

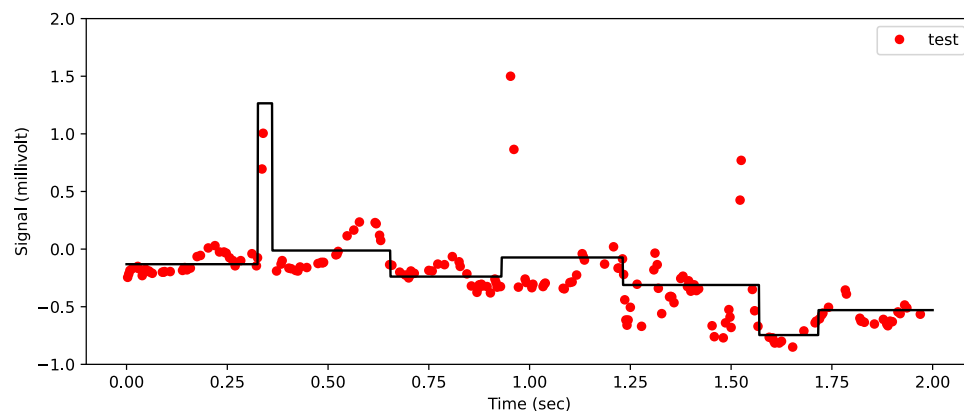
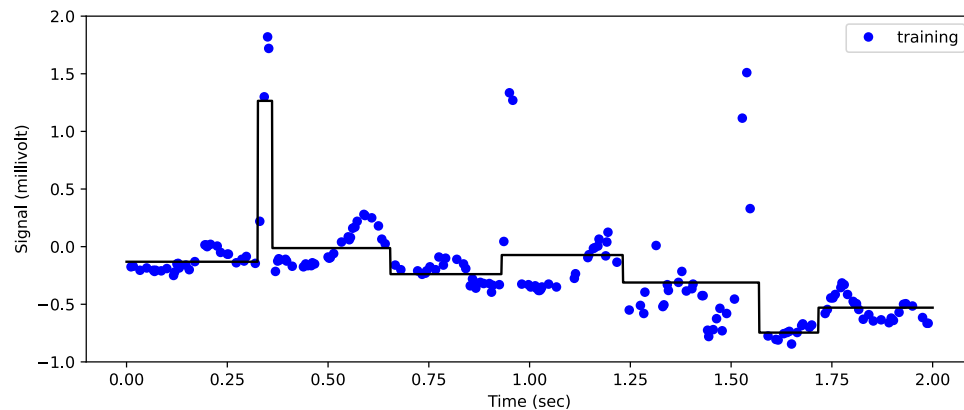
ENGR 421 / DASC 521: Introduction to Machine Learning

Homework 05: Decision Tree Regression

Deadline: April 22, 2022, 11:59 PM

In this homework, you will implement a decision tree regression algorithm in Python. Here are the steps you need to follow:

1. You are given a univariate regression data set, which contains 180 training data points in the file named `hw05_data_set_train.csv` and 180 test data points in the file named `hw05_data_set_test.csv`.
2. Implement a decision tree regression algorithm using the following pre-pruning rule: If a node has P or fewer data points, convert this node into a terminal node and do not split further, where P is a user-defined parameter. (40 points)
3. Learn a decision tree by setting the pre-pruning parameter P to 30. Draw training data points/test data points, and your fit in the same figures. Your figure should be like the following figures. (20 points)



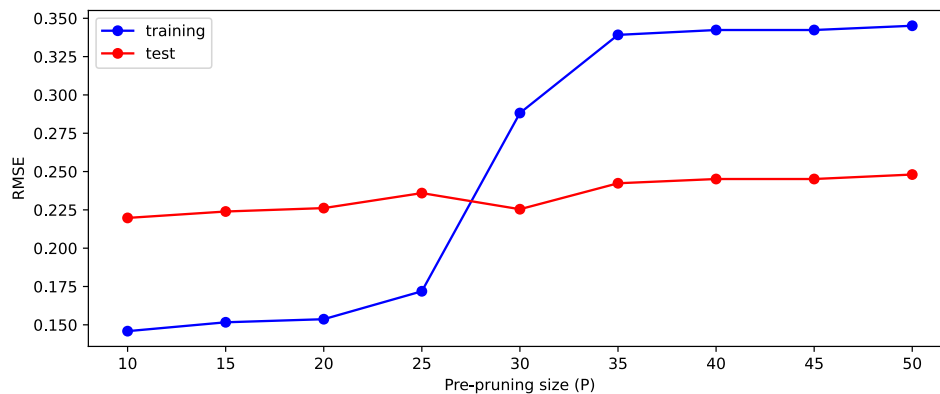
4. Calculate the root mean squared error for training and test data points. The formula for RMSE can be written as

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

Your output should be like the following sentences. (20 points)

```
RMSE on training set is 0.2881929342801884 when P is 30
RMSE on test set is 0.22542236010134897 when P is 30
```

5. Learn decision trees by setting the pre-pruning parameter P to 10, 15, ..., 50. Draw RMSE for training and test data points as a function of P . Your figure should be like the following figure. (20 points)



What to submit: You need to submit your source code in a single file (.py file) named as **STUDENTID.py**, where **STUDENTID** should be replaced with your 7-digit student number.

How to submit: Submit the file you created to Blackboard. Please follow the exact style mentioned and do not send a file named as **STUDENTID.py**. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.