CSC 4792 – DATA MINING

PROGRAMMING EXERCISE 2//2019

In this exercise you will carry out regression analysis in order to develop a suitable model that predicts miles per gallon (mpg as y) a given automobile with some amount of horsepower (x) is expected run. You have been provided with a dataset, Auto.csv. The first five (5) data instances when you view the dataset is as follows:

```
-----------------------------------------------------------------------

Number of records: 397


     mpg  cylinders  displacement horsepower  weight  acceleration  year
\
0  18.0          8         307.0         130    3504          12.0    70
1  15.0          8         350.0         165    3693          11.5    70
2  18.0          8         318.0         150    3436          11.0    70
3  16.0          8         304.0         150    3433          12.0    70
4  17.0          8         302.0         140    3449          10.5    70

   origin                         name
0        1  chevrolet chevelle malibu
1        1          buick skylark 320
2        1         plymouth satellite
3        1              amc rebel sst
4        1                ford torino
--------------------------------------------------------------------------
```

Complete the following tasks:

1. Load the dataset for your analysis
2. Preprocess the data (If there is a need)
3. Plot the datapoints (*mpg* vs. *horsepower*) to get sense of the distribution.
4. Using the closed-form solution from programming exercise I, find the weight vector $\boldsymbol{w}$ of parameters and write down the model in this form: $h(x_i) = w_0 + w_1 x_i$
5. Using the sum of squared errors function compute the in-sample error with the parameters in 4 above.
6. Plot the model found to visualize how it fits the training data.
7. If you think there is a problem with the way the model is fitting the training data, see if you can find ways of solving it.

Link to the dataset: https://drive.google.com/open?id=1FtTkLVYIR6UYOtcJDbNZzLITqppotqB2