COMP 421 – HW04

Cem Ergin
0034296

During this homework, I tried to implement the formula specified in the book's section 8.8; Nonparametric Regression: Smoothing Models.

My RunningMeanSmoother and KernelSmootherMap functions work like the g-hat functions explained in section 8.8. They take a training set, a univariate data point to apply the algorithm and bin width. I used numpy's vectorize feature with these functions to apply these two functions the test data.

I did something different while trying to implement Regressogram function while trying to use my work and my computer's work more efficient. Since bins are predetermined by the user and the average y value of any bin won't change after training. I created a Regressogram function that takes training data, bin width, maximum value for bins and minimum value for bins to output the average y value for each bin as an array and the starting value of each bin. By storing these data, I can just check which bin any testing data point falls into using the bins array and return the average y-value from the Regressogram array.

I couldn't device any simplification technique like this for the other two but all my results seem to be in accord with the ones shown on the homework text. I observed that the results found, vary tremendously according the training data set. When there are data taken from every bin and training data represents the general behavior of the complete data set, the accuracy of regression seems to increase whereas as some less representative data can return really interesting results and high RMSE scores. In my test runs, I got RMSE scores varying between 20 and 30 which verified this observation.

One thing I would like to note here, my Regrossogram function was returning NaN values when there were no points from the training data that fall into a specific bin from time to time. So, I fixed this by setting the y-average value of that bin to the previous bin or to 0 if this happens during the first iteration. After adding this small code, I stopped getting errors about NaN values in my Regrossogram function.