COMP 551 - Assignment 1

Report

Q1.1 Pseudocode:

Define a function f (no# days):

Initialize four variables for different routine: R1, R2, R3, R4

for (some x in range(n): #n is the number of days

generate a random number y between [0,1)

#for probability [0.0, 0.1), increment
#Playing sample points

if (y = [0.0, 0.1)):

increment R3

#for probability [0.1, 0.3), increment
#Movies sample points

else if (y = [0.1, 0.3)):

increment R1

#for probability [0.3, 0.6), increment
#Studying

else if (y = [0.3, 0.6)):

increment R4

#for probability [0.6, 1.0), increment
#COMP-551

else:

increment R2

return an array of [R1, R2, R3, R4] which are the sample points obtained from each activity for n days

1.2. For n = 100:

Movies: 18 / 100 = 0.18

COMP-551: 38 / 100 = 0.38

Playing: 9 / 100 = 0.09

Studying: 35 / 100 = 0.35

For n = 1000

Movies: 185 / 1000 = 0.185

COMP-551: 392 / 1000 = 0.392

Playing: 106 / 1000 = 0.106

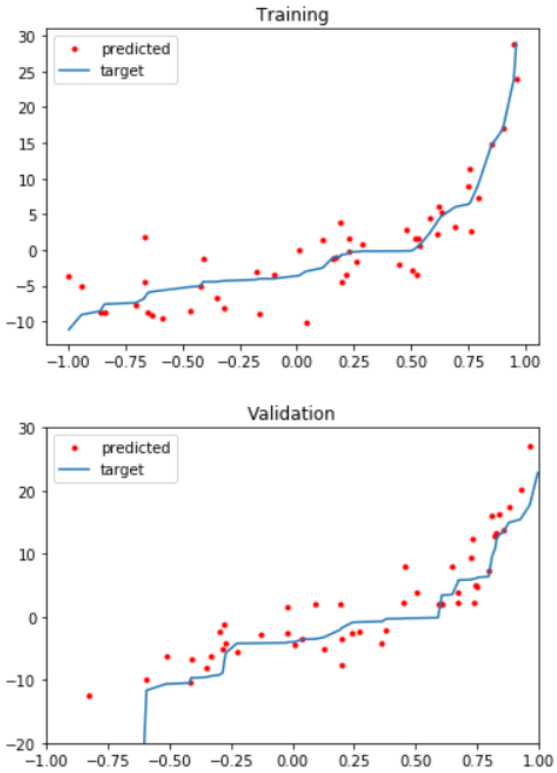
Studying: 317 / 1000 = 0.317

Comparison: As the sample points increase, the error decreases and therefore yields a better estimation.

Q2.1. a. MSE of Training Data: 6.474

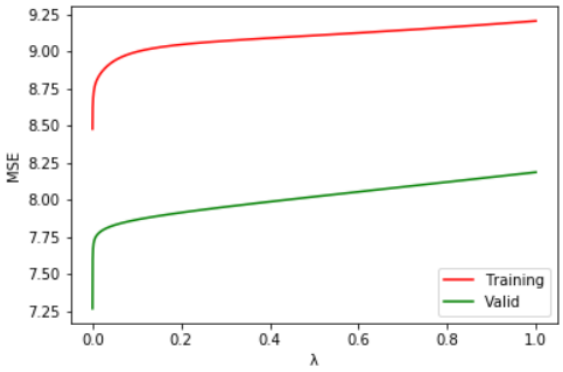
MSE of Validation Data: 1422.692

b.



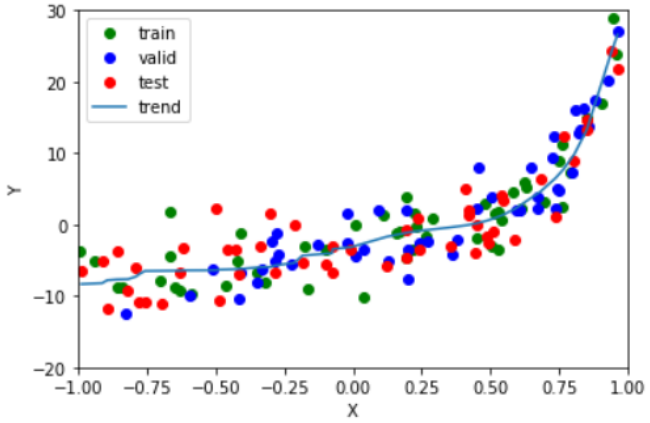
c. The model is overfitting because it is fitting too well with the training dataset instead of generalizing the general pattern. Due to this it does not fit well with the validation data set and hence yields high MSE.

2.2 a.



b. MSE for λ = 0.0001 for the test set is: 11.1802

c.



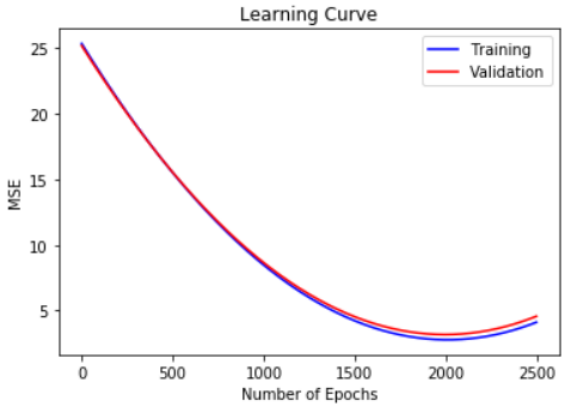
d. The model is neither overfitting nor underfitting. It is well modelled along the points.

2.3 The degree of the source polynomial is 3

3.1 a MSE for Validation Set on 5000 epochs:

```
25.19355702
25.18647866
25.17944864
...
4.51871471
4.51750569
4.51629343
```

b



3.2 a

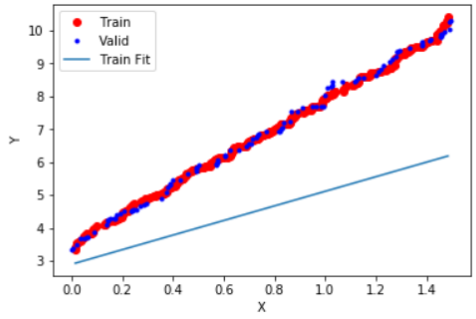
Step Size	MSE for 10 epochs
0.1	5058.13567063 21776.13131168 50262.96281889 90618.14478577 142725.81324745 207228.85378308 282735.17325151 370448.89279453 468484.3861287 577697.46305834
0.01	11.555188 114.43423706 333.81284372 673.57276139 1128.07938641 1699.46642411 2385.40420006 3185.97071892 4106.40558325 5121.06729019
0.001	18.57412605 13.10417429 8.82452395 5.73886011 3.81293147 3.05458473 3.46603625 5.04595453 7.80348673 11.64639179
0.0001	24.49159363 23.79125311 23.09716063 22.42292336 21.76706304 21.1153402 20.48268796 19.85628065 19.23783974 18.63309653
1e-05	25.12924933 25.05733424 24.98510392 24.91303672 24.84167885 24.77013397 24.69868087 24.62734715 24.55639693 24.48557603

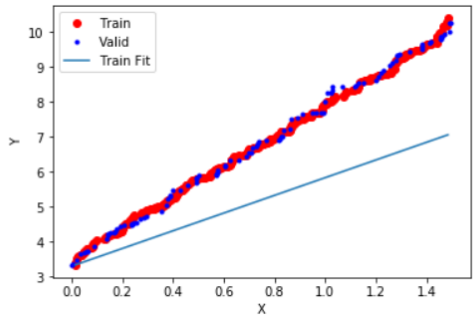
1.0000000000000002e-06	25.19339289 25.18628812 25.1790696 25.17192623 25.16476121 25.15754917 25.15043366 25.14321502 25.13604647 25.12887617

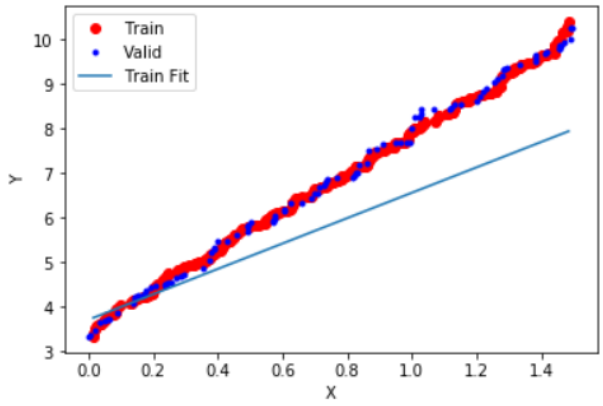
b. MSE of Test Set for step-size: 0.001 for 10 epochs is:

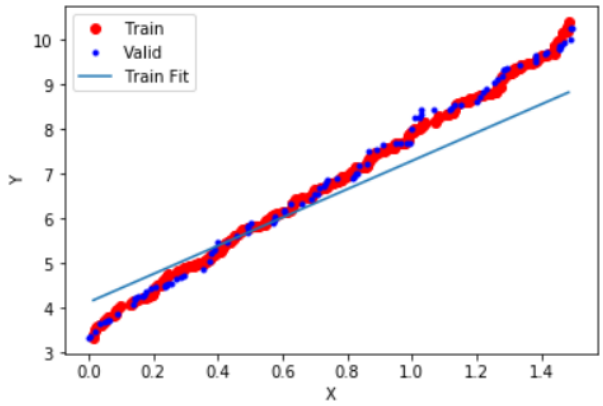
24.89865077
24.87631008
24.85404494
24.83171244
24.80955342
24.78727809
24.76512965
24.74295877
24.72079929
24.69872422

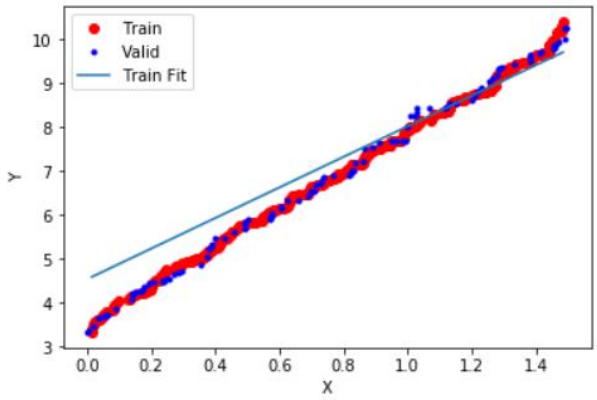
3.3






Train
Valid
Train Fit
Y
X
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Train
Valid
Train Fit
Y
X
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Train
Valid
Train Fit
Y
X
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

4.1 **a** No, because there might be outliers that will result increase the variance and bias, resulting in improper approximations.

b Using median

c Using sample median gives a better approximation of the missing attributes. Since the median is not affected by the outliers, it does not affect the variance and yields better approximations than using the sample mean.

4.2 Parameters learnt using SGD from the training set:

1. w0 = 3.395e-05 w1 = 1.17515317e-05
2. w0 = 5.939e-05 w1 = 2.05708073e-05
3. w0 = 8.702e-05 w1 = 3.03282033e-05
4. w0 = 0.00012281 w1 = 4.2882826e-05
5. w0 = 0.0001504 w1 = 5.26665382e-05

4.3 **a**

b No because ridge regression minimizes all the predictor coefficient and estimates them towards zero, based on the size of the regularization hyperparameter λ. Although this reduces the MSE but does not contribute towards choosing the best feature.

c

d Performance with reduced features will result in better performance. This will eliminate the issue of dimensionality resulting in complex decision boundaries. It also reduces overfitting. Lesser features also imply fewer correlated variables which improves the learning algorithm.