Information Engineering and Technology Faculty

German University in Cairo

EX1 Take2 Machine Learning

Name: **Yasmine Walid Ibrahim Helal**

ID: **40-9594**

Introduction

The demand of machine learning in our everyday life is highly increasing. Machine learning is the preferred approach in Speech recognition, Natural language processing, Computer vision, Medical outcomes analysis, Robot control, Computational biology, Sensor networks, etc. There are two types of machine learning; supervised learning which is known as predictive learning and unsupervised learning which is known as descriptive learning.

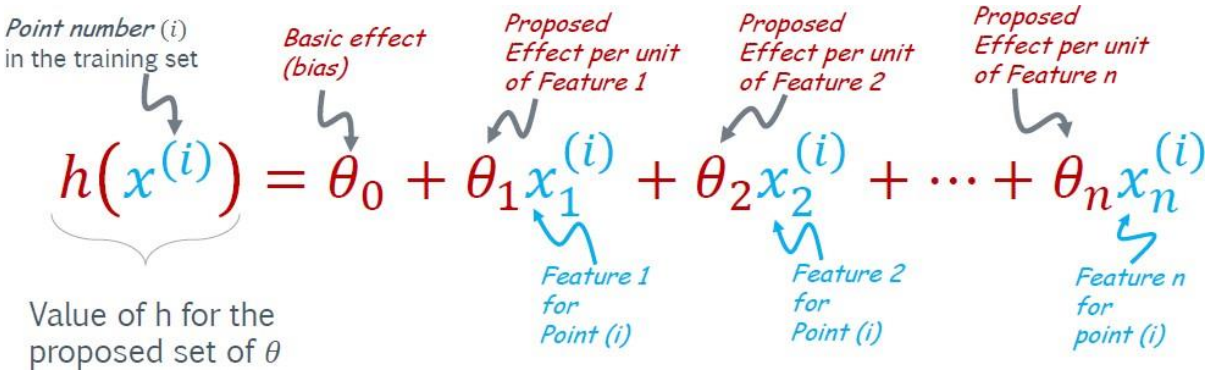
Supervised Learning consists of two types which are regression for continuous value output and classification for discrete value output.

Linear Regression is a regression supervised learning algorithm. Linear Regression is a linear approach to model the relation between a scalar response and one or more explanatory variables (dependent and independent variables). The relationships between the variables are modeled using linear predictor functions whose unknown model parameters are estimated from the data.

In this report we will focus on Linear Regression with multiple variables and its different techniques.

Linear regression with multiple variables techniques and results:

The hypothesis function of linear regression with multiple variable which means we have n features. So we have (x1,x2,...xn) and (θ1,θ2,...θn).



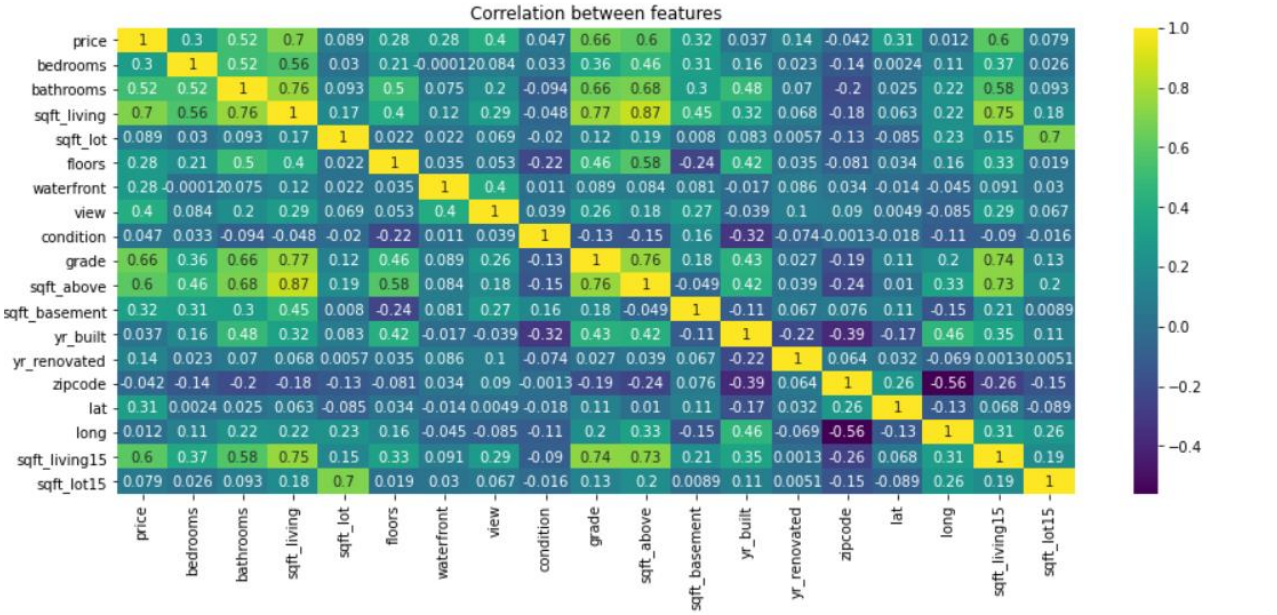
Here are the steps done in the code:

1- At first, the dataset is loaded. The dataset contains of 18 features and the price.
2- The nan values are dropped.

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zip
0	221900.0	3.0	1.00	1180.0	5650.0	1.0	0.0	0.0	3.0	7.0	1180.0	0.0	1955.0	0.0	98
1	538000.0	3.0	2.25	2570.0	7242.0	2.0	0.0	0.0	3.0	7.0	2170.0	400.0	1951.0	1991.0	98
2	180000.0	2.0	1.00	770.0	10000.0	1.0	0.0	0.0	3.0	6.0	770.0	0.0	1933.0	0.0	98
3	604000.0	4.0	3.00	1960.0	5000.0	1.0	0.0	0.0	5.0	7.0	1050.0	910.0	1965.0	0.0	98
4	510000.0	3.0	2.00	1680.0	8080.0	1.0	0.0	0.0	3.0	8.0	1680.0	0.0	1987.0	0.0	98
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
17994	320000.0	2.0	1.00	1802.0	11225.0	1.0	0.0	0.0	3.0	7.0	1802.0	0.0	1961.0	0.0	98
17995	1990000.0	5.0	3.00	4480.0	5000.0	2.5	0.0	0.0	5.0	12.0	3420.0	1060.0	1902.0	0.0	98
17996	253000.0	2.0	1.00	1310.0	7128.0	1.0	0.0	0.0	4.0	7.0	940.0	370.0	1980.0	0.0	98
17997	630000.0	3.0	2.50	2320.0	32772.0	2.0	0.0	0.0	3.0	9.0	2320.0	0.0	1992.0	0.0	98
17998	216000.0	2.0	1.00	1130.0	12500.0	1.0	0.0	0.0	4.0	7.0	1130.0	0.0	1953.0	0.0	98

17999 rows × 19 columns

3- The data correlation is calculated and only the features with correlation with the price which is higher than 0.5 are taken.



4- Split the data into a Training Set (60%), a Cross Validation (CV) Set (20%) and a Test Set (20%).

$$x_i = \frac{x_i - \mu_i}{S_i}$$

```
Train dimensions :(10799, 5)
Cross Validation :(3599, 5)
Test :(3599, 5)
```

5- Feature Normalization of the data each feature is subtracted from the mean and then we divided it by the standard deviation.

6- After the featureNormalize function is tested, we now add the intercept term to normalized

8- functions computeCostMulti and gradientDescentMulti to implement the cost function and gradient descent for linear regression with multiple variables.

9- The gradientDecentMulti is calculated for first, second and third degree with different alphas

Graphs of the training data

- alpha=0.03 number of iterations=100 degree 1

As shown in the graph, by increasing the number of iterations the cost decreased. So increasing iterations in this case is beneficial to minimize cost function for different values of theta but the cost is almost constant starting from iteration 80.

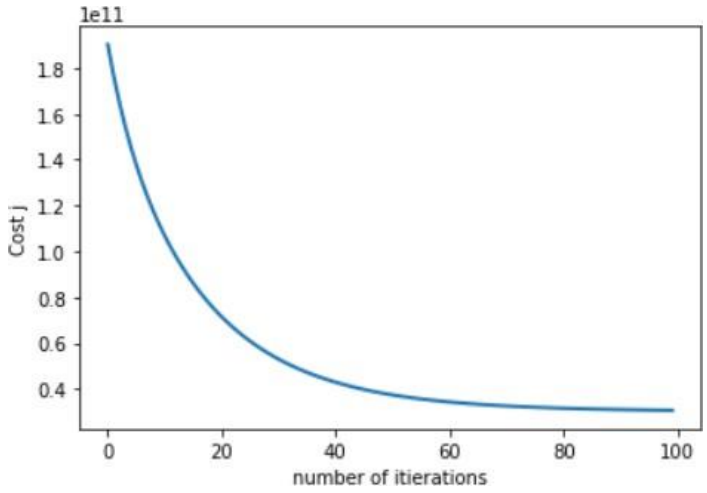
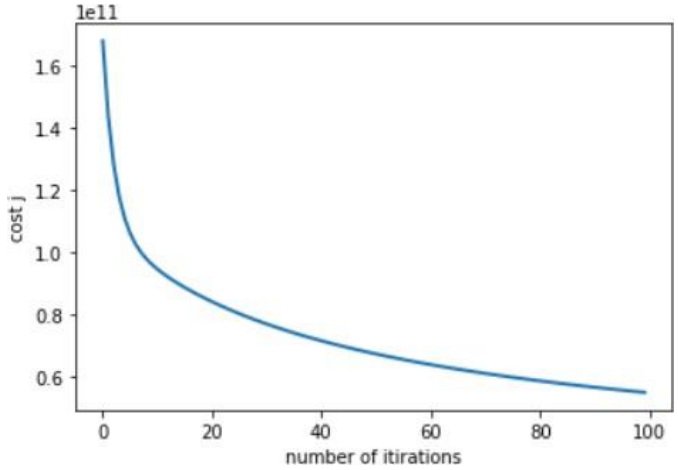


Figure 1 alpha 0.03 degree 1

-
-
-

- alpha 0.01 , degree 2, number of iterations=100



- *Figure 2 alpha 0.01 degree 2*

- The cost function decreases slower than in degree 1.

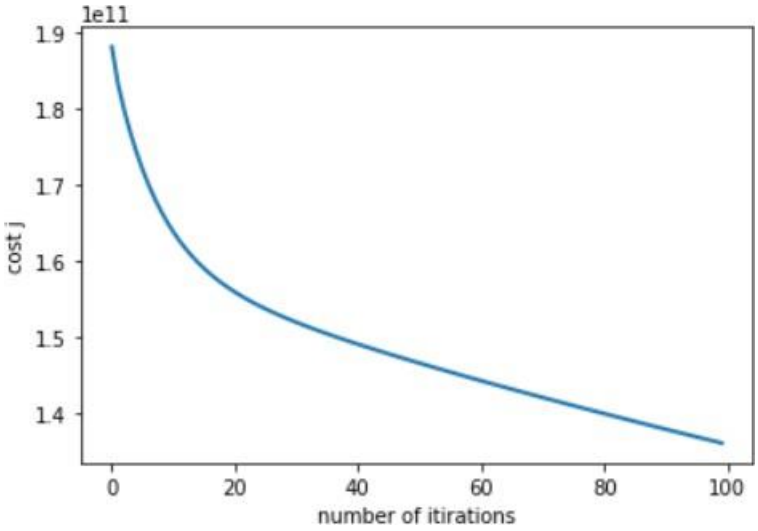
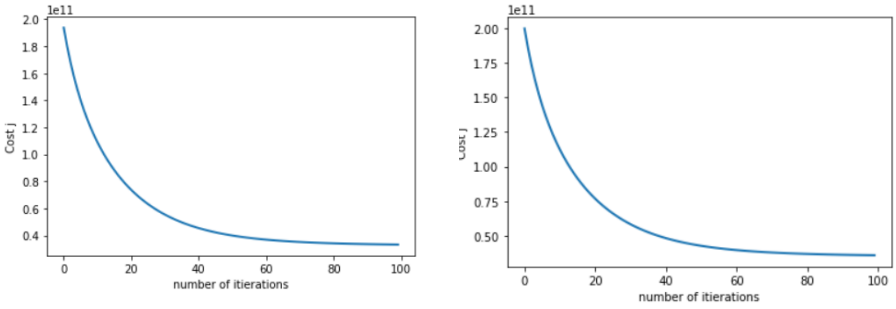


Figure 3 alpha 0.001 degree 3

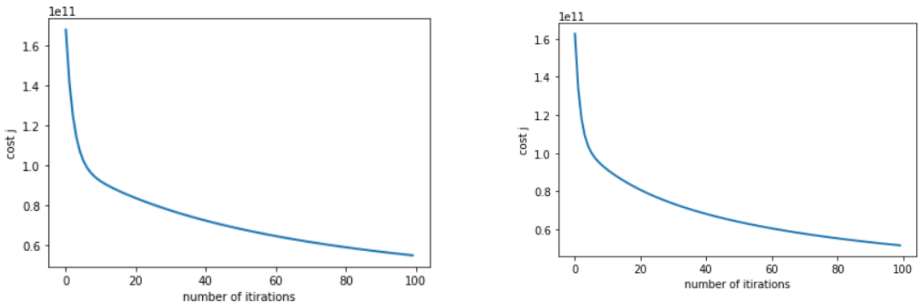
- Increasing the degree of the function and changing alpha didn't improve the cost function .

10- Do the same for validation and test data then plot the graphs.

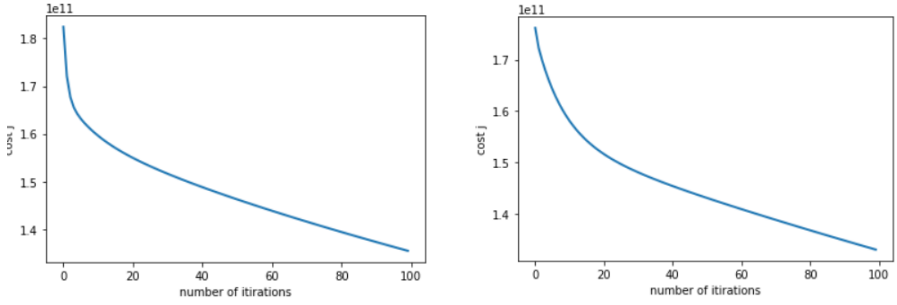
a- alpha=0.03, number of iterations=100, and degree 1 cross validation and test data cost function respectively.



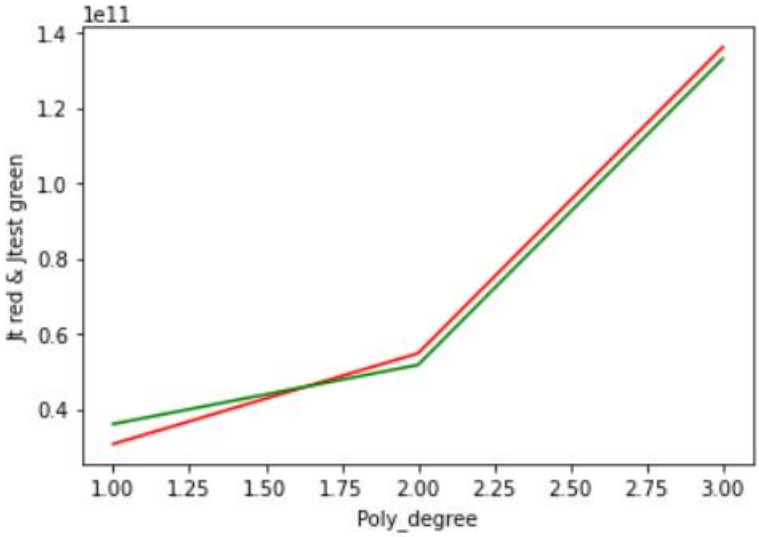
b- alpha=0.01, number of iterations=100, and degree 2 cross validation and test data cost function respectively.



c- alpha=0.001, number of iterations=100, and degree 3 cross validation and test data cost function respectively.



10- Plotting_Jtrain_&_Jtest_against_the_degree



In conclusion, the best hypothesis is the one with degree 1 because the graph shows that it's the one with the minimum cost in relation with the other degrees of the hypothesis.