ENGR 421/DASC 521: Introduction to Machine Learning
Homework 4: Decision Tree Regression
Deadline: December 12, 2022, 11:59 PM

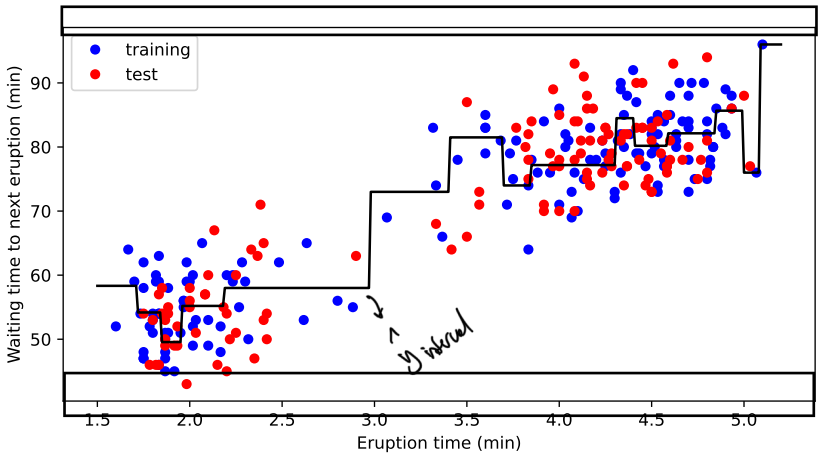
In this homework, you will implement a decision tree regression algorithm in Python. Here are the steps you need to follow:

1. You are given a univariate regression data set, which contains 272 data points about the duration of the eruption and waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA (`https://www.yellowstonepark.com/things-to-do/about-old-faithful`), in the file named `hw04_data_set.csv`.

2. Divide the data set into two parts by assigning the first 150 data points to the training set and the remaining 122 data points to the test set.

3. 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)

4. Learn a decision tree by setting the pre-pruning parameter P to 25. Draw training data points, test data points, and your fit in the same figure. Your figure should be similar to the following figure. (20 points)



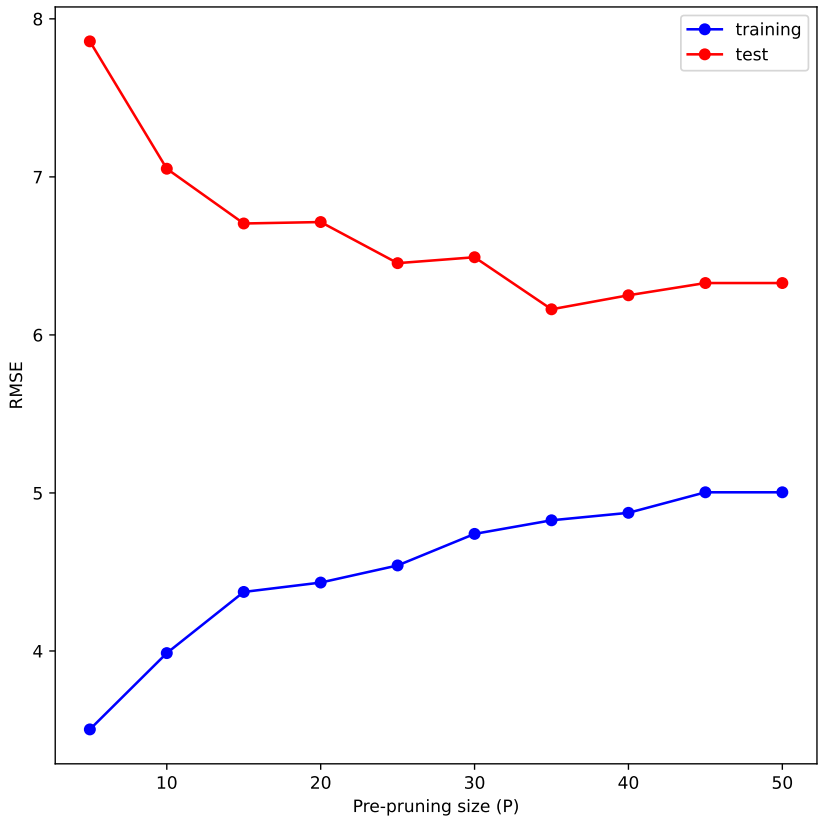
5. Calculate the root mean squared error for training and test data points. The formula for RMSE can be written as

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

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

```
RMSE on training set is 4.541214189194451 when P is 25
RMSE on test set is 6.454083413352087 when P is 25
```

6. Learn decision trees by setting the pre-pruning parameter P to $5, 10, 15, \ldots, 50$. Draw RMSE for training and test data points as a function of P. Your figure should be similar to 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.
