

Variables

x_train: the values of the training set. (1*100)

x_test: the values of the test set. (1*33)

y_train: the labels of the training set. (1*100)

y_test: the labels of the test set . (1*33)

p: the parameter for pre pruning. = 10

y_predicted: the predicted labels of the test data.

rmse: root mean square error calculated the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_{test}} (y_i - \hat{y}_i)^2}{N_{test}}}$$

p_testing: the list of rmse of the different pre-pruning parameter runs.

Functions

train: the loop for training the data which I got from the lab code. I changed the terminal node condition and the scoring function.

tree: Also it is now a function that gets a dataset to return a y_prediction value with triggering the train function.

In this homework we implanted a Decision Tree regression, with the pre-pruning parameter. The code resembles the code we did in the lab session, but it is a regression algorithm instead of classification. So, the scoring must be different from the lab session. Also, in addition we have a pre-pruning parameter that if the split node has less values than the pre-pruning parameter. We stop splitting that node. I had to cover the loop where we train the data in a function because of the 5th part of the homework. I had to run that loop 20 times so with the function I was able to re run the loop.