DASC521: Introduction to

Machine Learning

Fall 2021

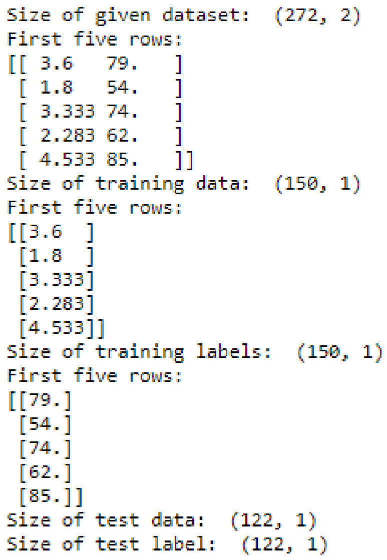
Instructor: Mehmet Gönen

Homework 04- Nonparametric Methods

Gamze Keçibaş- 60211

24.11.2021

There are three main steps in the homework as importing and preparing datasets, writing three functions to make predictions and calculating errors. Firstly, provided dataset that is "*The Old Faithful geyser in Yellowstone National Park, Wyoming, USA"* is imported to script as pandas dataframe. After that, the set is seperated two parts as training and testing set using suggested size as seen below:



```
Size of given dataset:  (272, 2)
First five rows:
[[ 3.6   79.   ]
 [ 1.8   54.   ]
 [ 3.333 74.   ]
 [ 2.283 62.   ]
 [ 4.533 85.   ]]
Size of training data:  (150, 1)
First five rows:
[[3.6  ]
 [1.8  ]
 [3.333]
 [2.283]
 [4.533]]
Size of training labels:  (150, 1)
First five rows:
[[79.]
 [54.]
 [74.]
 [62.]
 [85.]]
Size of test data:  (122, 1)
Size of test label:  (122, 1)
```

Figure 1: Size of datasets and first five rows of first three sets

Secondly, three methods are used to predict Waiting Time to Next Eruption in minutes. The methods are regressogram (1), running mean smoother (2) and kernel smoother (3).

$$g(x) = \frac{\sum_{i=1}^{N} b(x, x_i) y_i}{\sum_{i=1}^{N} b(x, x_i)}$$

where

$$b(x, x_i) = \begin{cases} 1 & \text{if } x_i \text{ is in the same bin with } x \\ 0 & \text{otherwise} \end{cases}$$

(1)

$$g(x) = \frac{\sum_{i=1}^{N} w\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{N} w\left(\frac{x - x_i}{h}\right)}$$

where

$$w(u) = \begin{cases} 1 & \text{if } |u| \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

(2)

$$g(x) = \frac{\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)}$$

where

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{3}$$

Their results are presented below:

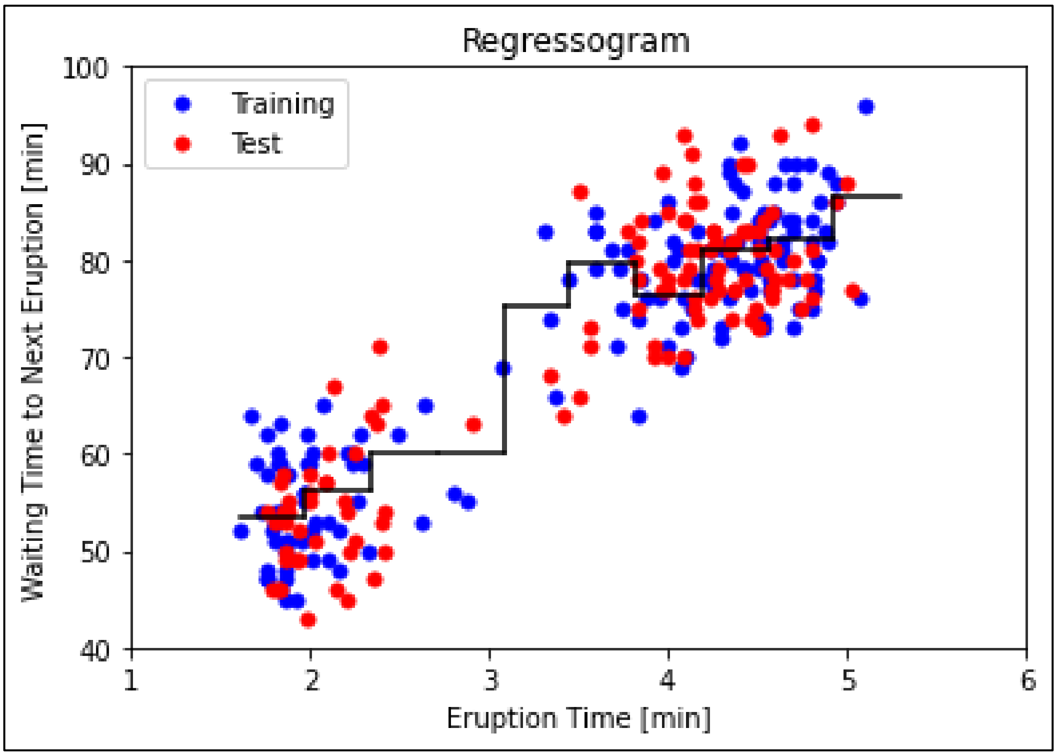


Figure 2: Prediction trend by regressogram where bin width= 0.37 and origin= 1.5

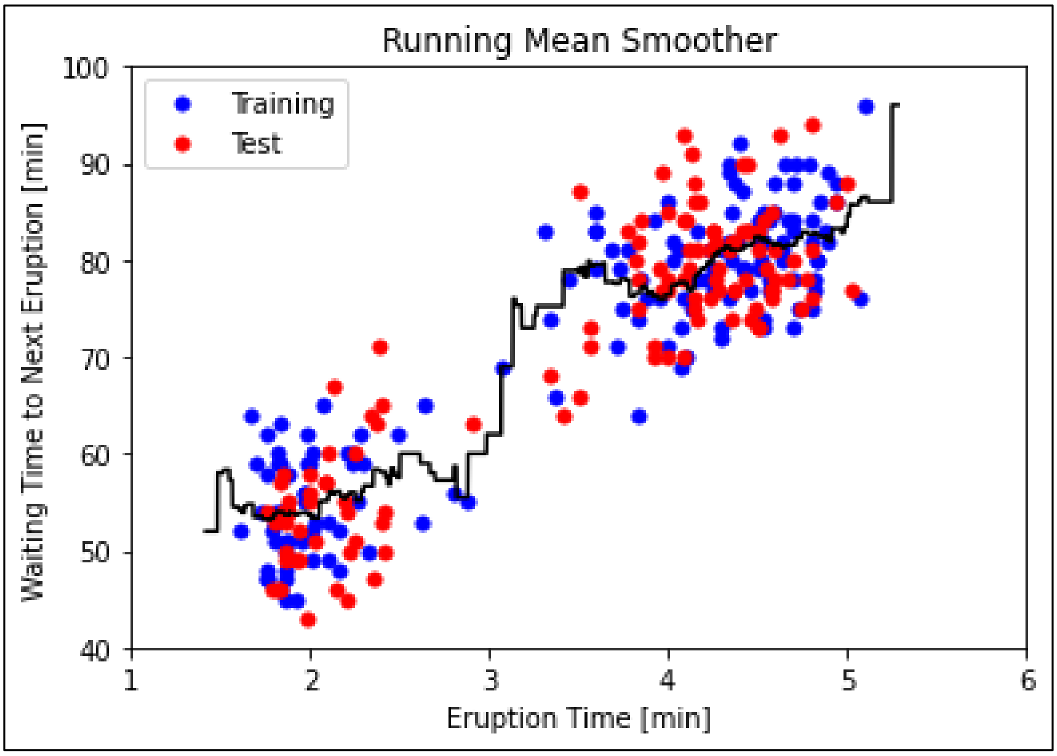


Figure 3: Prediction trend by running mean smoother where bin width= 0.37 and origin= 1.5

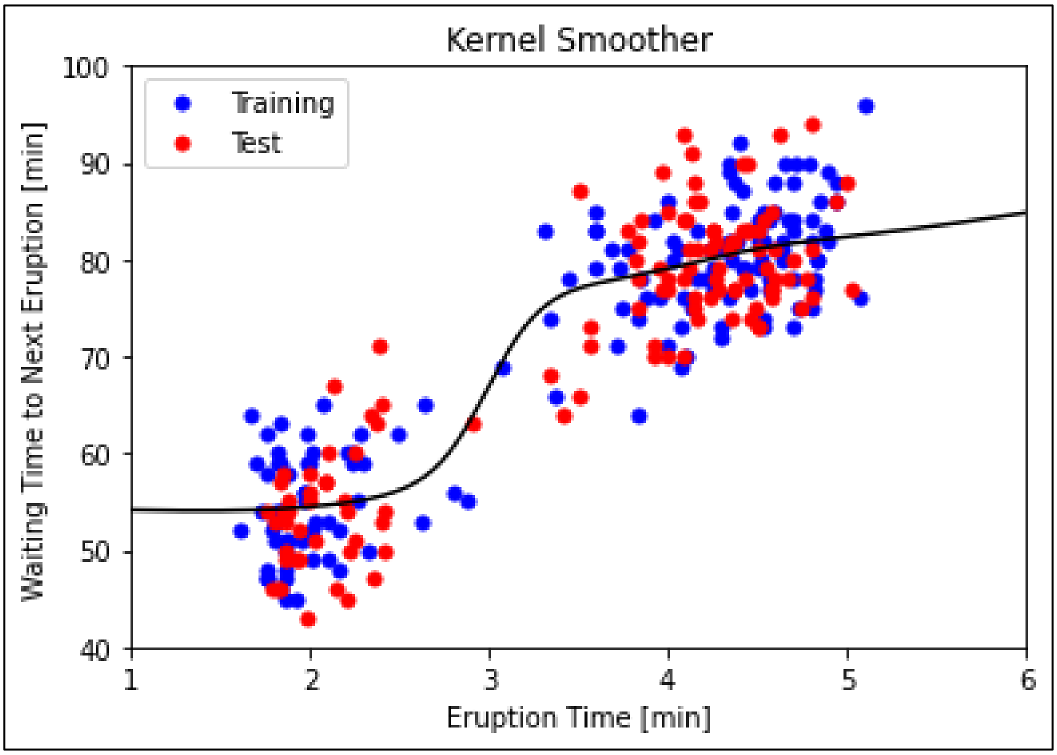
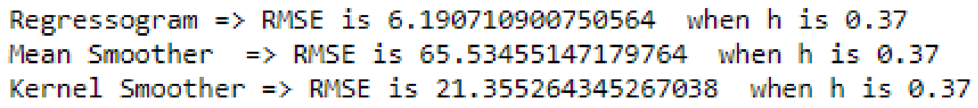


Figure 4: Prediction trend by kernel smoother where bin width= 0.37 and origin= 1.5

Their errors are calculated by Root Mean Squared Error (RMSE) (4) formula:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{4}$$

Calculated errors are represented below Figure. However, there is an issue about RMSE of running mean squared error. I have a trouble to implement running mean squared in my opinion but I could not find the bug. Probably, bin width and origin selections are not appropriate fort he method.



```
Regressogram => RMSE is 6.190710900750564  when h is 0.37
Mean Smoother  => RMSE is 65.53455147179764  when h is 0.37
Kernel Smoother => RMSE is 21.355264345267038  when h is 0.37
```

Figure 5: RMSE values of three methods