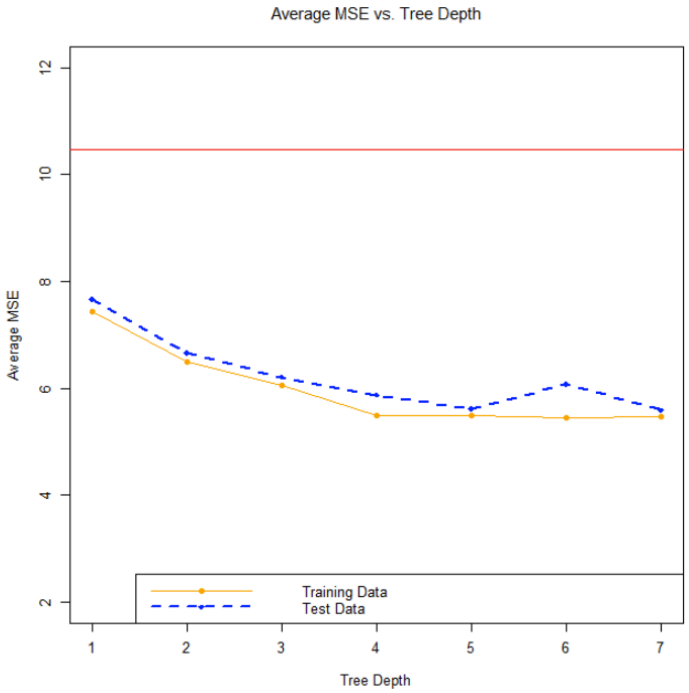
Average Train MSE	Average Test MSE
10.38495	10.469

PART B

Average Train R^2	Average Test R^2	Average Train MSE	Average Test MSE
0.5158684	0.4782119	5.022978	5.47295

PART C


Average R^2 vs. Tree Depth
Average R^2
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
Tree Depth
Training Data
Test Data


Average MSE vs. Tree Depth
Average MSE
2
4
6
8
10
12
1
2
3
4
5
6
7
Tree Depth
Training Data
Test Data

PART D

Number of Trees	Average Train R^2	Average Test R^2	Average Train MSE	Average Test MSE
10	0.9025596	0.5483427	1.014328	4.604571
30	0.9021141	0.5649803	1.014701	4.612964
100	0.9023596	0.5554627	1.011707	4.697928
300	0.9027545	0.5429270	1.014086	4.567754

REFERENCES

1. https://www.guru99.com/r-random-forest-tutorial.html
2. https://www.geeksforgeeks.org/decision-tree-for-regression-in-r-programming
3. https://tomroth.com.au/regression/
4. https://www.r-bloggers.com/2012/06/how-do-i-create-the-identity-matrix-in-r/