CS 334 – Homework 3

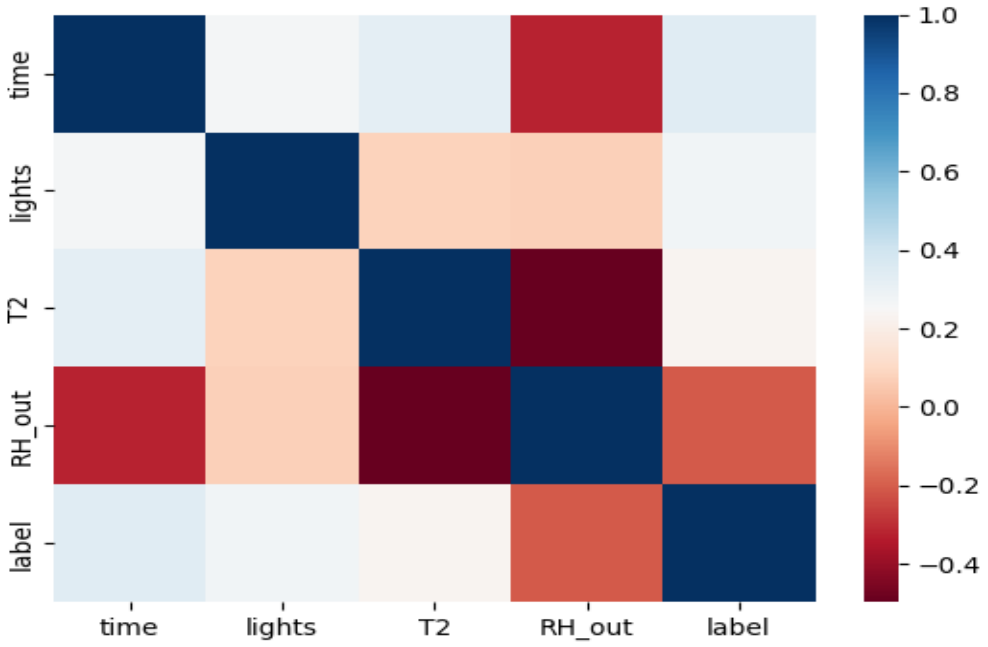
Abdullah Hamid

October 15th 2021

Question 1: Feature Extraction + Feature Selection

a. Time separation of date and time into two features was made. However, the dates and times were converted into a day of the year and into minute of the day respectively. For example, a March 1st date would be written as the 60th day of the year. This way, both features are processable numbers.

b. The heatmap is shown in the figure below.



c. I used RH.out, lights, time and T2 as the features I kept because they had a Pearson correlation of either greater than 0.2 or less than -0.2.

d. I applied MinMaxScaler to the remaining features as my pre-processing step. All code can be found in selFeat.py

Question 2: Linear Regression: Single Unique Solution

The code for this question can be found in lr.py and standarLR.py.

Question 3: Linear Regression using SGD

a. The code for the linear regression using SGD is located in sgdLR.py.

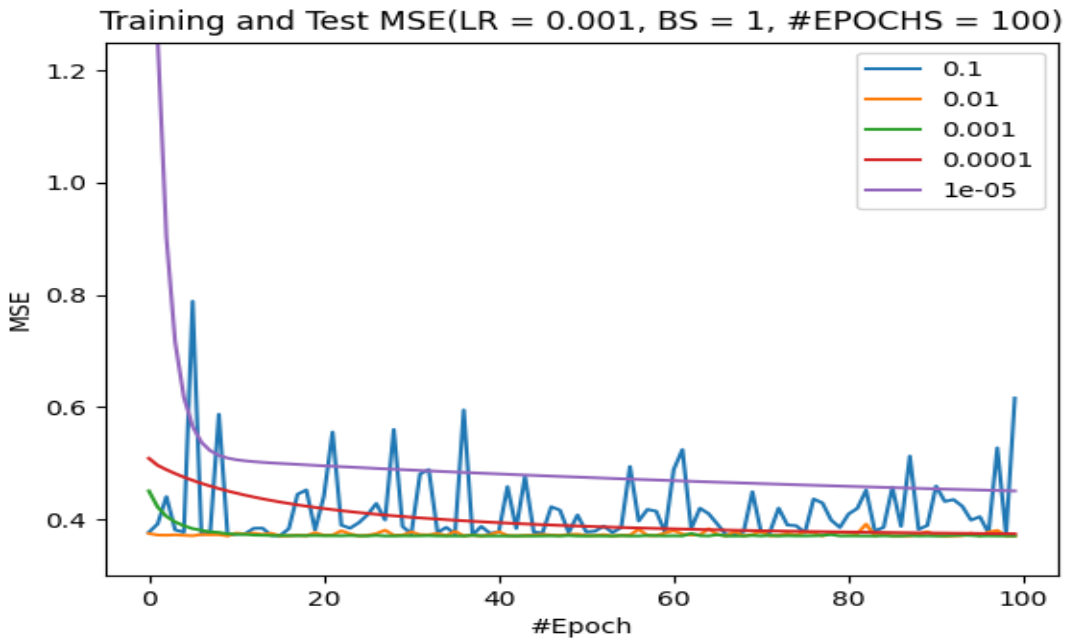
b. I chose these as my parameters

Optimal Learning Rate: 0.001

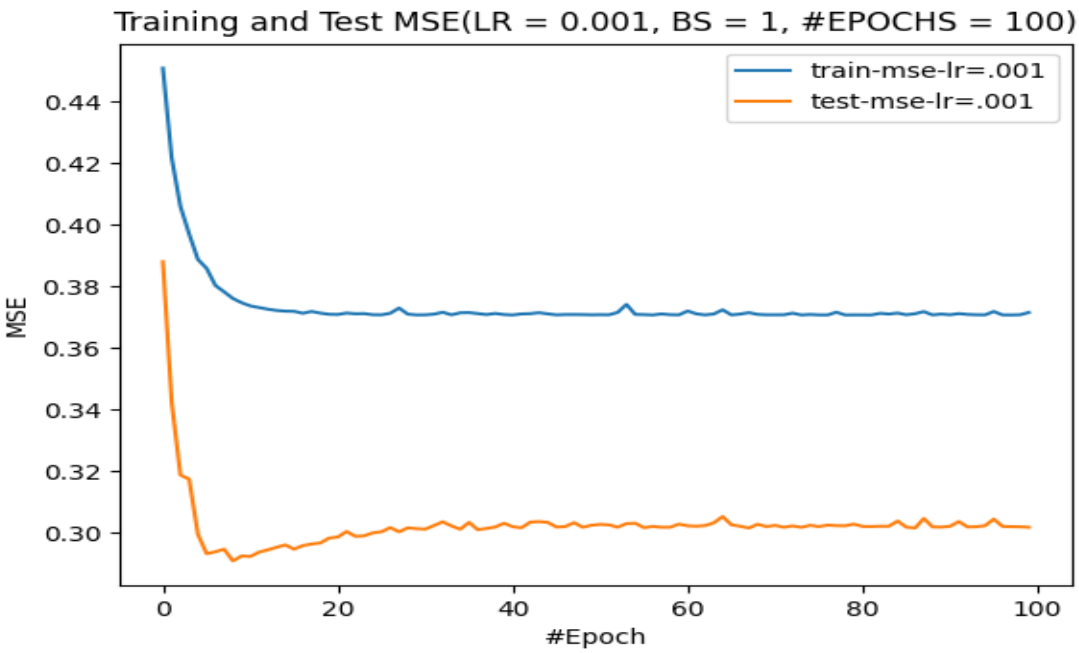
Batch Size: 1

# of Epochs: 100

I chose 0.001 as the optimal learning rate because it had the lowest MSE compared to all other learning rates. Furthermore, the graph looks more cleaner and smoother with 0.001. The code generated plot is shown below.

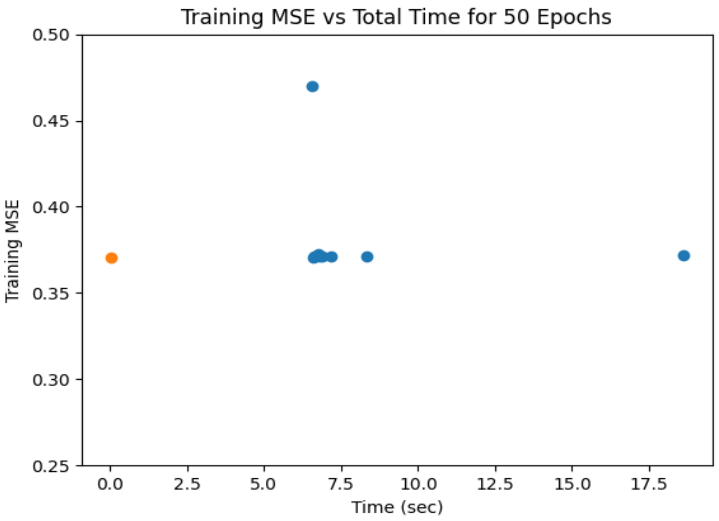


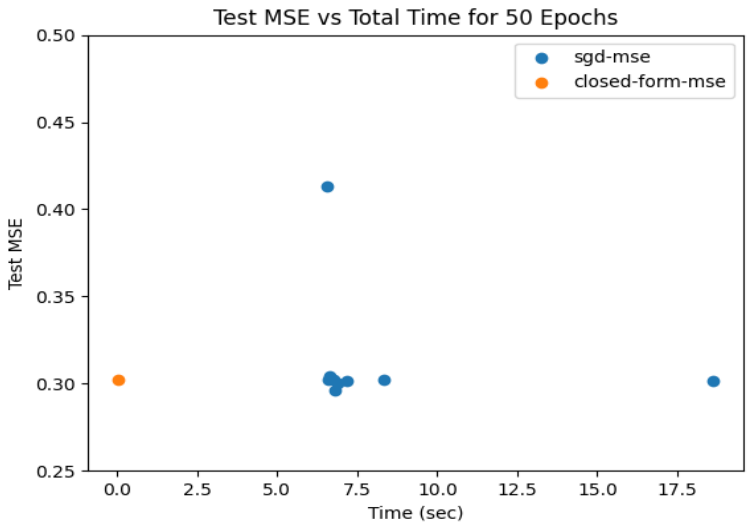
c. The code that generates this plot is located in sgdLR.py



Question 4: Comparison of Linear Regression Algorithms using SGD and closed form solutions.

a. I generated plots for each of the batch sizes which plots curves of the train-mse for different learning rates. The batch sizes were {1,20,40,60,80,120,180,210,240,16770(length of xTrain)}. This code generates 10 plots (commented out in my **sgdLR.py** implementation under method 4A). By using these values, I plotted the total time taken against the train-mse and the test-mse. It also includes the time and MSE's from the closed form solution. All code can be found in **sgdLR.py**. The plots to find the optimal learning rates are not included here because they would take too much space. But if you want to see them, please uncomment part of my method **4A** in **sgdLR.py**.





b. First, the point in the top is for N sized epoch and to the far right is for size 1 epoch. These points are not optimal when compared to the closed form solution or other batch sizes because the MSE is too high, or they take too long to run. From the plots above, we can see that mini batches are much better to get a good MSE when compared to the closed form although they take more time.