CS534 – Machine Learning

Homework Assignment 1*

Fall 2018

Ali Raza 933620601 razaa@oregonstate.edu
Sheekha Jariwala 933620505 jariwalas@oregonstate.edu
Vivswan Shitole 933591615 shitolev@oregonstate.edu

Oregon State University
October 013, 2018

* Due 11:59PM Oct 13th, 2018

We worked as a team of three students and contribution of each member is 1/3. We code our implementation in Python 3. The effects of different learning rates, regularization, and normalization are discussed in the following parts.

Part 0

a)

If we include the ID feature in learning, we will be modeling for specific home units. We are not interested in predicting price for specific homes. It will not answer the general question of predicting the houses' price as ID feature is not related in that regards.

b)

Date feature can be separated in multiple ways. We can use any of these time frames:

1. Day of the week
2. Day of the month
3. Day of the year
4. Month of the year
5. Year

Splitting the date of sale of house could be used to analyze the seasonal trends observed according to the year, month and date of the year. Number of reasons like holiday season and employment trend over the year, salary disbursement date over the month are some of the reasons that could be used to identify their effect on sale of houses over the measured time frame.

c)

Required statistics of the features are given below:

Feature Name	Mean	Standard Deviation	Range
id	NA	NA	NA
month	6.59240000e+00	3.11127984e+00	1.100000e+01
day	1.58021000e+01	8.62133027e+00	3.000000e+01
year	2.01431850e+03	4.65894570e-01	1.000000e+00
bedrooms	3.37520000e+00	9.43199321e-01	3.200000e+01
bathrooms	2.11887500e+00	7.65089854e-01	7.250000e+00
sqft_living	2.08022320e+03	9.11288790e+02	9.520000e+03
sqft_lot	1.50892014e+04	4.12018347e+04	1.650787e+06
floors	1.50370000e+00	5.42619858e-01	2.500000e+00
view	2.29400000e-01	7.55893934e-01	4.000000e+00
condition	3.40910000e+00	6.53557335e-01	4.000000e+00
sqft_above	1.79309930e+03	8.30823890e+02	8.490000e+03
sqft_basement	2.87123900e+02	4.34983513e+02	2.720000e+03
yr_built	1.97112490e+03	2.94791197e+01	1.150000e+02
yr_renovated	8.12267000e+01	3.94360085e+02	2.015000e+03
zip code	9.80782931e+04	5.35157154e+01	1.980000e+02
lat	4.75598142e+01	1.38643657e-01	6.217000e-01
long	-1.22213287e+02	1.41397615e-01	1.195000e+00
sqft_living15	1.99432610e+03	6.91865705e+02	5.650000e+03
sqft_lot15	1.27463234e+04	2.82398309e+04	8.705400e+05

Frequency for categorical features is given below:

Number of waterfront	**Frequency**
0	9930
1	70

Grade	**Frequency**
0	0
1	0
2	0
3	0
4	11
5	105
6	933
7	4130
8	2838
9	1182
10	547
11	210
12	39
14	5

Condition	**Frequency**
0	0
1	13
2	76
3	6530
4	2569
5	812

d)

Feature	**Description**	**Importance**
dummy	1	for controlling bias of the model
grade	Overall grade given to the housing unit. 1 poor ,13 excellent.	Grade of house indicates its category. Statistics show the distribution of available houses across various categories on sale. Availability of houses drive the demand-supply chain which in turn influence the price of house.
sqft_above	square footage of house apart from basement	This feature indicates the available living space above ground level. This feature is a better indicator of available living space than the number of bedrooms as it can help buyers estimate availability of multi-purpose space.
sqft_living	square footage of the home	Indicates the area of available land with the house including the garden or backyard area which helps buyers factor in available outdoor space for use with the house.
bathrooms	Number of bathrooms/bedrooms	Standard deviation of bathrooms is one of the least for all the features which will help predict the price uniformly.
waterfront	House which has a view to a waterfront	Presence of waterfront as an accessory influences the price of the house.

yr_renovated	Year when house was renovated (0 if n/a)	Values of this feature is either the year in which the house was renovated or 0 instead of a default reference year. The mean generated as 81 doesn't correctly help determine the sale price of the house.
yr_built	Built Year	Relation between year _built and year_renovation (if any) is a better indicator of sale price of house rather than year_built or year_renovation alone.

e)

NORMALIZATION OF FEATURES BETWEEN 0 TO 1:

As all numeric features in the dataset have drastically different scales and ranges, normalization of data is done to scale all the features to lie in the range of 0 to 1.

Formula used for normalization of numerical features is follows:

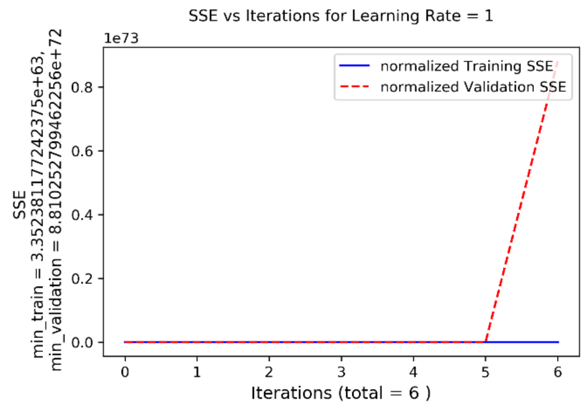
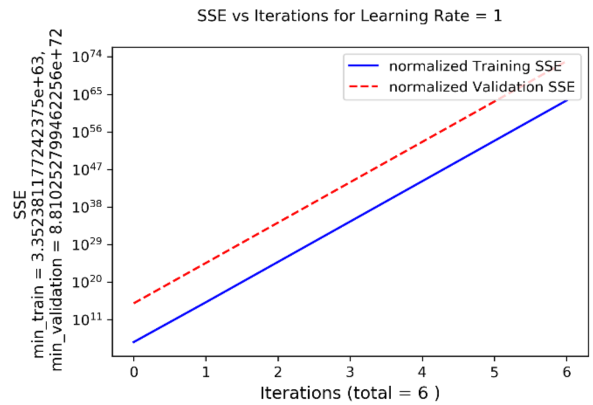
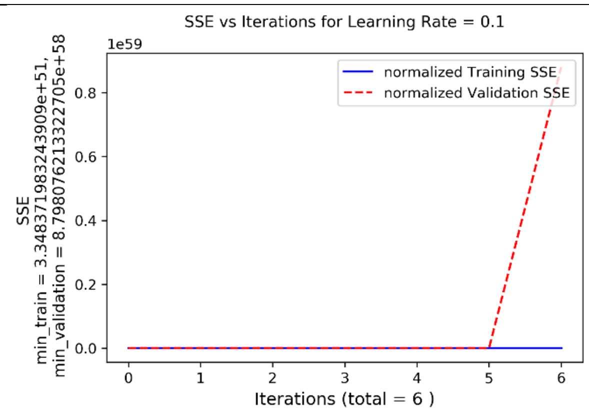
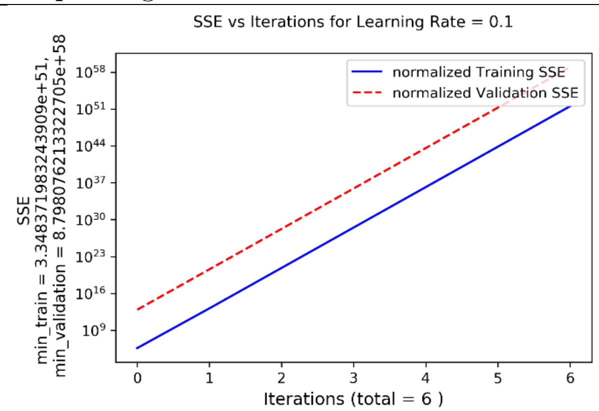
$$Normalized\ value\ =\ \left[\frac{actual\ value - min\_value}{max\_value\ -\ min\_value}\right]$$

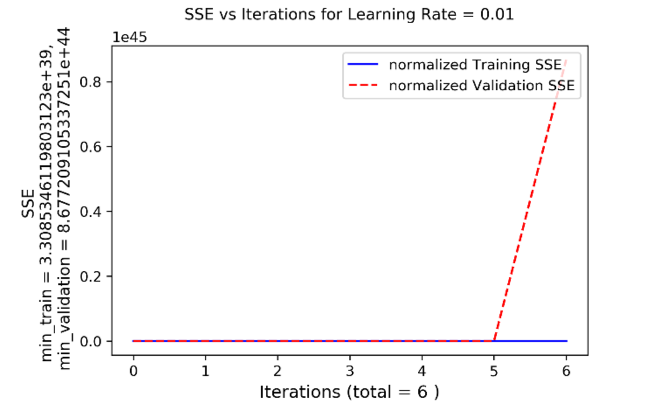
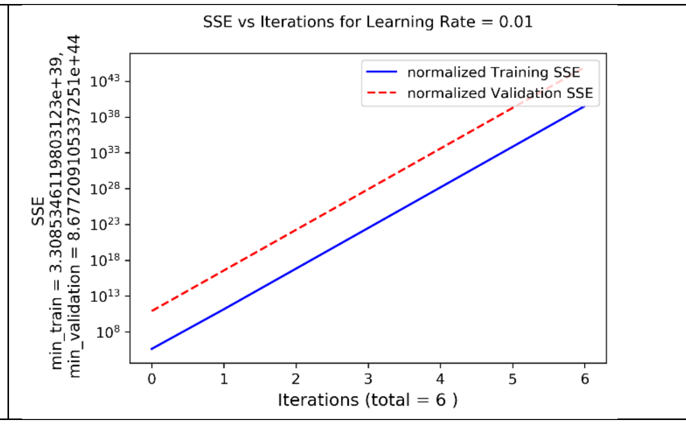
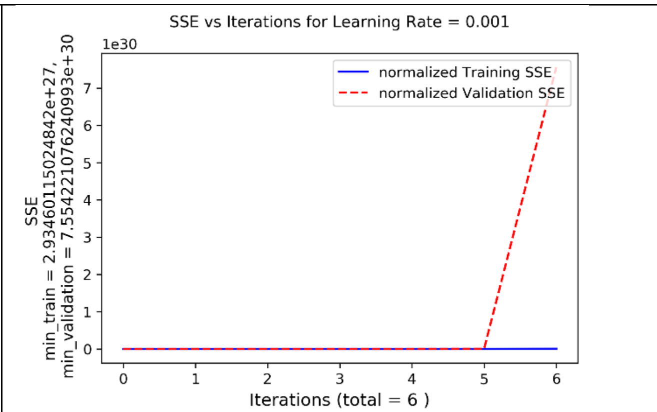
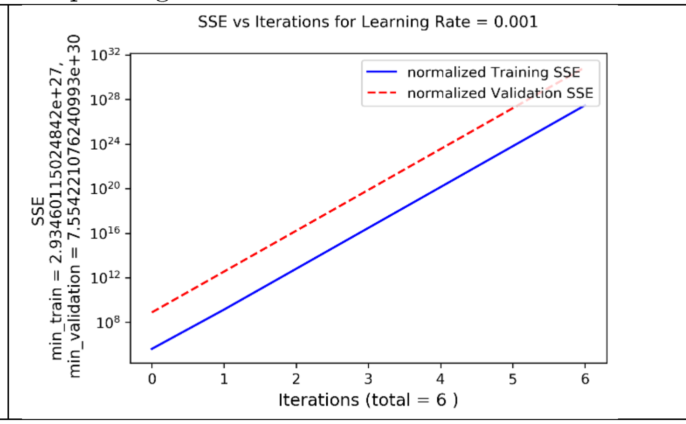
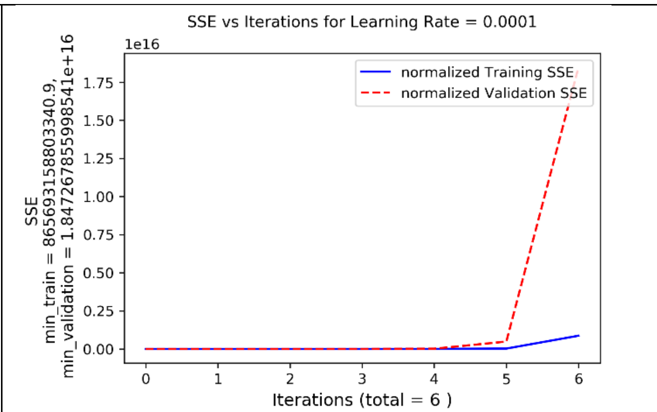
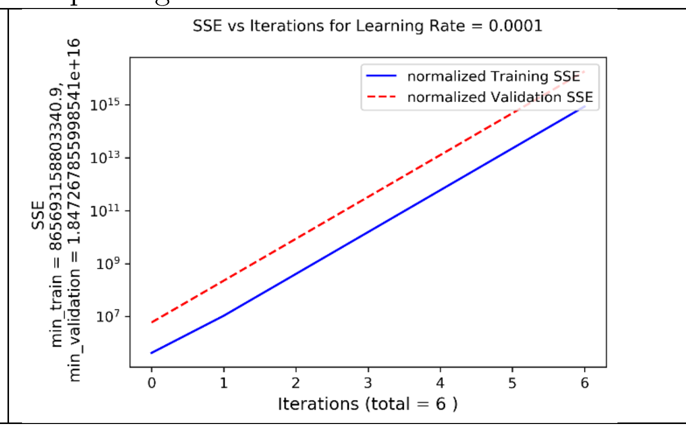
Where, min_value and max_value are the minimum and maximum value from the training data. As we'd have values of only training data at the time of creating the model we consider the maximum and minimum values of the training data.

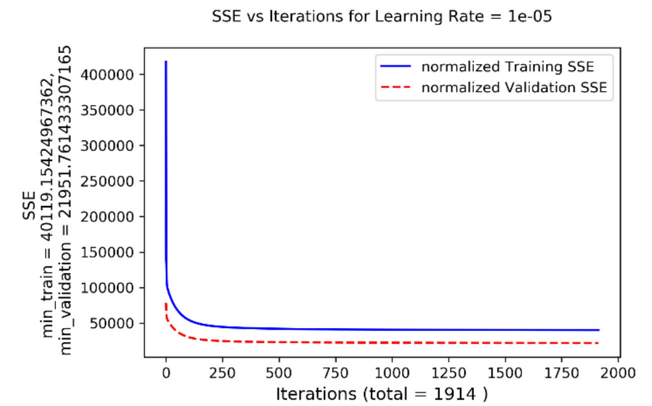
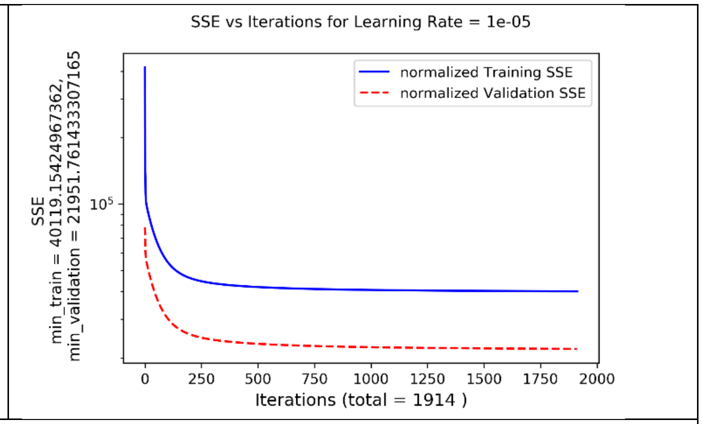
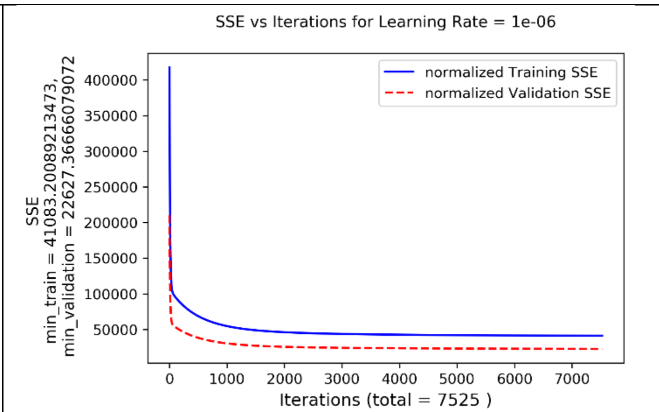
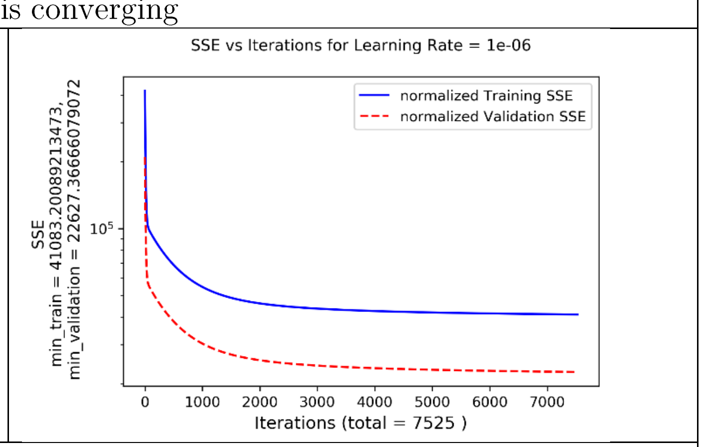
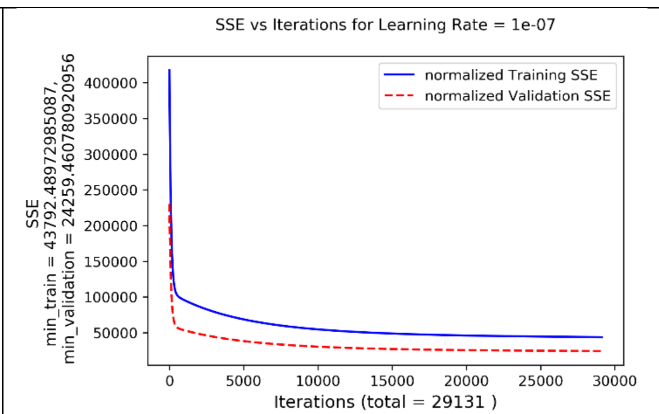
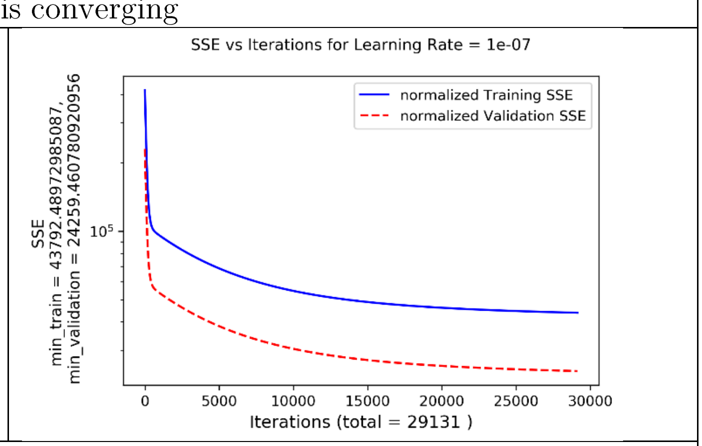
Part 1

a)

The behavior of SSE for the different values of Learning Rates (α) is shown below:

<table>
<tr><th>α</th><th>Normal axis</th><th>Log axis</th></tr>
<tr><td>1</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding</td></tr>
<tr><td>10^{-1}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding</td></tr>
</table>

<table>
<tr><td>10^{-2}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding</td></tr>
<tr><td>10^{-3}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding</td></tr>
<tr><td>10^{-4}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding</td></tr>
</table>

<table>
<tr><td>10^{-5}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is converging</td></tr>
<tr><td>10^{-6}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is converging</td></tr>
<tr><td>10^{-7}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is converging</td></tr>
</table>

We can see from the charts that gradient descent is exploding for learning rates:

$0, 10^{-1}, 10^{-2}, 10^{-3}$, and 10^{-4}.

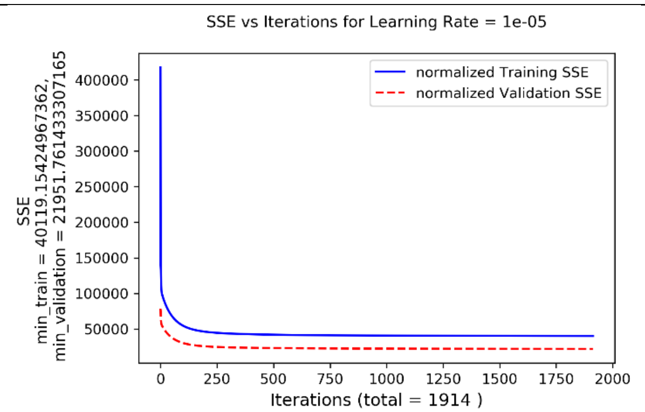
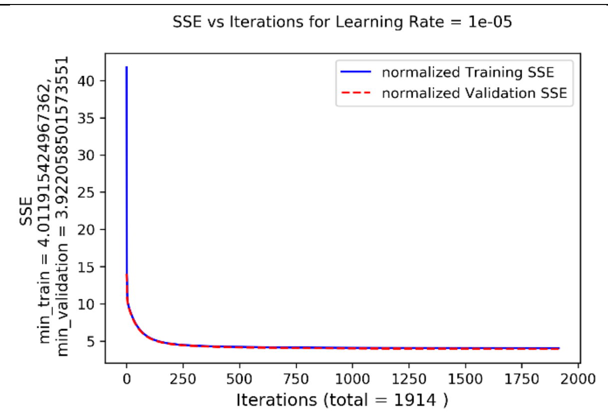
Whereas, it is converging for learning rates:

$10^{-5}, 10^{-6}, 10^{-7}$.

b)

Learning Rate	10^{-5}	10^{-6}	10^{-7}
Training SSE	1914	7525	29131
Validation SSE	21951	22627	24259
Iterations	1914	7527	29131

It can be seen that 10^{-5} is a better learning rate. Furthermore, note that Validation SSE is less than Training SSE. This is because we are using SSE and it depends on the number of samples used. We are given $10K$ samples for training and $\sim 5K$ samples for validation. Consequently, Training SSE is higher than Validation SSE. We also calculated MSE (Mean Square Error) as given below:

<table>
<tr><th>α</th><th>SSE</th><th>MSE</th></tr>
<tr><td>10^{-5}</td><td>
</td><td>
</td></tr>
</table>

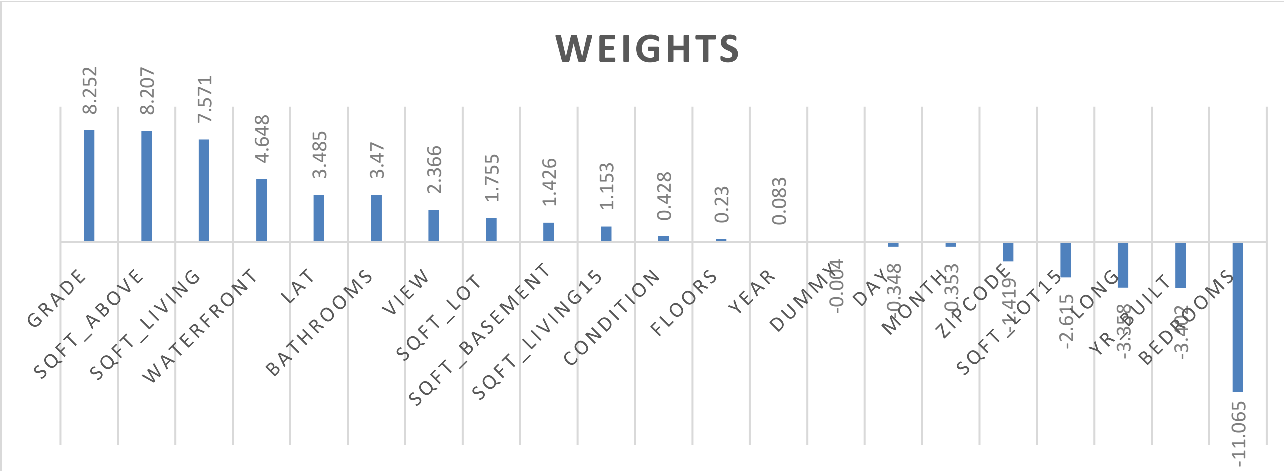
Learning Rate	10^{-5}
Training SSE	4.01
Validation SSE	3.92

We can see that MSE is almost the same for both the Training Dataset and the Validation Dataset. However, our model is performing better for Validation Data. Consequently, we can conclude that straight line model can fit the Validation Data better than the Training Data.

c)

We are using Learning Rate = $\mathbf{10^{-5}}$ for our final model. Behavior of the model for Validation Dataset for learning rates has been shown in the figurse in part (a) and it has been discussed in part (b). Learned weights in decreasing order are given below:

Features	**Weights**
grade	8.252
sqft_above	8.207
sqft_living	7.571
waterfront	4.648
lat	3.485
bathrooms	3.470
view	2.366
sqft_lot	1.755
sqft_basement	1.426
sqft_living15	1.153
condition	0.428
floors	0.230
year	0.083
dummy	-0.004
day	-0.348
month	-0.353
zipcode	-1.419
sqft_lot15	-2.615
long	-3.358
yr_built	-3.402
bedrooms	-11.065

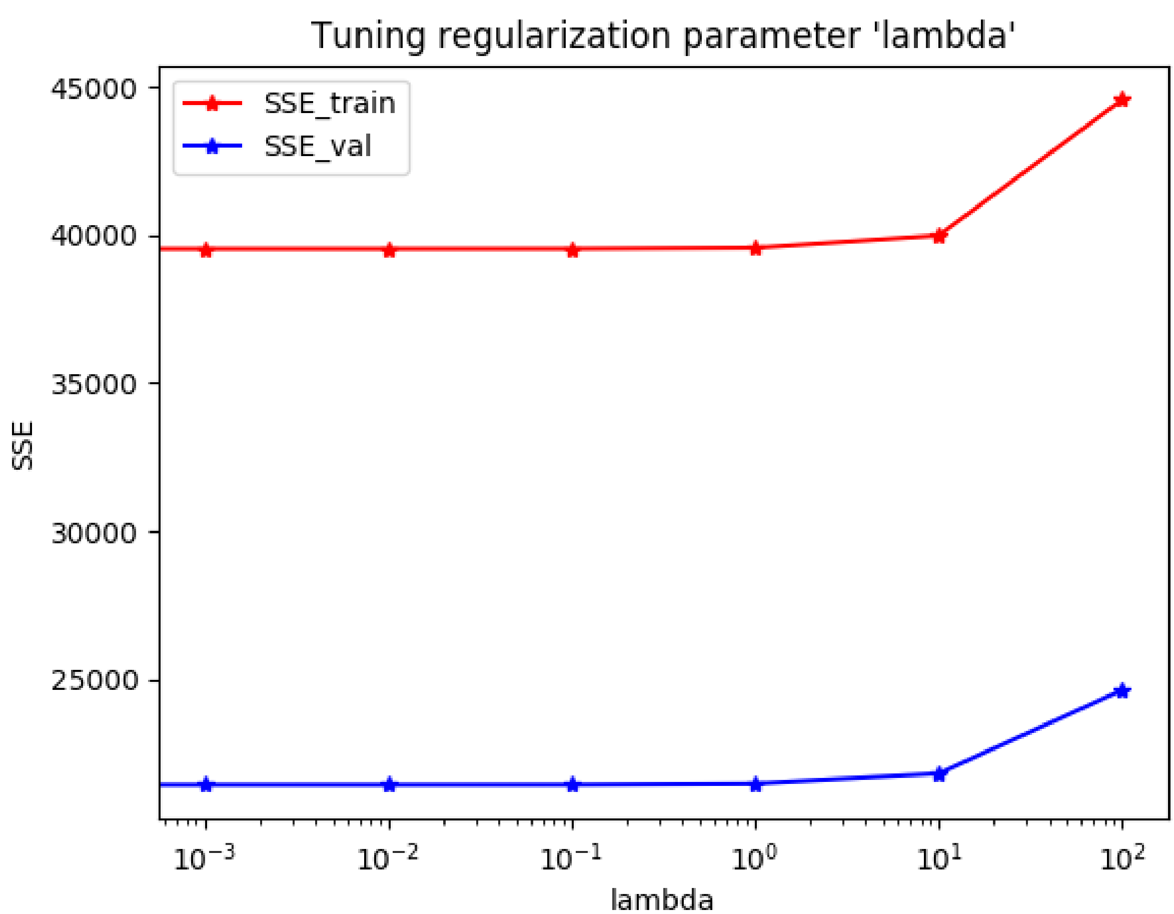


We can see that in our model grade, sqft_above, sqft_living, waterfront, lat, and bathrooms are the most important features. We included most of these features in our discussion in Part(0)-d. Interestingly bedrooms, yr_built, and yr_renovated do not carry much carry weights; whereas waterfront, lat, and bathroom carry significant weights.

Part 2

We experimented with different values of the regularization parameter (λ) as shown below:

Regularization Factor (λ)	**SSE for Training Data Set**	**SSE Validation Data**
0	39532	21466
10^{-3}	39532	21466
10^{-2}	39532	21466
10^{-1}	39535	21470
1	39566	21503
10	39979	21844
100	44560	24658



Weights for different values of λ are given below:

Feature	$\boldsymbol{\lambda = 0}$	$\boldsymbol{\lambda = 0.1}$	$\boldsymbol{\lambda = 0.01}$	$\boldsymbol{\lambda = 1}$	$\boldsymbol{\lambda = 10}$	$\boldsymbol{\lambda = 100}$
bias	-4.78153239E-03	-4.79175504E-03	-4.78255938E-03	-4.87923411E-03	-5.44887260E-03	-4.78153239E-03
month	-4.07036162E-01	-4.08028801E-01	-4.07135980E-01	-4.16432879E-01	-4.65092960E-01	-4.07036162E-01
day	-4.07036162E-01	-3.67844199E-01	-3.67502470E-01	-3.71069178E-01	-3.90301101E-01	-3.67464274E-01
year	5.04102023E-02	4.97548647E-02	5.03443394E-02	4.41718466E-02	9.42456061E-03	5.04102023E-02
bedrooms	-7.24211557E+00	-7.17657605E+00	-7.23552585E+00	-6.62088485E+00	-3.30329082E+00	-7.24211557E+00
bathrooms	3.23697496E+00	3.23560898E+00	3.23683778E+00	3.22385742E+00	3.13663592E+00	3.23697496E+00
sqft_living	7.27967297E+00	7.26475639E+00	7.27817550E+00	7.13609579E+00	6.23539975E+00	7.27967297E+00
sqft_lot	4.49736739E-01	4.46755997E-01	4.49436892E-01	4.21616196E-01	2.77585841E-01	4.49736739E-01
floors	2.66889842E-01	2.70299855E-01	2.67231653E-01	3.00203735E-01	5.39357753E-01	2.66889842E-01
waterfront	4.69802995E+00	4.69129704E+00	4.69735592E+00	4.63144157E+00	4.10330863E+00	4.69802995E+00
view	2.40915777E+00	2.41121737E+00	2.40936427E+00	2.42923057E+00	2.57034429E+00	2.40915777E+00
condition	3.11495640E-01	3.09968181E-01	3.11341812E-01	2.97253553E-01	2.37064778E-01	3.11495640E-01

grade	8.42259186E+00	8.41736268E+00	8.42206925E+00	8.37003875E+00	7.90065432E+00	8.42259186E+00
sqft_above	7.75155555E+00	7.73501040E+00	7.74989486E+00	7.59204356E+00	6.57511138E+00	7.75155555E+00
sqft_basement	1.28374268E+00	1.28317739E+00	1.28368507E+00	1.27911107E+00	1.30084928E+00	1.28374268E+00
yr_built	-3.43018953E+00	-3.42939381E+00	-3.43011033E+00	-3.42187401E+00	-3.32392693E+00	-3.43018953E+00
yr_renovated	6.29956522E-02	6.34470835E-02	6.30407058E-02	6.75933806E-02	1.13078371E-01	6.29956522E-02
zip code	-1.44073942E+00	-1.43994274E+00	-1.44065981E+00	-1.43272122E+00	-1.35948599E+00	-1.44073942E+00
lat	3.46344952E+00	3.46242963E+00	3.46334724E+00	3.45353799E+00	3.38443072E+00	3.46344952E+00
long	-3.44659414E+00	-3.44382603E+00	-3.44631781E+00	-3.41846539E+00	-3.14495903E+00	-3.44659414E+00
sqft_living15	1.17272282E+00	1.18458742E+00	1.17391304E+00	1.28772747E+00	2.04488913E+00	1.17272282E+00
sqft_lot15	-1.12000628E+00	-1.10846321E+00	-1.11884536E+00	-1.01088596E+00	-4.43771280E-01	-1.12000628E+00

a)

What trend do you observe from the training SSE as we change λ value?

Ans:

Training SSE remains almost constant at 39532 for lambda values: {0, 0.001 and 0.01}. As we increase lambda from 0.1 to higher values, training SSE increases exponentially.

b)

What tread do you observe from the validation SSE?

Ans:

Validation SSE follows a similar trend as that of the training SSE. It remains almost constant at 21466 for lambda values: {0, 0.001 and 0.01}. As we increase lambda from 0.1 to higher values, validation SSE increases exponentially.

c)

Provide an explanation for the observed behaviors.

Ans:

These trends for Training SSE and Validation SSE denote that the linear regression model fits the data well for lambda values: {0, 0.001, 0.01 and 0.1}. For higher values of lambda, the model starts suffering from under-fitting. The severity of under-fitting increases exponentially with higher lambda values.

d)

What features get turned off for $\lambda = 10$, $10\ -2$ and 0 ?

Ans:

The features corresponding to weights with drastically small magnitudes relative to the other weights are the ones that get turned off. Hence the analysis is as follows:

1. lambda = 10

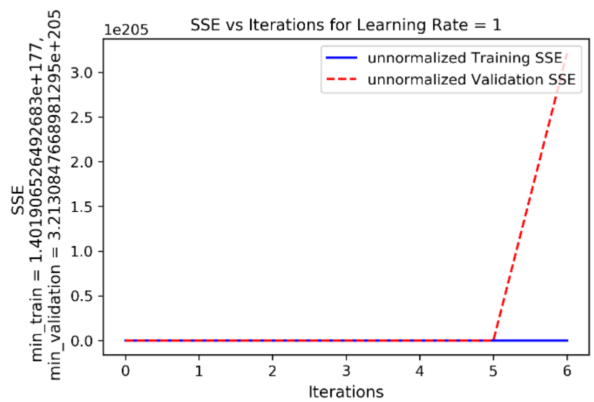
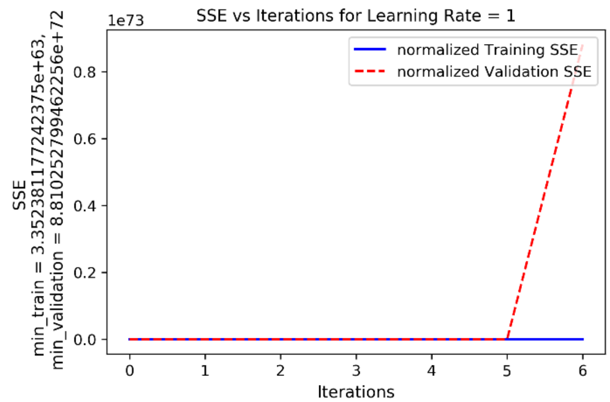
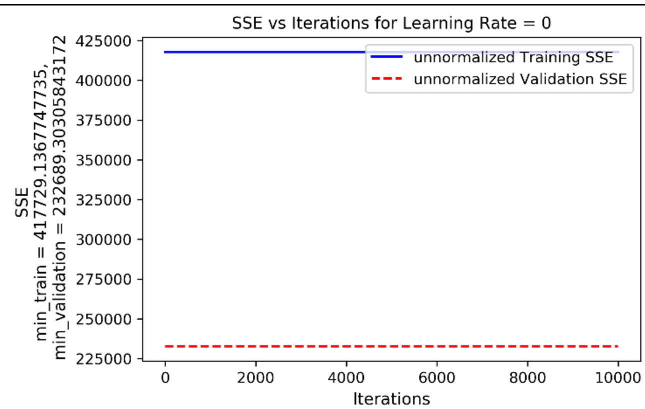
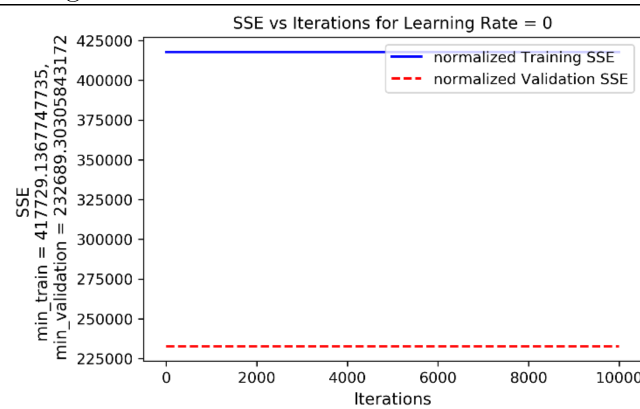
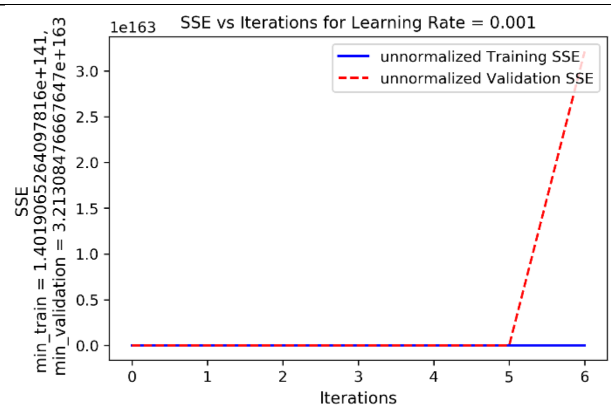
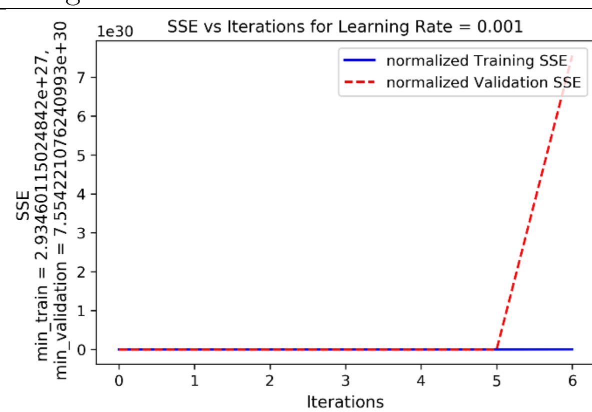
   Turned off/less weighted Features: year
2. lambda = 0.01

   Turned off/less weighted Features: yr_renovated, year
3. lambda = 0

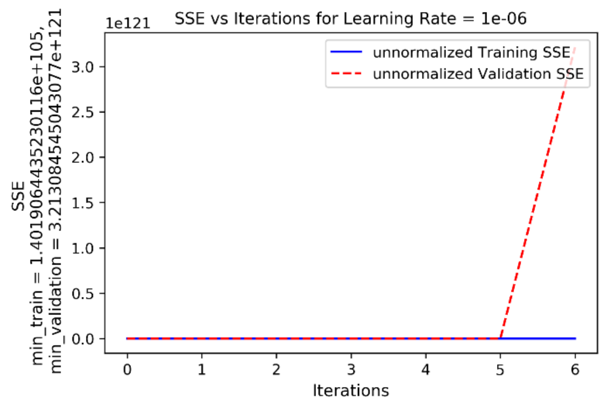
   Turned off/less weighted Features: yr_renovated, year

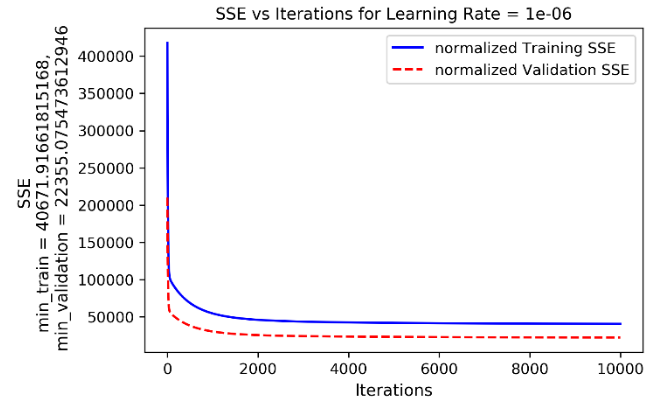
Part 3

Training SSE and Validation SSE for different values of the learning rate(α) are given below:

<table>
<tr><th>α</th><th>Non-normalized Data</th><th>Normalized Data</th></tr>
<tr><td>1</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding for both</td></tr>
<tr><td>0</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding for both</td></tr>
<tr><td>10^{-3}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is exploding for both</td></tr>
</table>

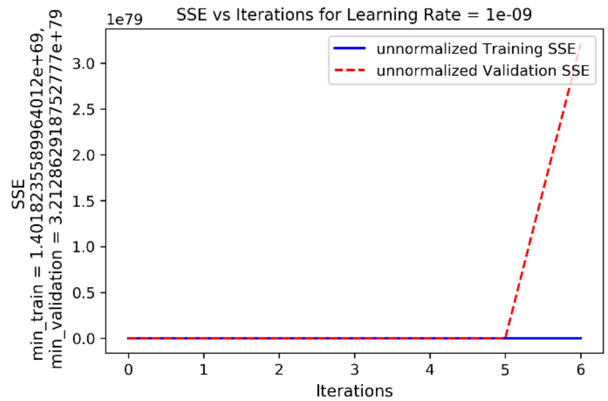
10^{-6}

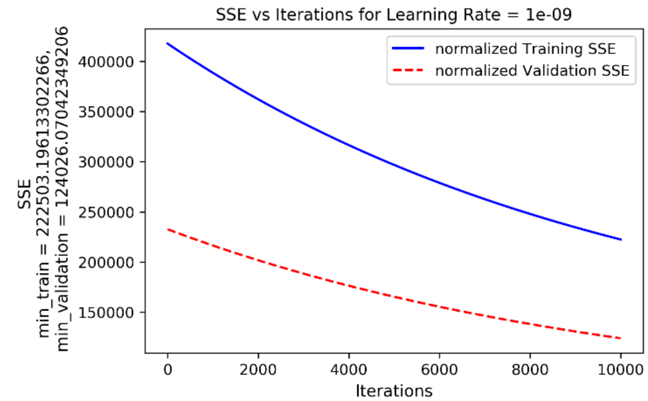




Gradient descent is still exploding for non-normalized data but it is converging for normalized data

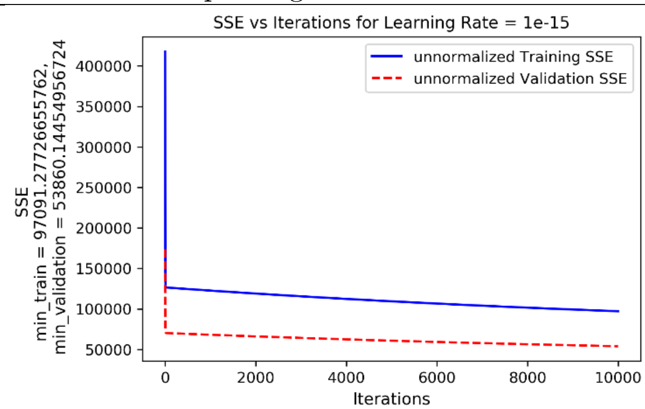
10^{-9}

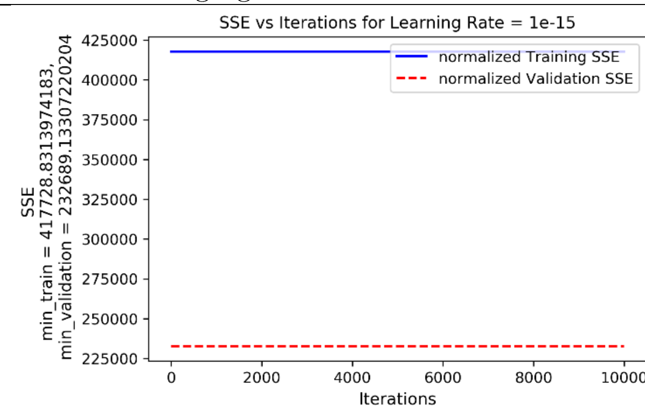




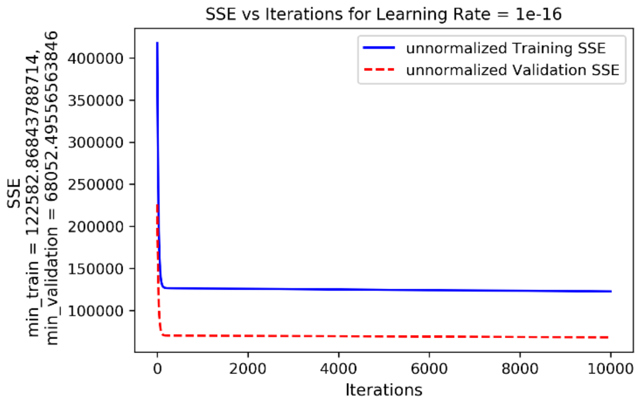
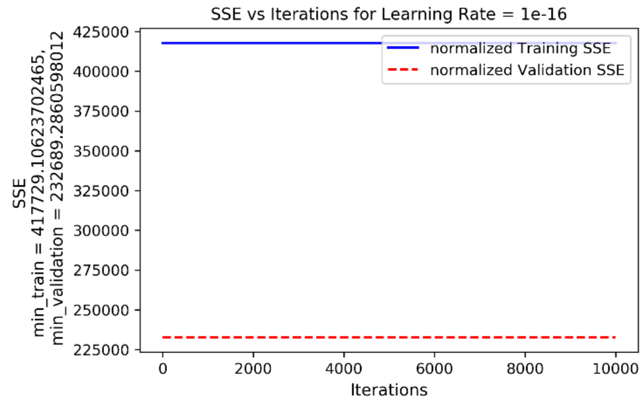
Gradient descent is exploding for non-normalized data but it is converging for normalized data

10^{-15}



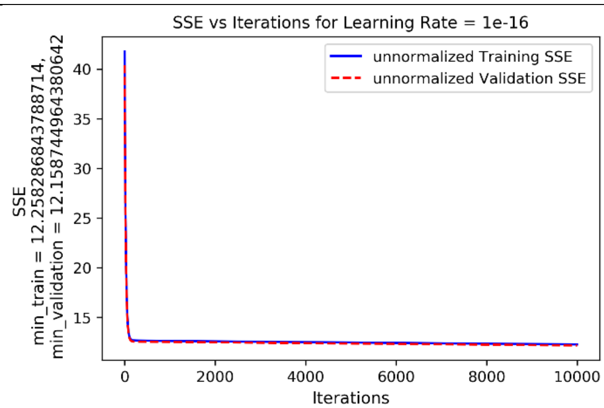
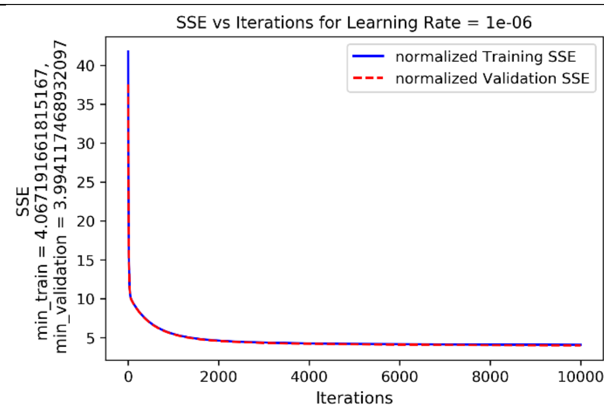


Gradient descent is converging for both

<table>
<tr><td>10^{-16}</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Gradient descent is converging for both</td></tr>
</table>

Discussion:

It is evident from the graphs that normalization helps to converge the linear regression faster. For fast learning rates, gradient descent is exploding with or without normalization. Gradient descent starts to converge at $\alpha = 10^{-6}$ for the normalization data sets but it is still exploding for the non-normalized data. Finally, gradient descent begins to converge at $\alpha =^{-15}$ for the non-normalized data.

<table>
<tr><td>
</td><td>
</td></tr>
<tr><td>MSE for Non-normalized Data</td><td>MSE for Normalized Data</td></tr>
</table>

Gradient descent reaches at an MSE of ~ 12 (it is still going down), after 10K iterations at a learning rate of 10^{-16}; whereas, it reaches ~ 4 after 10K iterations at a learning rate of 10^{-6}. Hence, normalization helps gradient descent to converge faster.

Conclusion

For the given Training Data set we generated a linear model with L2 regularization. We experimented with different values of Learning Rate and Regularization Term. Furthermore, we also analyze the effects of normalization on the convergence rate. In addition, we tested our model on the Validation data. In addition, prediction values for the Test Data is attached with this assignment. In our experience, more complex polynomial models will yield better results.