Assignment 4
DASC521 Fall 2019
Introduction to Machine Learning
Erhan Tezcan 0070881
09.11.2019

1. Task

We 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. We want to find a regression on this data.

2. Implementation

We will be using three different nonparametric regression algorithms:

- Regressogram with bin width $h = 0.37$ and origin 1.5.
- Running Mean Smoother with bin width $h = 0.37$.
- Kernel Smoother with bin width $h = 0.37$ and kernel function as Standart Normal (Gaussian) Distribution.

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

Our error function is Root Mean Squared Error (RMSE) function:

$$E(y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^{N_{test}} (y_i - \hat{y}_i)^2}{N_{test}}} \tag{1}$$

2.1. **Regressogram.** A regressogram has the prediction function defined as:

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} b(x, x^t) r^t}{\sum_{t=1}^{N} b(x, x^t)} \tag{2}$$

where

$$b(x, x^t) = \begin{cases} 1 & \text{if } x^t \text{ is in the same bin with } x \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Our results show that the RMSE for regressogram is 5.962617204275405. The plot is given in figure 1.

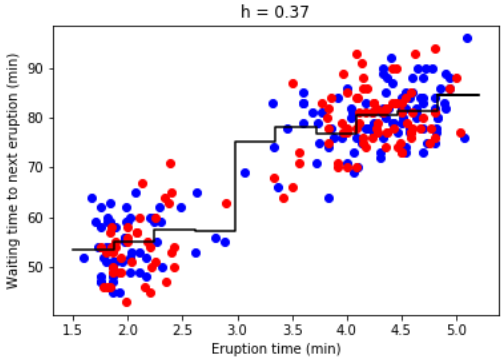
2.2. **Running Mean Smoother.** For the running mean smoother, our prediction function is defined as:

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} w(\frac{x-x^t}{h}) r^t}{\sum_{t=1}^{N} w(\frac{x-x^t}{h})} \tag{4}$$

where h is the bin width and our kernel function w is defined as:

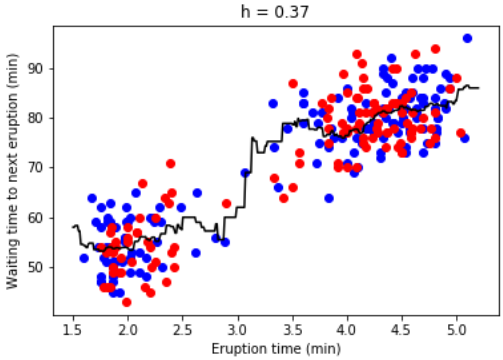
$$w(u) = \begin{cases} 1 & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

FIGURE 1. Regressogram Plot



This function is better in the way that it eliminates the need of an origin parameter. Our results show that the RMSE for Running Mean Smoother is 6.089003211720321. The plot is given in figure 2. For our implementation,

FIGURE 2. Running Mean Smoother Plot



we reformulated the kernel function. Looking at equations (4) and (5), we see that we have this inequality during our calculations:

$$\left|\frac{x - x^t}{h}\right| < 1 \tag{6}$$

For $x - x^t > 0$ we can write this as:

$$x - x^t < h \tag{7}$$

which yields $x^t > x - h$. For $x - x^t < 0$ we can write this as:

$$x^t - x < h \tag{8}$$

which yields $x^t < x + h$. Combining these two inequalities give us:

$$x - h < x^t < x + h \tag{9}$$

Note that this seems to double the bin width, so to compensate for that:

$$x - \frac{h}{2} < x^t < x + \frac{h}{2} \tag{10}$$

This is how we used it in our implementation. Also, we should note that for any NaN value that occurs we set it to be 0.

2.3. **Kernel Smoother.** The Kernel Smoother uses a similar prediction function but the kernel function is different.

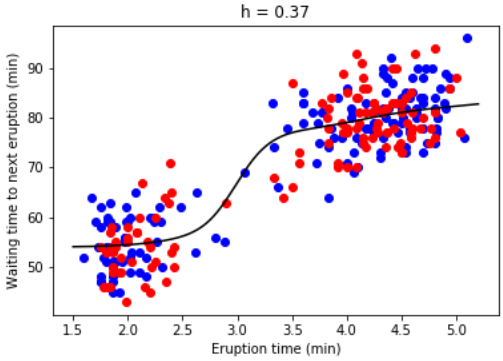
$$\hat{g}(x) = \frac{\sum_{t=1}^{N} K(\frac{x-x^t}{h}) r^t}{\sum_{t=1}^{N} K(\frac{x-x^t}{h})} \tag{11}$$

where h is the bin width and our kernel function K is the standart normal distribution function as a Gaussian kernel, which is:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{12}$$

Our results show that the RMSE for Running Mean Smoother is 5.874362846844968. The plot is given in figure 3. Also, we should note that for any NaN value that occurs we set it to be 0. Overall, the RMSE is smallest for Kernel

FIGURE 3. Kernel Smoother Plot



Smoother, then Regressogram, and the highest RMSE error is for Running Mean Smoother.