Machine Learning Assignment1
Linear Regression

Lovejeet Singh Parihar
(IMT2019048)

3 October 2021

1 Evaluation metrics used

The metric used for evaluating the performance of the model is mean squared error.

2 Data Preprocessing

2.1 Imports and visualisation of structure

We import the numpy, pandas and matplotlib libraries.
The data has 7385 rows and 12 columns with visibly no outliers and no invalid value.

2.2 Handling Null values

There are no null values in the data.

2.3 Handling Duplicate rows

There are 1103 rows that are duplicate of some other row. We will remove these extra rows and we have 6282 rows left.

2.4 Label Encoding the Model values

Model feature has about 2000 different labels, so one hot encoding is not feasible, hence we will Label Encode it.

2.5 Encoding non-numeric categorical data

We use one hot encoding for encoding the non-numeric features except Model as it gives better results than when we label encode them.

2.6 Normalisation

The normalisation for features is done here is done by subtracting the mean and dividing by the standard deviation.

It was crucial because without it, it was very difficult to set the hyperparameters in a way that achieves convergence. The learning rate had to be very small (even a small increase caused cost function to explode). But with small learning rates, even with large number of iterations, the cost was failing to converge.

2.7 Splitting into test and training set

The split was according to the parameters mentioned in the assignment pdf, i.e.

```
X_train ,X_test ,y_train ,y_test=
train_test_split(X,y,test_size=.20, random_state=5)
```

3 Correlation

There is a very high positive correlation of features like Fuel Consumption Comb (L/100 km), Fuel Consumption Hwy (L/100 km) and Fuel Consumption City (L/100 km).

Fuel Consumption Comb (mpg) has negative correlation with all other numeric quantities present.

4 Closed Form Solution

- For all features :

  MSE : 25.111016043996045

- For Transmission :

  MSE : 2667.30736471424

- For Fuel Type :

  MSE : 3367.40794772957

- For Fuel Consumption Comb (L/100 km) :

  MSE : 559.2187681792115

- For Engine Size(L) :

  MSE : 913.5233978600984

- For Cylinders :

  MSE : 1113.7201755521685

- For the combination of Transmission, Fuel Type, Fuel Consumption Comb (L/100 km), Engine Size(L), Cylinders

  MSE : 32.5580044075268

5 Gradient Descent Solution

- For all features :

  MSE : 25.072623695350575

- For Transmission :

  MSE : 2671.714707478549

- For Fuel Type :

  MSE : 3367.4079271962305

- For Fuel Consumption Comb (L/100 km) :

  MSE : 559.2187681792111

- For Engine Size(L) :

  MSE : 913.523397860098

- For Cylinders :

  MSE : 1113.7201755521694

- For the combination of Transmission, Fuel Type, Fuel Consumption Comb (L/100 km), Engine Size(L), Cylinders

  MSE : 32.55683787247225

6 Newton's Method Solution

- For all features :

  MSE : 25.1110160439795

- For Transmission :

  MSE : 2667.307364714239

- For Fuel Type :

  MSE : 3367.4079477295704

- For Fuel Consumption Comb (L/100 km) :

  MSE : 559.2187681792112

- For Engine Size(L) :

  MSE : 913.5233978600982

- For Cylinders :

  MSE : 1113.720175552169

- For the combination of Transmission, Fuel Type, Fuel Consumption Comb (L/100 km), Engine Size(L), Cylinders

  MSE : 32.5580044075277

7 Some observations

- Normalisation caused the gradient descent to converge in lesser number of iterations.
- The number of iterations using Newton's method for gradient descent is significantly lower than when we don't use it.
- For this case, the answer we get with multiple features is far better than choosing a single feature for prediction.
- The solutions from all the models are almost equal.
- The five selected features gave a fairly good answer.