ENGR 421/DASC 521: Introduction to Machine Learning
Homework 3: Nonparametric Regression
Deadline: December 5, 2022, 11:59 PM

In this homework, you will implement three nonparametric regression algorithms in Python. Here are the steps you need to follow:

1. Read Section 8.8 from the textbook.

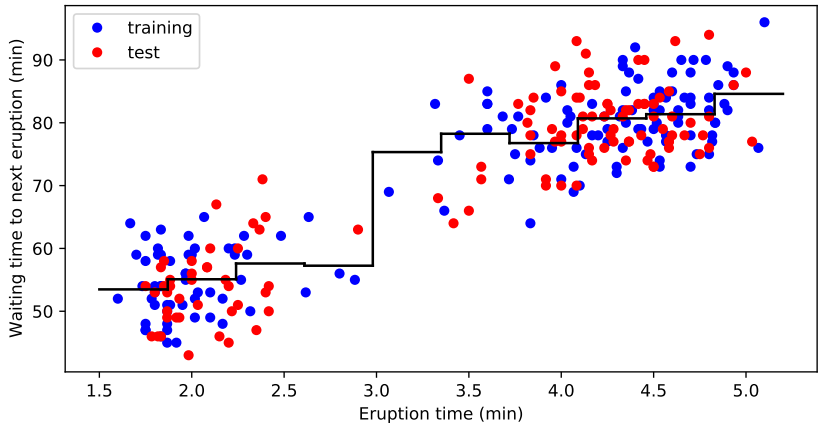
2. You are given a univariate regression data set, which contains 272 data points about the duration of the eruption and waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA (`https://www.yellowstonepark.com/things-to-do/about-old-faithful`), in the file named `hw03_data_set.csv`.

3. 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. (10 points)

4. Learn a regressogram by setting the bin width parameter to 0.37 and the origin parameter to 1.5.

$$g(x) = \frac{\sum_{i=1}^{N_{train}} b(x, x_i) y_i}{\sum_{i=1}^{N_{train}} b(x, x_i)} \qquad \text{where } b(x, x_i) = \begin{cases} 1 & \text{if } x_i \text{ is in the same bin with } x \\ 0 & \text{otherwise} \end{cases}$$

   Draw training data points, test data points, and your regressogram in the same figure. Your figure should be similar to the following figure. (20 points)



5. Calculate the root mean squared error (RMSE) of your regressogram for test data points. The formula for RMSE can be written as

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

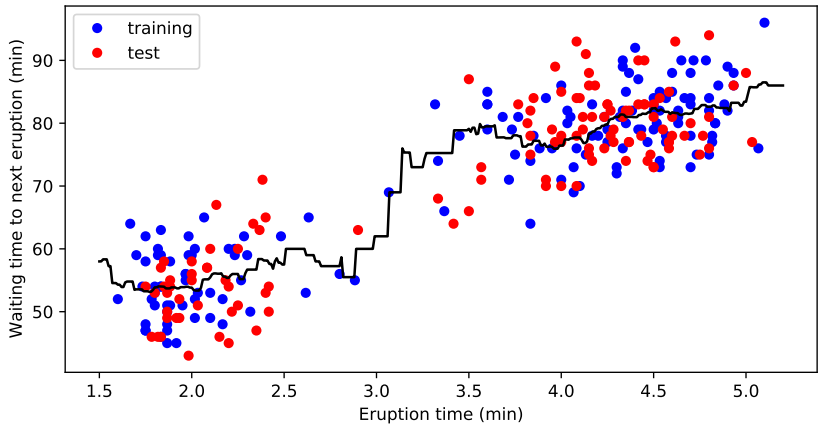
   Your output should be similar to the following sentence. (10 points)

```
Regressogram => RMSE is 5.962617204275405 when h is 0.37
```

6. Learn a running mean smoother by setting the bin width parameter to 0.37.

$$g(x) = \frac{\sum_{i=1}^{N_{train}} w\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{N_{train}} w\left(\frac{x - x_i}{h}\right)} \qquad \text{where } w(u) = \begin{cases} 1 & \text{if } |u| \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Draw training data points, test data points, and your running mean smoother in the same figure. Your figure should be similar to the following figure. (20 points)



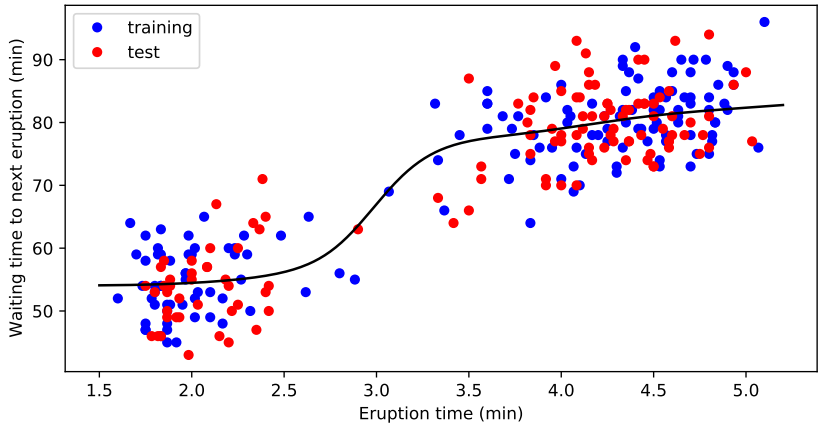
7. Calculate the RMSE of your running mean smoother for test data points. Your output should be similar to the following sentence. (10 points)

```
Running Mean Smoother => RMSE is 6.089003211720321 when h is 0.37
```

8. Learn a kernel smoother by setting the bin width parameter to 0.37 and the origin parameter to 1.5.

$$g(x) = \frac{\sum_{i=1}^{N_{train}} K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{N_{train}} K\left(\frac{x - x_i}{h}\right)} \qquad \text{where } K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

Draw training data points, test data points, and your kernel smoother in the same figure. Your figure should be similar to the following figure. (20 points)



9. Calculate the RMSE of your kernel smoother for test data points. Your output should be similar to the following sentence. (10 points)

```
Kernel Smoother => RMSE is 5.874362846844968 when h is 0.37
```

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.
