ENGR 421 / Homework 2: Discrimination by Regression

Umur Berkay Karakaş

April 19, 2021

In homework 3, we are given 272 data points abot the duration of the eruption and waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park.

First, I created xtraining variable from the first 150 data points in the data set and xtest variable from the last 122 points in the data set. I assigned their corresponding class labels to ytraining and ytest. I also calculated number of classes by using np.max. Then I set bin width to 0.37 and origin to 1.5.

With the corresponding values of bin width and origin, I created arrays of left borders and right borders and then I created regressogram results array by using following formula for each x from origin to maximum value:

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} b(x, x^t) r^t}{\sum_{t=1}^{N} b(x, x^t)} \quad (8.24)$$

where

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

Figure 1: Regressogram score function

Then, I got the following regressogram plot:

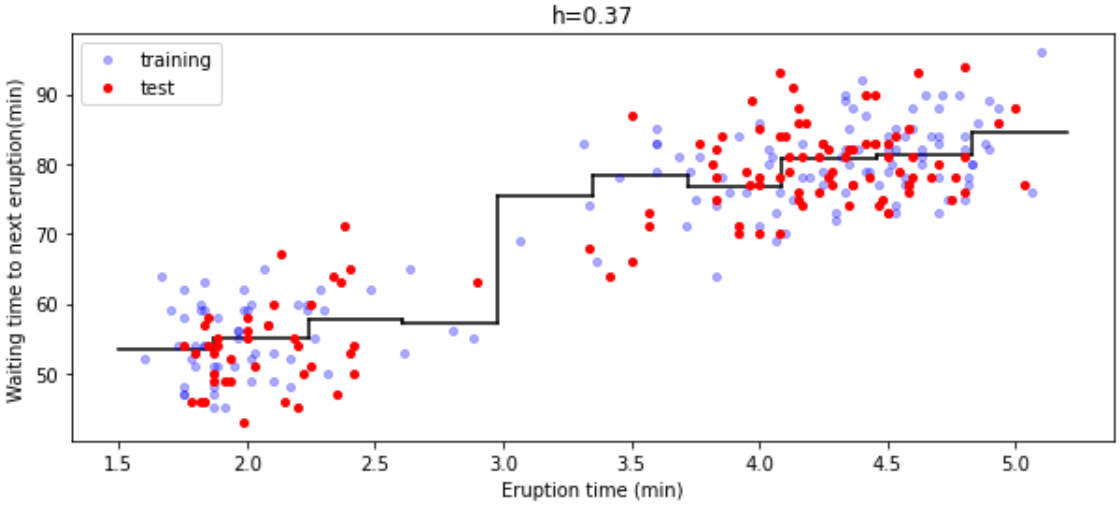

Figure 2: Regressogram plot

Then I created another array which consists of the corresponding score function of each value in xtest and by using the values in the array and ytest, I calculated the RMSE for regressogram:

Regressogram => RMSE is 5.9626 when h is 0.37.

For the RMS smoother, I created an array for the results by using following formula for each x from origin to maximum value:

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} w\left(\frac{x-x^t}{h}\right) r^t}{\sum_{t=1}^{N} w\left(\frac{x-x^t}{h}\right)} \tag{8.25}$$

where

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

Figure 3: RMS score function

Then, I got the following RMS plot:

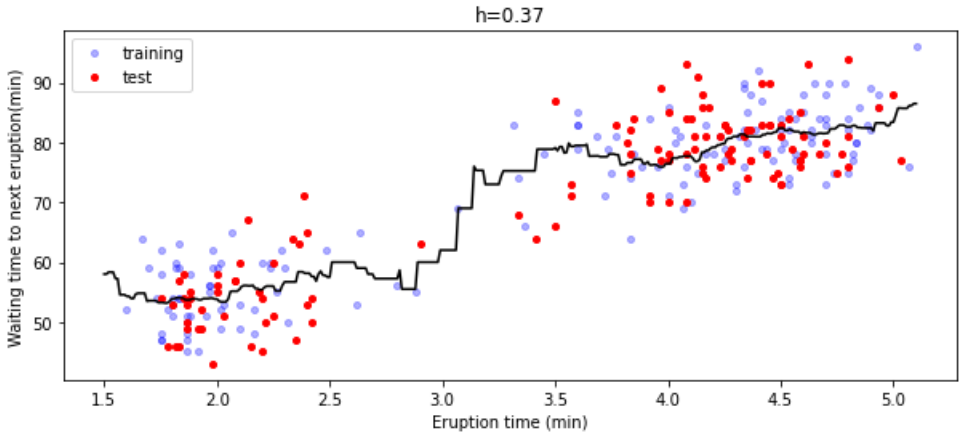

Figure 4: RMS plot

Then I created another array which consists of the corresponding score function of each value in xtest and by using the values in the array and ytest, I calculated the RMSE for RMS:

Running Mean Smoother => RMSE is 6.0890 when h is 0.37.

For the kernel smoother, I created an array for the results by using following formula for each x from origin to maximum value:

$$\hat{g}(x) = \frac{\sum_t K\left(\frac{x-x^t}{h}\right) r^t}{\sum_t K\left(\frac{x-x^t}{h}\right)} \quad (8.26)$$

Figure 5: Kernel smoother score function

Then, I got the following kernel smoother plot:

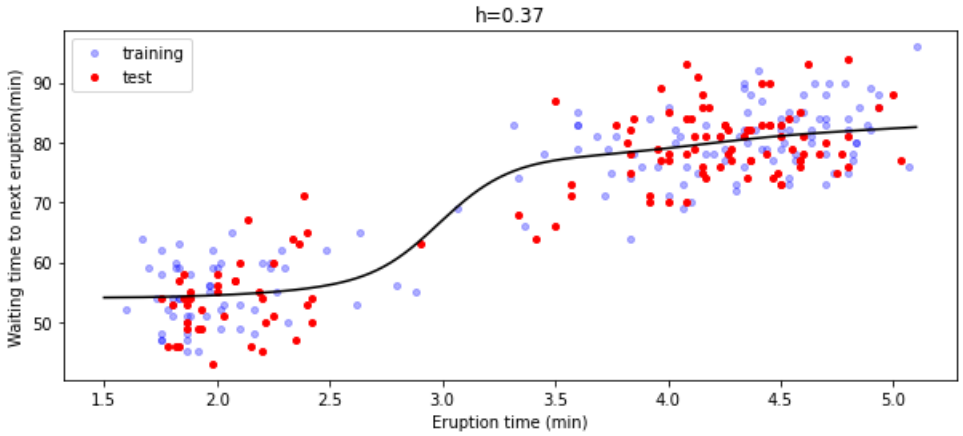

Figure 6: Kernel smoother plot

Then I created another array which consists of the corresponding score function of each value in xtest and by using the values in the array and ytest, I calculated the RMSE for kernel smoother:

Kernel Smoother => RMSE is 5.8744 when h is 0.37.