Machine Learning Assignment 1

Vaibhavi Lokegaonkar

December 31, 2021

1 Predicting the price of an automobile (Regression)

In this task, given all the details of an automobile, we need to predict the price of an automobile. The dataset given contains all these details along with the price of each automobile. We use the linear regression model to predict the price of an automobile given it's details.

1.1 Data Exploratory Analysis

Although currently, all the details of the automobile are considered to predict the price, we can further increase the model accuracy by deleting some columns. The columns that are deleted have very little or no dependence on the price of the automobile. We can determine this correlation by using a correlation matrix.

Correlation Matrix

	car_ID	symboling	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
car_ID	1.00	-0.15	0.13	0.17	0.05	0.26	0.07	-0.03	0.26	-0.16	0.15	-0.02	-0.20	0.02	0.01	-0.11
symboling	-0.15	1.00	-0.53	-0.36	-0.23	-0.54	-0.23	-0.11	-0.13	-0.01	-0.18	0.07	0.27	-0.04	0.03	-0.08
wheelbase	0.13	-0.53	1.00	0.87	0.80	0.59	0.78	0.57	0.49	0.16	0.25	0.35	-0.36	-0.47	-0.54	0.58
carlength	0.17	-0.36	0.87	1.00	0.84	0.49	0.88	0.68	0.61	0.13	0.16	0.55	-0.29	-0.67	-0.70	0.68
carwidth	0.05	-0.23	0.80	0.84	1.00	0.28	0.87	0.74	0.56	0.18	0.18	0.64	-0.22	-0.64	-0.68	0.76
carheight	0.26	-0.54	0.59	0.49	0.28	1.00	0.30	0.07	0.17	-0.06	0.26	-0.11	-0.32	-0.05	-0.11	0.12
curbweight	0.07	-0.23	0.78	0.88	0.87	0.30	1.00	0.85	0.65	0.17	0.15	0.75	-0.27	-0.76	-0.80	0.84
enginesize	-0.03	-0.11	0.57	0.68	0.74	0.07	0.85	1.00	0.58	0.20	0.03	0.81	-0.24	-0.65	-0.68	0.87
boreratio	0.26	-0.13	0.49	0.61	0.56	0.17	0.65	0.58	1.00	-0.06	0.01	0.57	-0.25	-0.58	-0.59	0.55
stroke	-0.16	-0.01	0.16	0.13	0.18	-0.06	0.17	0.20	-0.06	1.00	0.19	0.08	-0.07	-0.04	-0.04	0.08
compressionratio	0.15	-0.18	0.25	0.16	0.18	0.26	0.15	0.03	0.01	0.19	1.00	-0.20	-0.44	0.32	0.27	0.07
horsepower	-0.02	0.07	0.35	0.55	0.64	-0.11	0.75	0.81	0.57	0.08	-0.20	1.00	0.13	-0.80	-0.77	0.81
peakrpm	-0.20	0.27	-0.36	-0.29	-0.22	-0.32	-0.27	-0.24	-0.25	-0.07	-0.44	0.13	1.00	-0.11	-0.05	-0.09
citympg	0.02	-0.04	-0.47	-0.67	-0.64	-0.05	-0.76	-0.65	-0.58	-0.04	0.32	-0.80	-0.11	1.00	0.97	-0.69
highwaympg	0.01	0.03	-0.54	-0.70	-0.68	-0.11	-0.80	-0.68	-0.59	-0.04	0.27	-0.77	-0.05	0.97	1.00	-0.70
price	-0.11	-0.08	0.58	0.68	0.76	0.12	0.84	0.87	0.55	0.08	0.07	0.81	-0.09	-0.69	-0.70	1.00

Figure 1: Correlation Matrix for the Automobile Dataset

The table above shows that out of all columns, wheelbase, carlength, carwidth, curbweight, enginesize, horsepower are some of the columns with the highest correlation with

the price column. In their plots with price, the points are scattered almost linearly. Hence, if we consider only these columns, we might get a higher accuracy (lesser error) from the model.

Price distribution

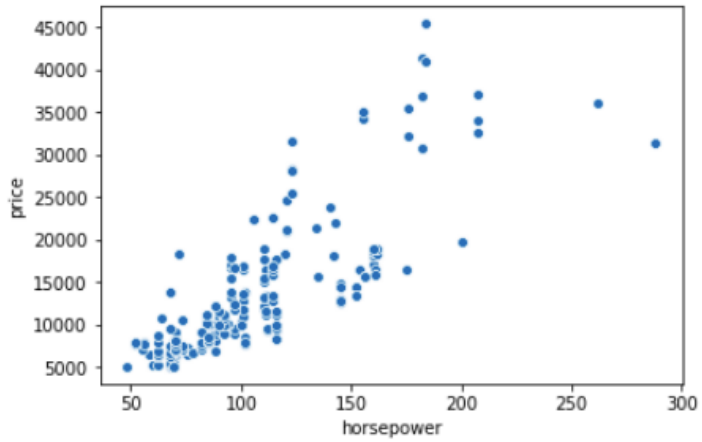

Figure 2: Horse power vs. price

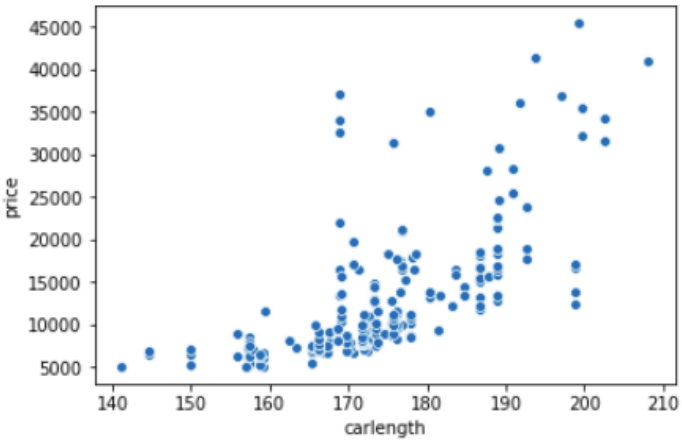

Figure 3: Car length vs. price

1.2 Data Preprocessing

1.2.1 Encoding Categorical Columns

The following columns have non numerical data:

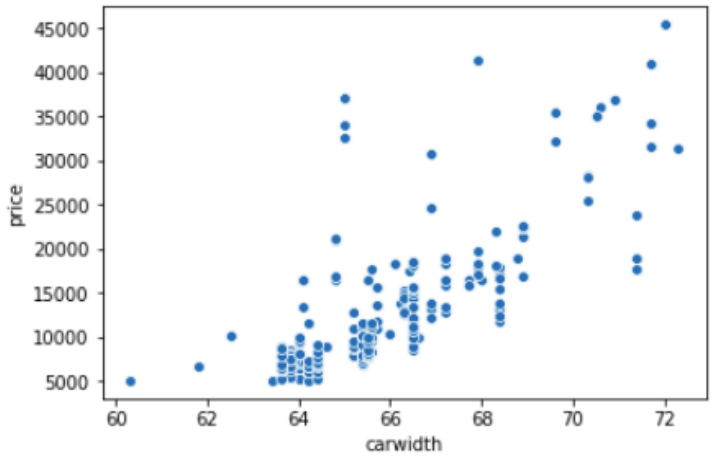

Figure 4: Car width vs. price

1. CompanyName
   We'll have to use one hot encoding due to the high number of unique values.

2. fueltype
   This has only 2 unique values. Hence, we can use label encoding in this column.

3. aspiration
   This has only 2 unique values. Hence, we can use label encoding in this column.

4. doornumber:
   This has 2 values: two and four. Hence, the column can be encoded to the digits 2 and 4 respectively.

5. carbody
   There are more than 2 unique values in this column. Hence, it has to be one hot encoded.

6. drivewheel
   There are more than 2 unique values in this column. Hence, it has to be one hot encoded.

7. enginelocation
   This has only 2 unique values. Hence, we can use label encoding in this column.

8. enginetype
   There are more than 2 unique values in this column. Hence, it has to be one hot encoded.

9. cylindernumber:
   This again has values in words instead of in numbers. Hence, we could replace the words with their repsective numeral form.

10. fuelsystem
    There are more than 2 unique values in this column. Hence, it has to be one hot encoded.

1.3 Train test split

We shuffle the preprocessed dataset and split it in 2 parts train set and test set. The train set contains 70% of the data, while the test set contains the remaining 30%. We add another column in the X matrix to accomodate for the constant term in linear regression

1.4 Learning Algorithm: Linear Regression

In the dataset, we have 64 features for an input x. The dataset has a total of 205 rows. We know that in linear regression, the weight vector will have 65 values for this dataset, which includes the constant term in the below expression:

$$w_0 + w_1x_1 + w_2x_2 + ... + w_{64}x_{64} = y_{pred}$$

1.4.1 Optimisation By Closed Form

In this method the weights are calulated from the below matrix multiplication:

$$w = X^T X X^T y$$

1.4.2 Optimisation By Gradient Descent

We calculate the weights using the following until it converges.

$$w_{new} = w_{old} - \mu \frac{dJ|}{dw|}_{w_{old}}$$

where, μ is the learning rate and J is the loss function given by:

$$J = \sum_{n=0}^{205} (y_i - y_{pred_i})^2$$

1.4.3 Optimisation By Newton's Method

We calculate the weights using the following until it converges.

$$w_{new} = w_{old} - \frac{J(w)'|}{J(w)''|}_{w_{old}}$$

1.5 Inference Algorithm

The value is predicted using the following relation:

$$w_0 + w_1x_1 + w_2x_2 + ... + w_{64}x_{64} = y_{pred}$$

$$w^T x = y_{pred}$$

We use w calculated from the above methods in order to predict the price of the automobile.

1.6 Accuracy Metrics

We calculate the error by using the mean square error given below:

$$\frac{\sum_{n=0}^{205}(y_i - y_{pred_i})^2}{205}$$

1.7 Results: Accuracy on train and test sets

The following table summarises the results for the model:

Optimization Method	**Train set accuracy**	**Test set accuracy**
Closed Form	899155.415	2673641.216
Gradient Descent	175610757.037	64170722.454
Newton's Method	899155.414	2673624.883

Table 1: Accuracy of the test and train set