ML Assignments Report

25.04.2021

—

Salma Saleh
40-8992

1.Introduction

This report discusses applying the diagnostic techniques introduced in lecture 5,6 ,on a new dataset which consists of 18000 rows and 21 columns.Each entry in the dataset represents a house along with its 21 features including its selling price.Since the target we want to estimate is the price which is a continuous variable then the linear regression model would be used.

2.Goals

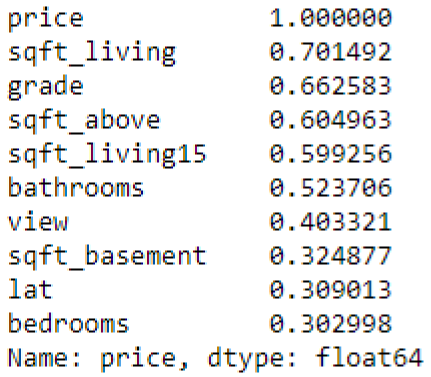
1. Enhancing the testing error then choosing the best combination of theta's and hypothesis degree that minimizes the cross-validation error.
2. Putting the diagnostic methods introduced in the lectures into use to train and test our model efficiently.

3.Specifications

3.1 Data exploring and cleaning

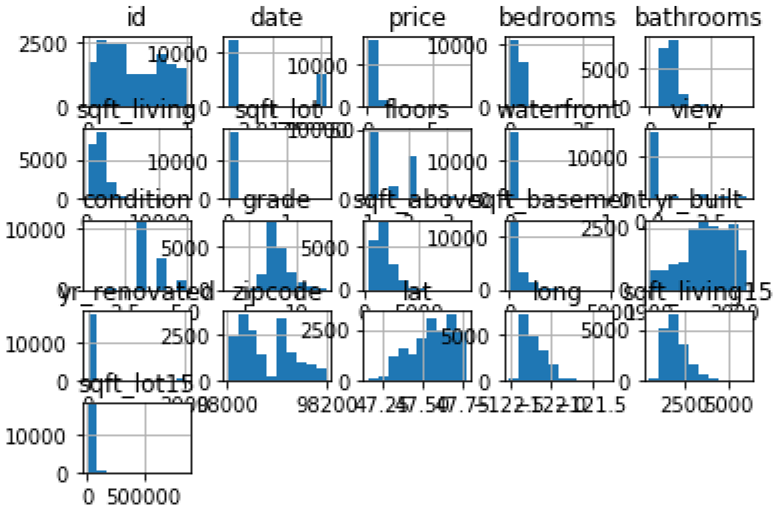
❑ Upon importing the csv file inside the jupyter notebook nan values appeared due to having empty cells ,then this method was used to remove them **`df.dropna().`**

- After analyzing the features the first feature was the id of the house which doesn't really have an effect on estimating the price,so this column was excluded from the X matrix along with the price.
- The price column was placed in the y-vector for it to be our target.
- Further analysis indicated that the number of bathrooms in X was of type float which doesn't make sense so converted it to integer.
- Then just to sense how each feature affects the price by measuring the correlation between the price and the 19 features excluding the id of the house.This was the output of the top ten correlated features with the price.



```
price            1.000000
sqft_living      0.701492
grade            0.662583
sqft_above       0.604963
sqft_living15    0.599256
bathrooms        0.523706
view             0.403321
sqft_basement    0.324877
lat              0.309013
bedrooms         0.302998
Name: price, dtype: float64
```

- Then a histogram was plotted of all the features against the features to check their distribution.



- As we can see all the features are different from each other and each of them should be normalized based on its own scale and unit.

3.2 Data Splitting

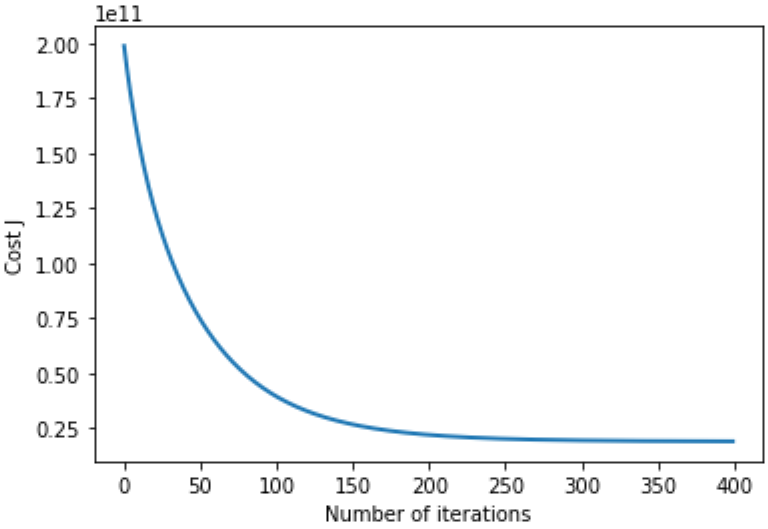
- ❏ If normalization was implemented before splitting would affect the performance of our model as we would know the range of the output it would be biased,with that being said i splitted that data **before** normalization.
- ❏ Now let's get to how the splitting was done .First off we decided to split the data into 3 subsets:The training set,testing set and cross validation set.
- ❏ The k-fold concept is used to ensure that the splitting is done randomly and that the testing set is representative of all the dataset .Then the stratified k-fold cross-validation technique is used to make sure every class in the dataset is involved when training or testing for imbalanced classification **but** since all features have the same size there is no imbalance in the classes ,so only k-fold is used with 10 folds. A part of the data-set is dedicated for cross-validation to ensure that the model is not too optimistic or too pessimistic.
- ❏ The k-fold splits the data differently in each iteration but just to give an overview on how the dataset is splitted. The X_train was given **14580** entries from the dataset,the X_test was given **1799** entries and the X_cv was given **1620** entries**.**
- ❏ Then normalization was implemented on each of the subsets.

3.3 Data Normalization

- ❏ After implementing the feature normalization function on the X matrices from scratch I compared the output from my method to the output of a predefined function that normalizes a given matrix **scaler.fit_transform( )** and they both had the same output just to make sure everything is working perfectly.

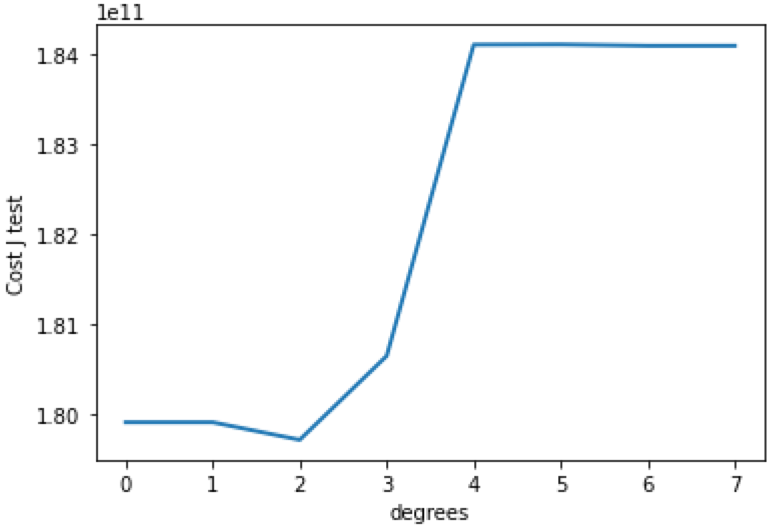
3.4 Data Training

- ❏ The cost function was implemented to calculate the MSE of the model on the training data only.Then gradient Descent functions were called on the X_train for 400 times with learning rate (alpha) equals 0.01 to obtain the best theta's that minimizes the cost function.When plotting the cost functions result against the number of iterations the following figure resulted, which indicates that when increasing the number of iterations the MSE decreases.

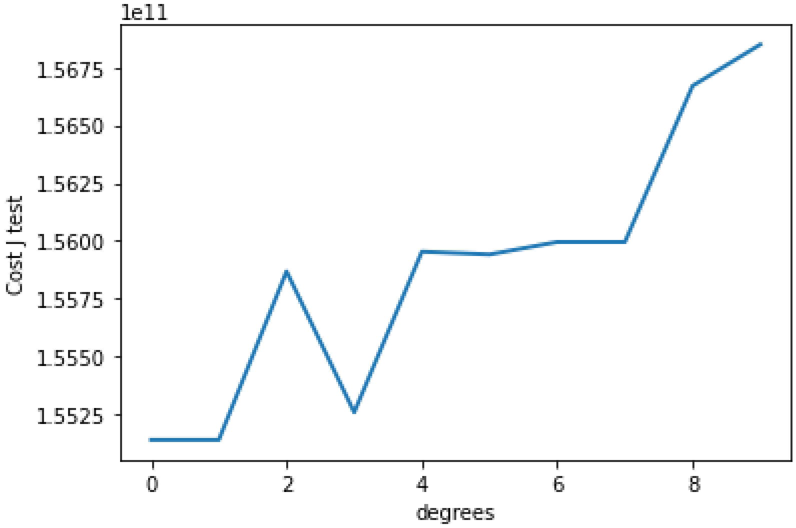


3.5 Data Testing

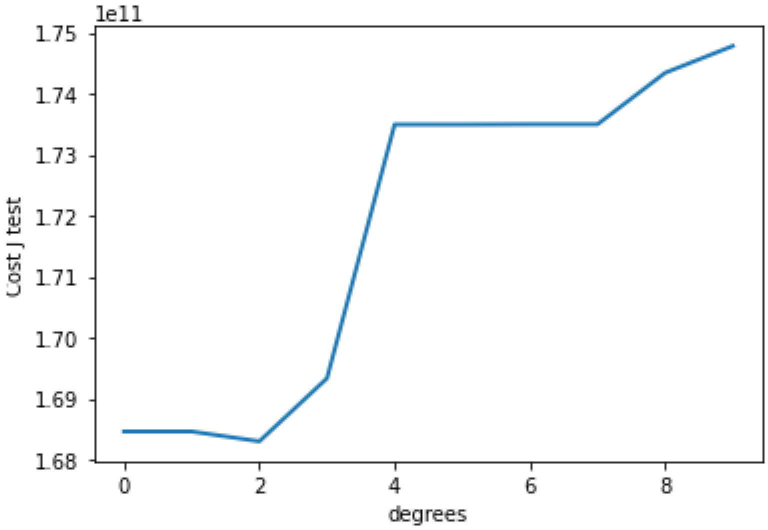
- ❑ Since the model was trained using linear regression , but the model performance isn't the best and it suffers from a relatively high error,so to explore the different degrees of the hypothesis and what is their effect on the cost function.
- ❑ So, a function was implemented that does the following:
  - ❖ It takes 2 parameters: a matrix of any size and an integer representing the degree desired to put the matrix in x_transform(var, degrees).
  - ❖ Inside the function we raise the matrix starting from giving the bias term a power of zero, since the 1st column is ones anyway.Then the 2nd column has a degree of one and so on depending on the degree given to the function.
  - ❖ Then we normalize each column on its own after raising it to its respective power.
  - ❖ After this we call the cost function on the matrix after raising a particular column and normalize it, to measure the error produced from this degree.
  - ❖ Then we plot the array having all the cost function results of each degree against its corresponding degree of the testing subset.The following figure was the result of this method being called x_transform(X_test, 8).



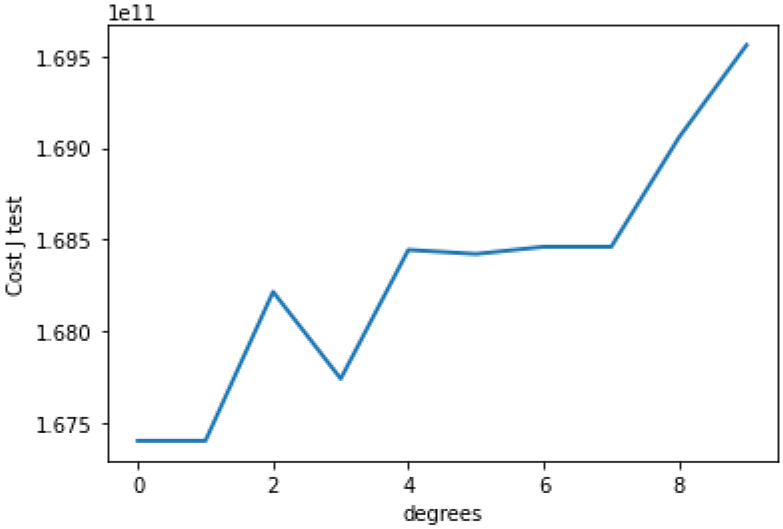
- ❑ From the above figure we conclude that the best degree to choose for our hypothesis is degree 2 to minimize the J-cost of testing.
- ❑ When increasing the range of degree to 10 x_transform(X_test, 10).



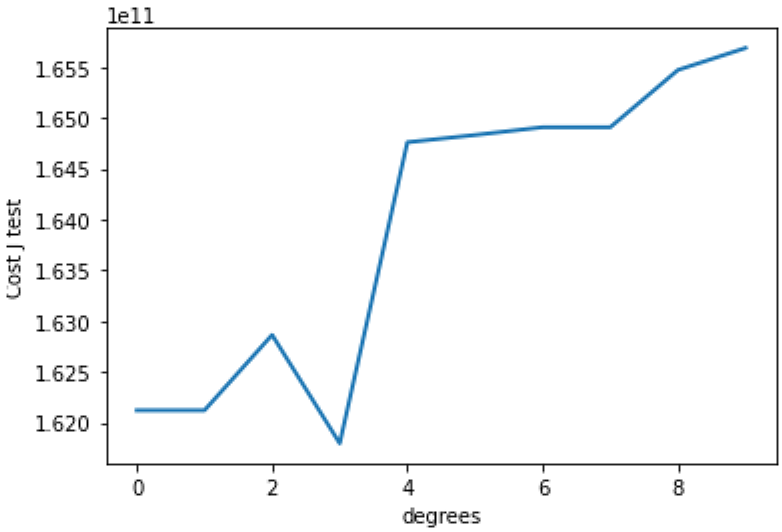
- ❑ When applying the method on the cross validation subset, we got the following figure x_transform(X_cv, 8).



- ❑ But if we increased the range to include a higher degree x_transform(X_cv, 10)



- ❑ When applying the same method but with degree = 10 on the training set we got the following figure.Which indicated that the best degree for hypothesis would be of degree 3 .



- ❑ From the previous analysis we conclude that the best degree for the hypothesis function to avoid overfitting and underfitting is degree 3 .
- ❑ The rest of the assignment is normalization using the closed form equation which won’t be efficient to use since our data this time is very large, so implementing it would be very expensive with the number of theta’s we have.