CSE 474/574: Introduction to Machine Learning

University at Buffalo

Keming Kuang (kemingku)

personal #: 50161776

October 9, 2018

Contents

1 Description

1.1 Objective

In this project, we will learn to solve the Learning to Rank (LeToR) Problem. By formulating this into a linear regression problem, we can solve it by using closed-form solution or stochastic gradient descent.

1.2 Task

With given dataset and target value, we will Train a linear regression model on them by using a closed-form solution and stochastic gradient descent.

1.3 Plan of Work

Closed Form Solution: We will train the model parameter with a given group of hyper parameter on the training set, validation set, testing set. We will get the corresponding accuracy and Erm by evaluating the data.

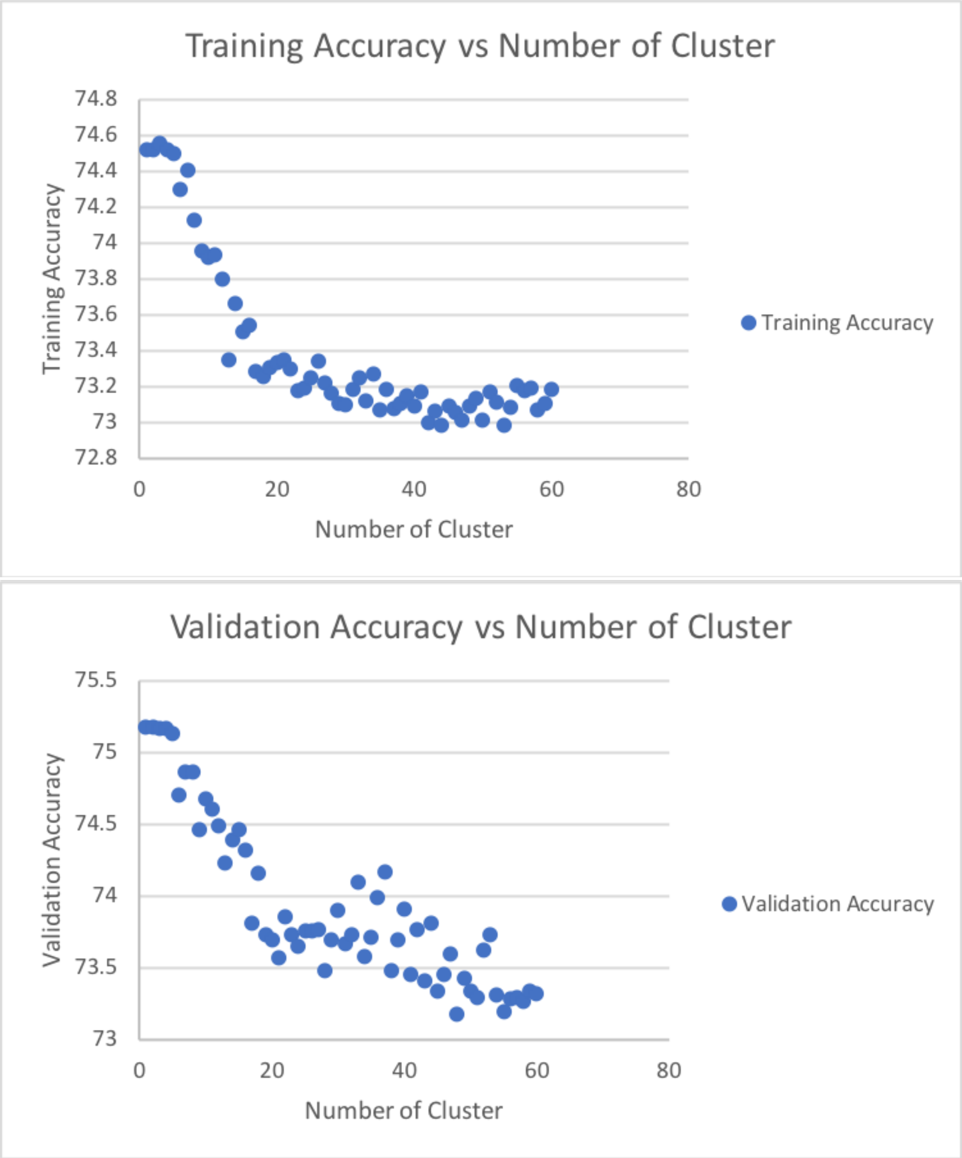
Stochastic Gradient Decent: We will train the model parameter with a given group of hyper parameter on the training set, validation set, testing set. In addition, we will study the training the learning rate and iteration on the dataset. By Evaluating the result, we will study the performance by compare the accuracy and Erms.

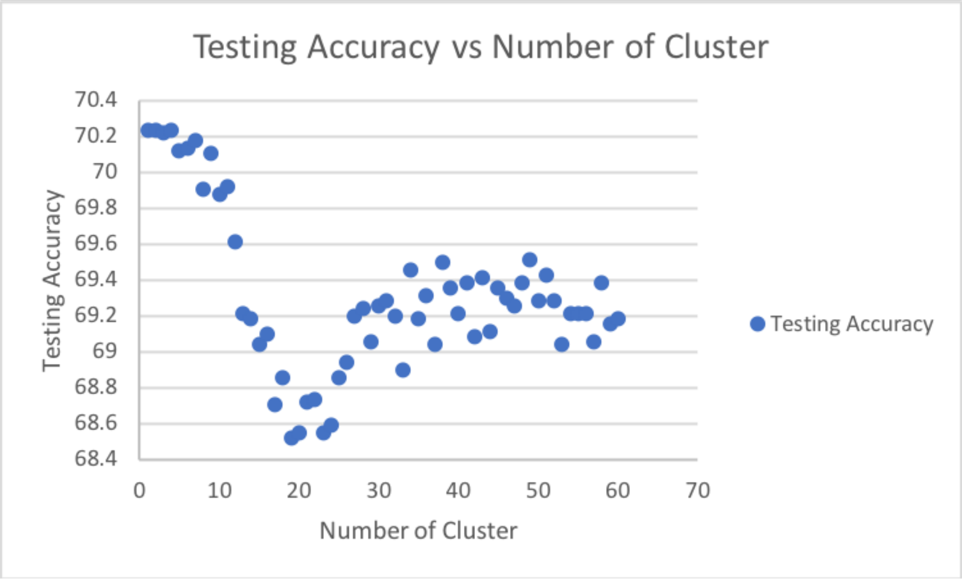
2 Closed Form Solution

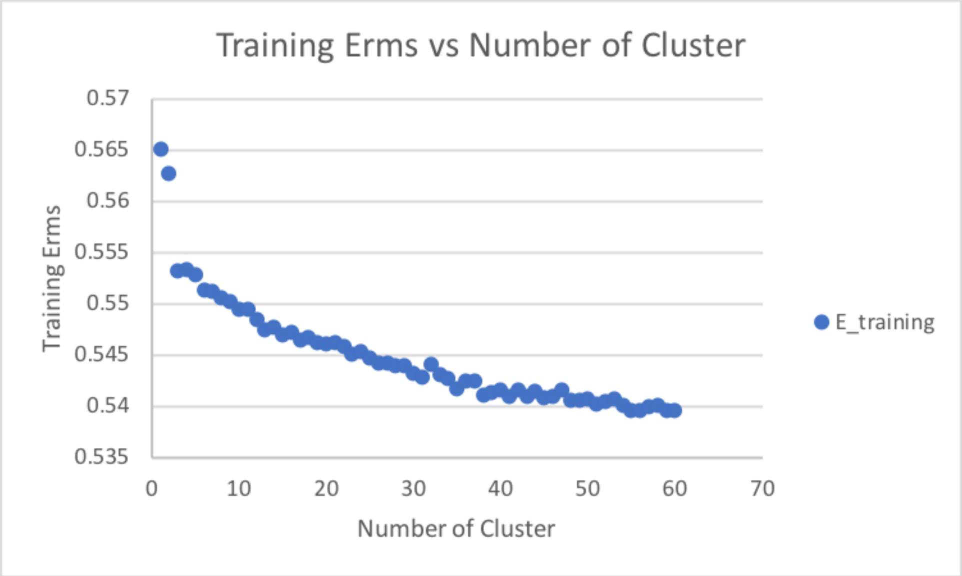
In this section, I will study the Closed Form Solution performance when the hyper parameter change.
By default, I will use
k-means = 10
Lambda = 0.03

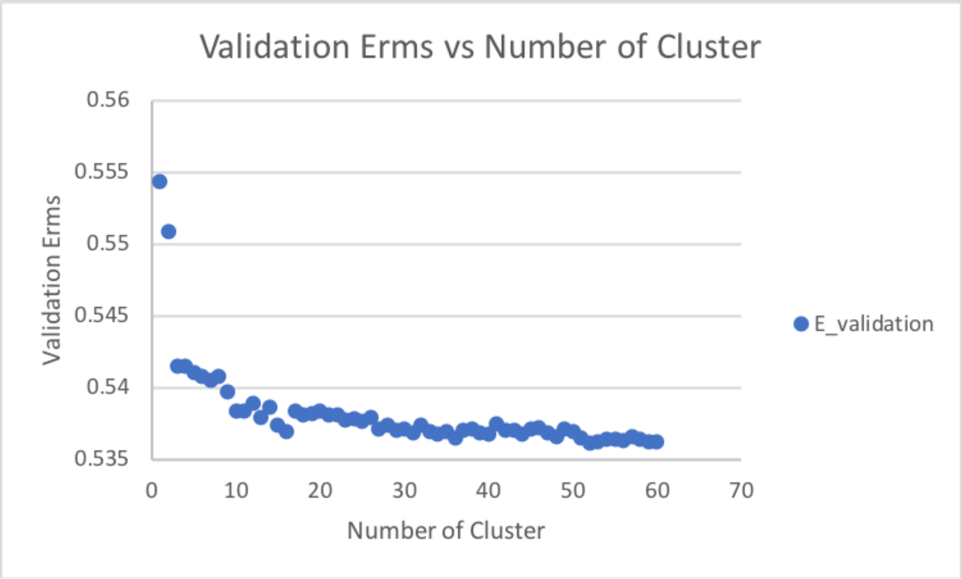
2.1 Number of Clusters

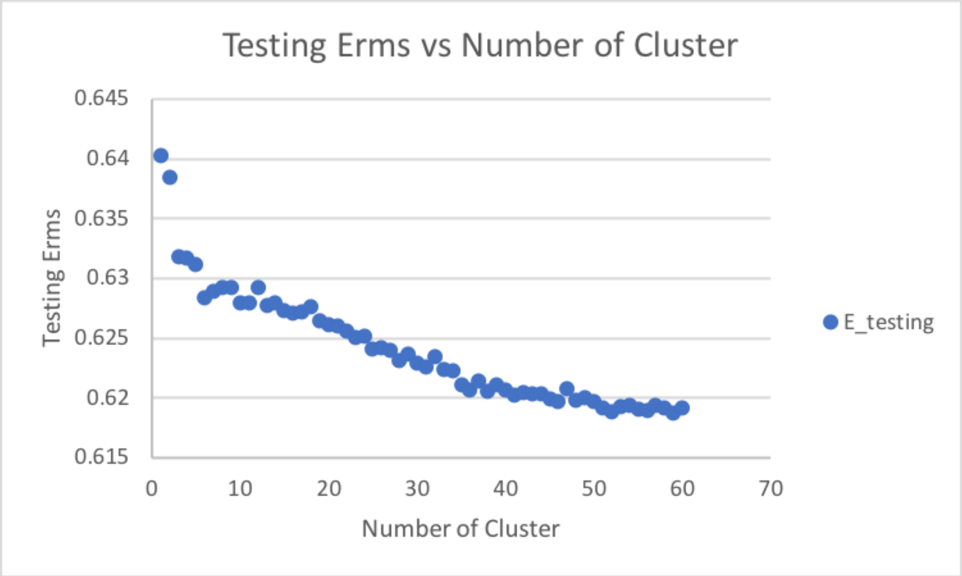
A cluster is a collection of data which are similar to each other and dissimilar to data in other clusters. As we are formatting the Learning to Rank Problem to linear regression, we will use a specific clustering approach: K-means clustering. By using k-means clustering, it allow us to add one more step before the unsupervised learning. In addition, k-means clustering partitions the given data into k cluster and determines all clusters at once. In this session, we will compare how the accuracy and Erms vary under different number of clusters.




Testing Accuracy vs Number of Cluster
Testing Accuracy
70.4
70.2
70
69.8
69.6
69.4
69.2
69
68.8
68.6
68.4
0
10
20
30
40
50
60
70
Number of Cluster
● Testing Accuracy


Training Erms vs Number of Cluster
Training Erms
0.57
0.565
0.56
0.555
0.55
0.545
0.54
0.535
0
10
20
30
40
50
60
70
Number of Cluster
● E_training



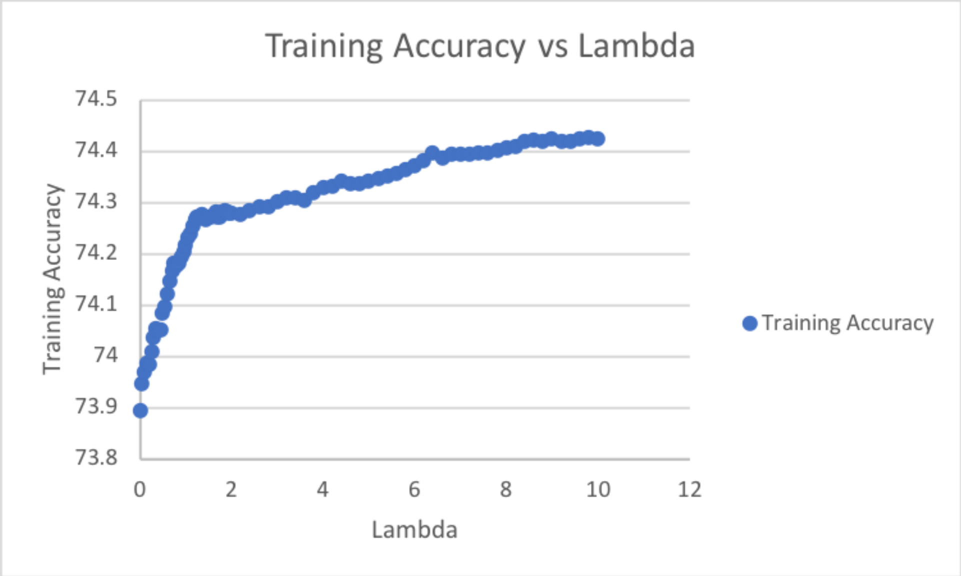


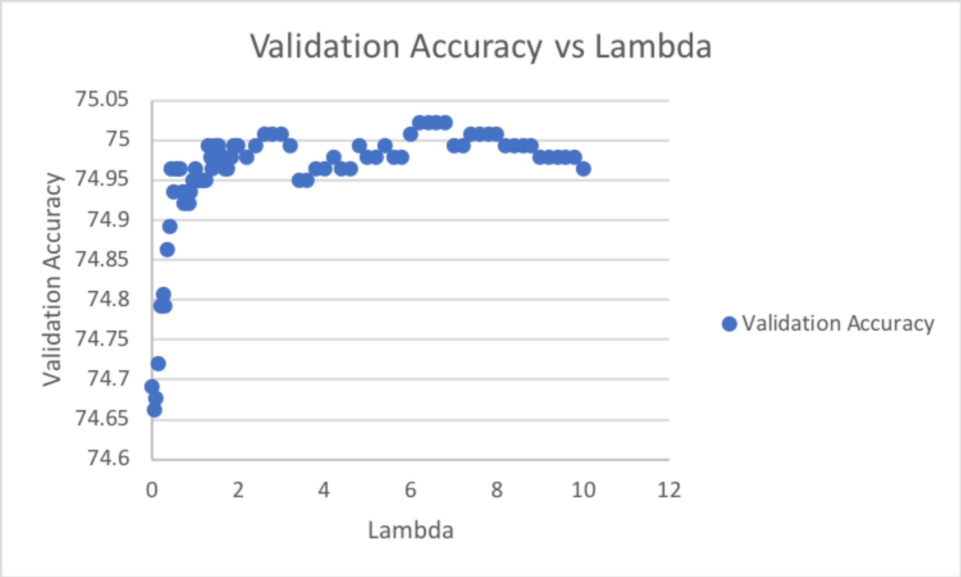
2.1.1 Conclusion

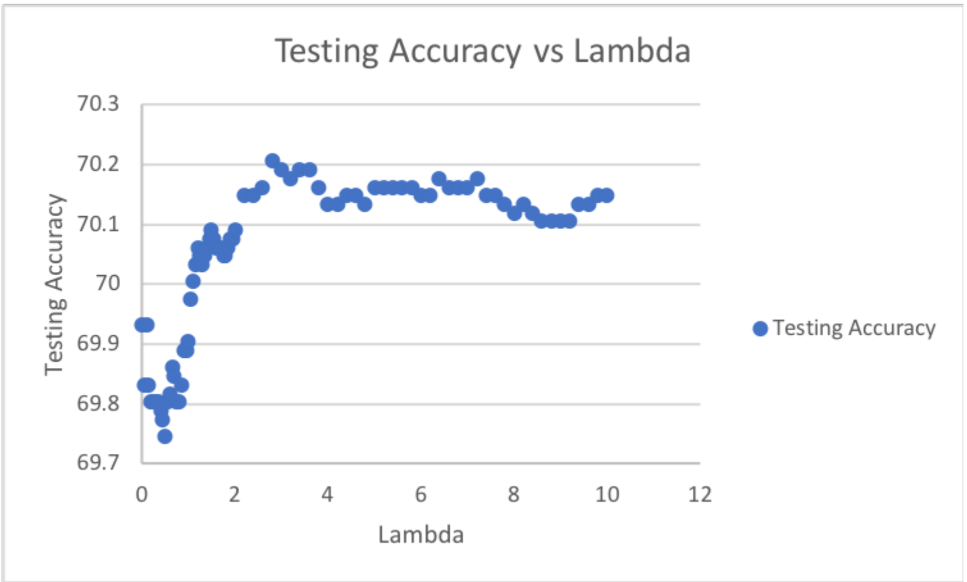
As shown above, we tested the accuracy and Erms under the k-means clusters of 1 to 60. Since we partitioned the data into T=training data, V=validation data, and testing data, we can get their corresponding data accuracy and data erms. While in common that all accuracy decrease with the number of clusters increments, the trend of testing accuracy claims that the testing accuracy has positive increase in general after 20 clusters. Erms also decrease slightly along with the increasing clusters, but they follows so small pace in decreasing that they nearly forms an linearly line.

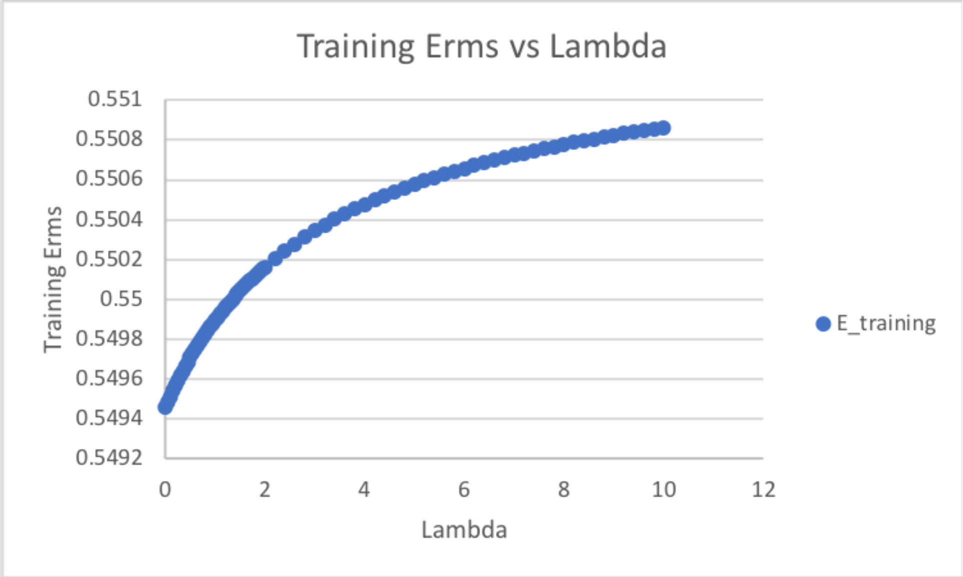
2.2 Lambda

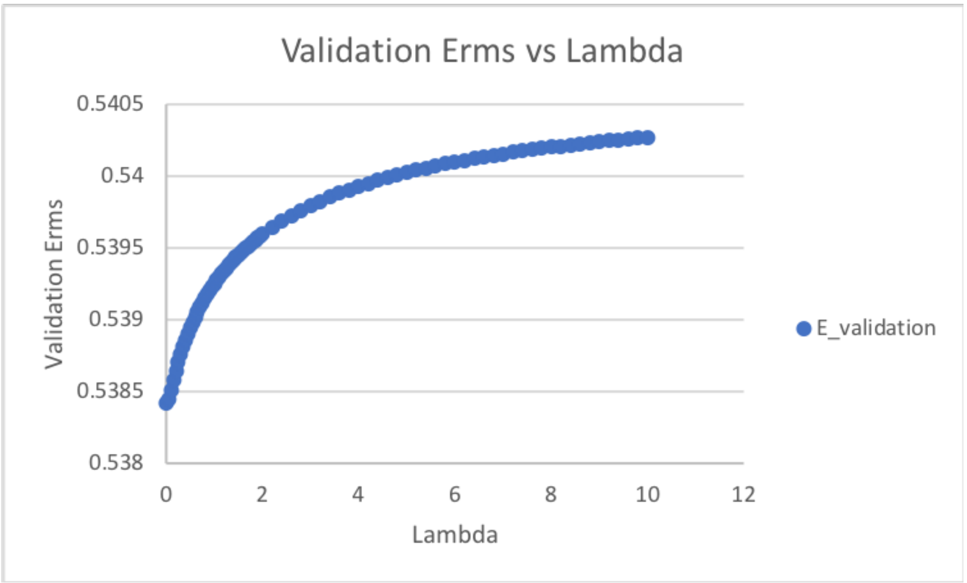
Since the criterion used for selecting the model is not the same as the criterion used to judge the suitability of the model. In our problem solving, overfitting may occur when the model capture points that particularly has significant difference within its main data group, and this usually results in capturing noise in the model. To avoid overfitting, we will apply Lambda to our getweight function to punish the data. By using the regularization, we should be able to see a more clear graphic about the accuracy and erms we are getting. By using regularization, we will using lambda and weight decay regularization, and we understand that this will affect the final result slightly. As acknowledged, we may get a lower accuracy and a high error.

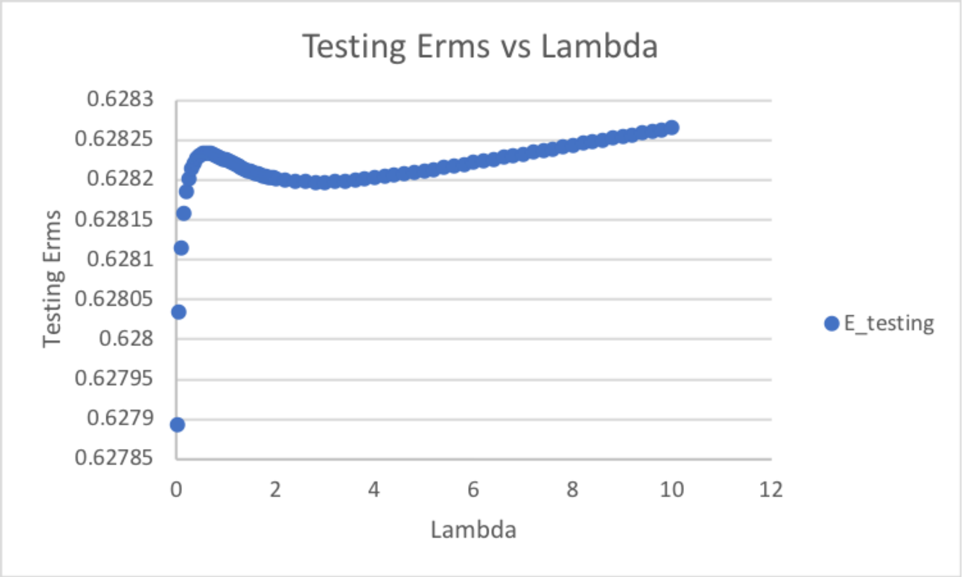





Testing Accuracy vs Lambda
70.3
70.2
70.1
70
69.9
69.8
69.7
Testing Accuracy
0
2
4
6
8
10
12
Lambda
● Testing Accuracy


Training Erms vs Lambda
0.551
0.5508
0.5506
0.5504
0.5502
0.55
0.5498
0.5496
0.5494
0.5492
Training Erms
0
2
4
6
8
10
12
Lambda
● E_training





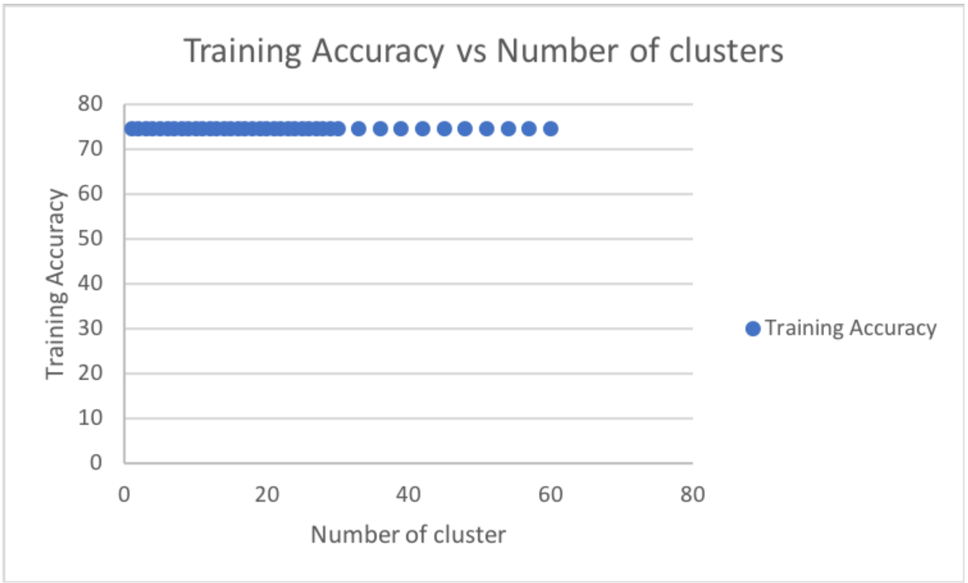
2.2.1 Conclusion

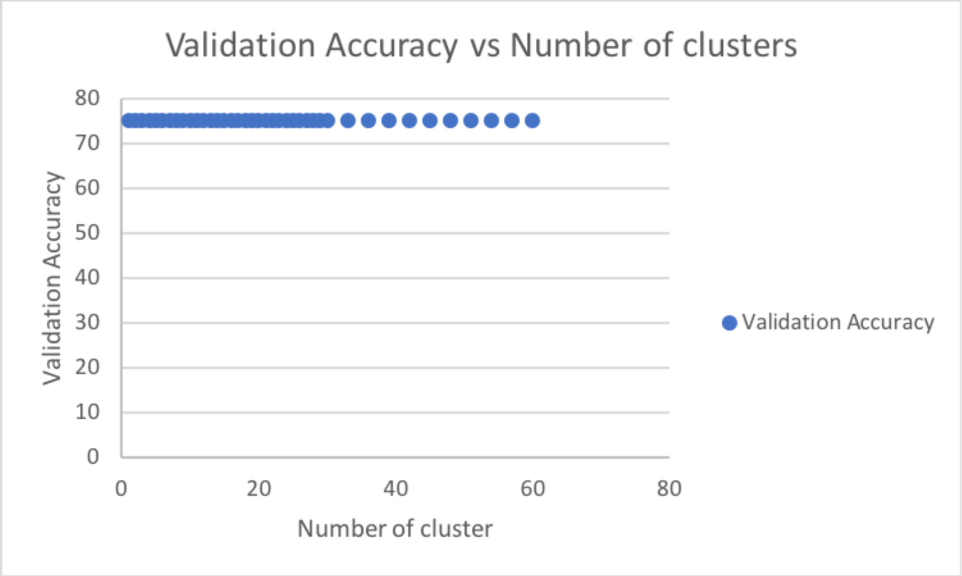
As shown above, we have value tested in an intervals of 0.05 Lambda in the range of Lambda 0 to 2 and an intervals of 0.2 Lambda in the range of Lambda 2 to 10. This is because I realized all accuracy and erm has and significant change at the point lambda 0.5, and I also want to show the long term changing from Lambda 2 to 10. As a result, all accuracy and erm have an increasing trend in general. Training accuracy has an more constant and steady increase comparing to validation accuracy and testing accuracy. Moreover, both testing accuracy and testing erms have an huge change in trend at the point 0.5, and the increasing trends for both training erms and validation erms tends to slow down.

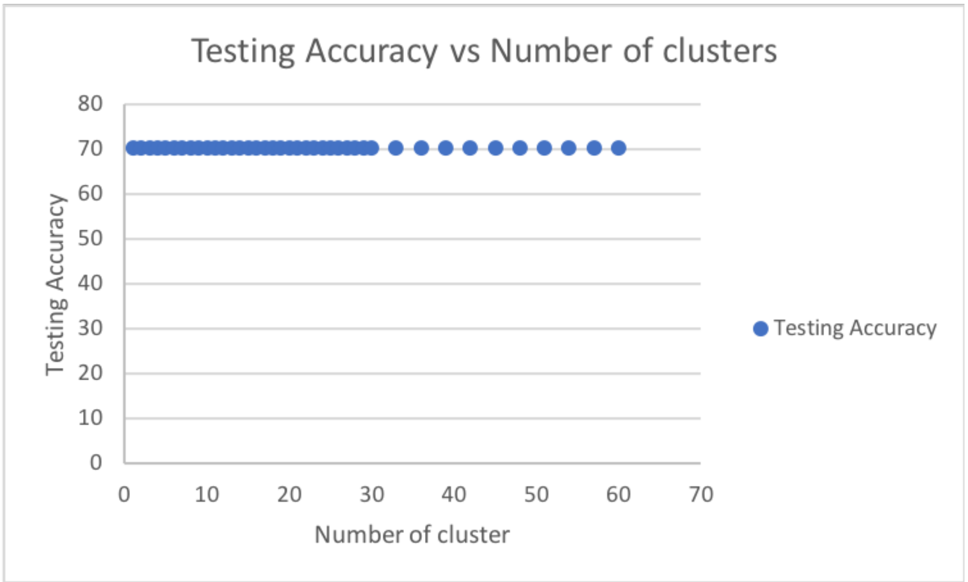
3 Stochastic Gradient Decent

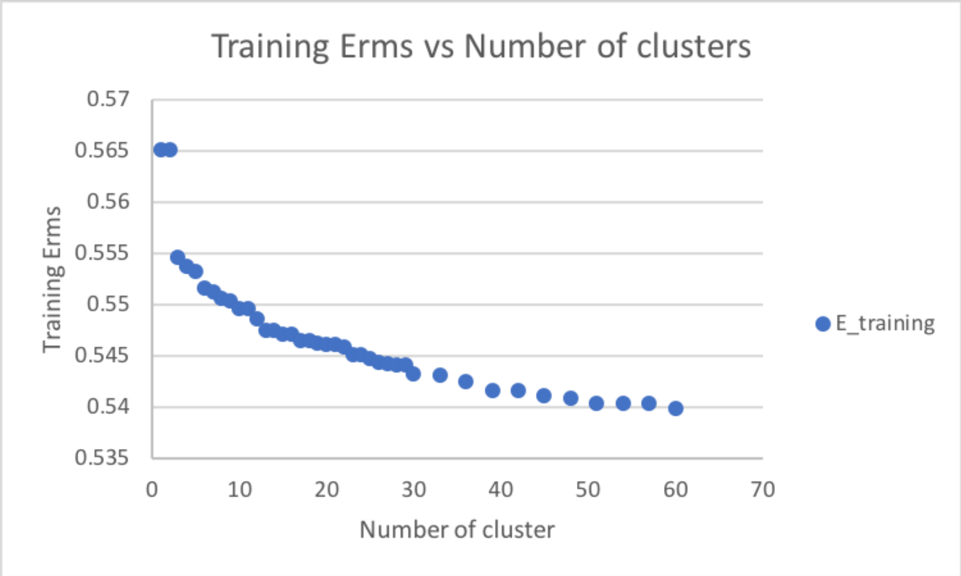
In this section, I will study the Closed Form Solution performance when the hyper parameter change.
By default, I will use
k-means = 10
Lambda = 2
Learning rate = 0.01

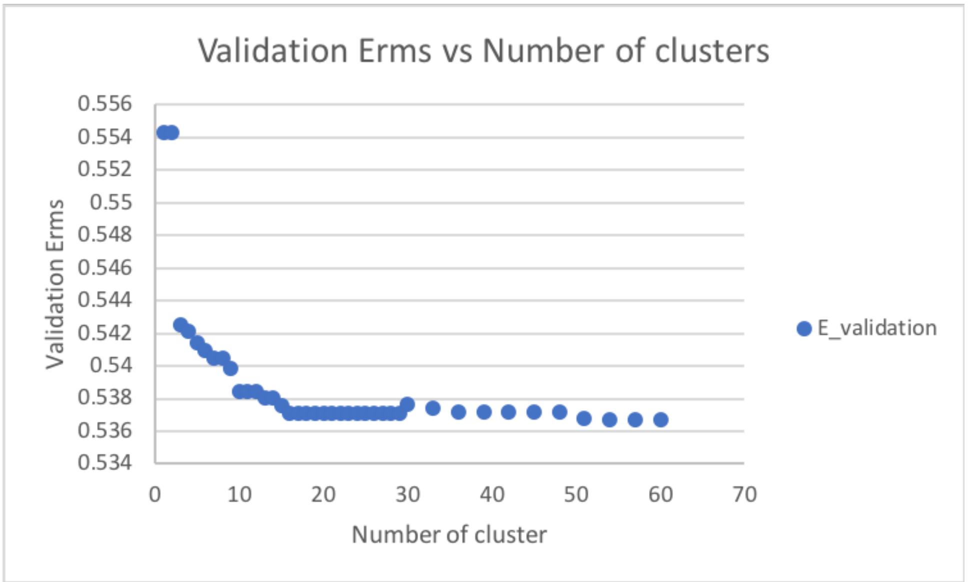
3.1 Number of Clusters

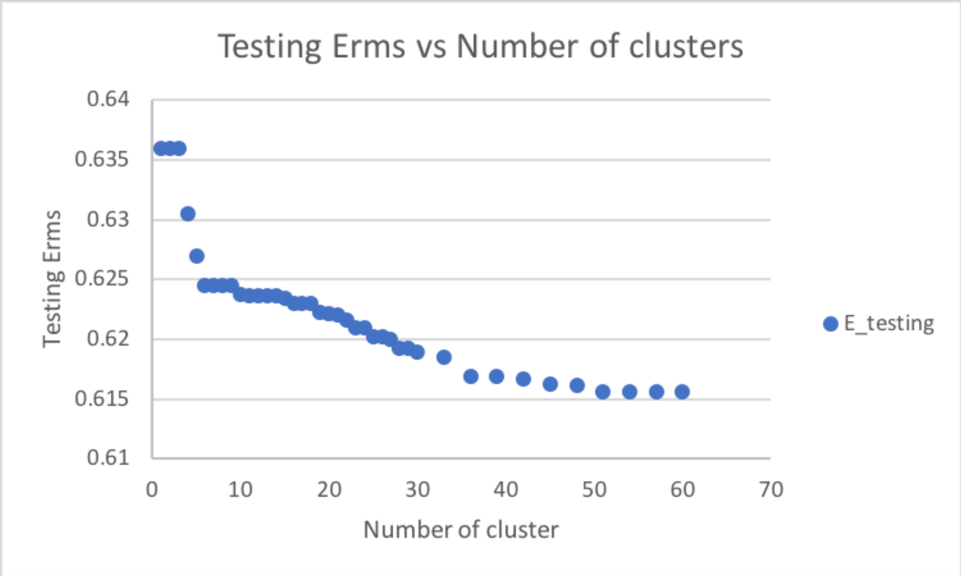





Testing Accuracy vs Number of clusters
Testing Accuracy
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
Number of cluster
● Testing Accuracy


Training Erms vs Number of clusters
Training Erms
0.57
0.565
0.56
0.555
0.55
0.545
0.54
0.535
0
10
20
30
40
50
60
70
Number of cluster
● E_training

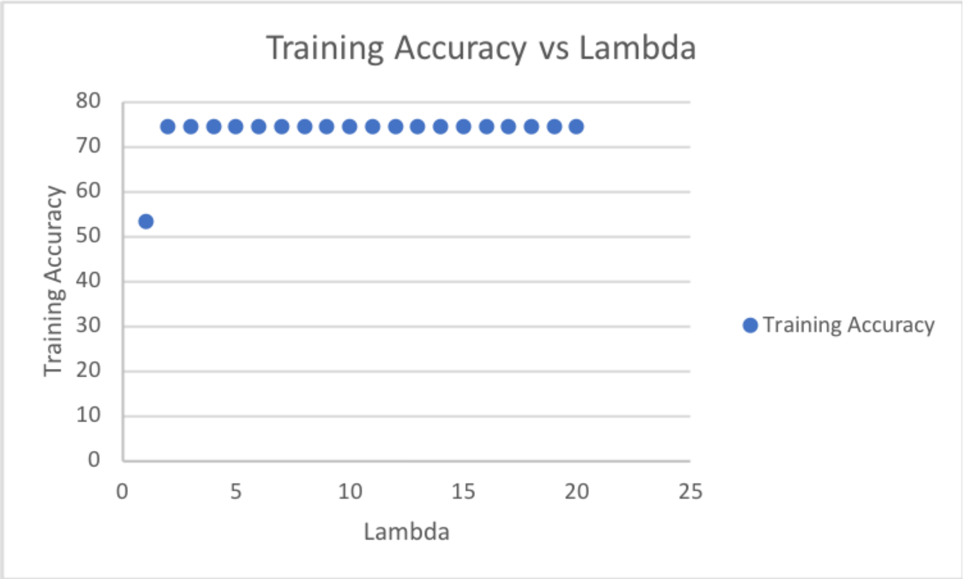


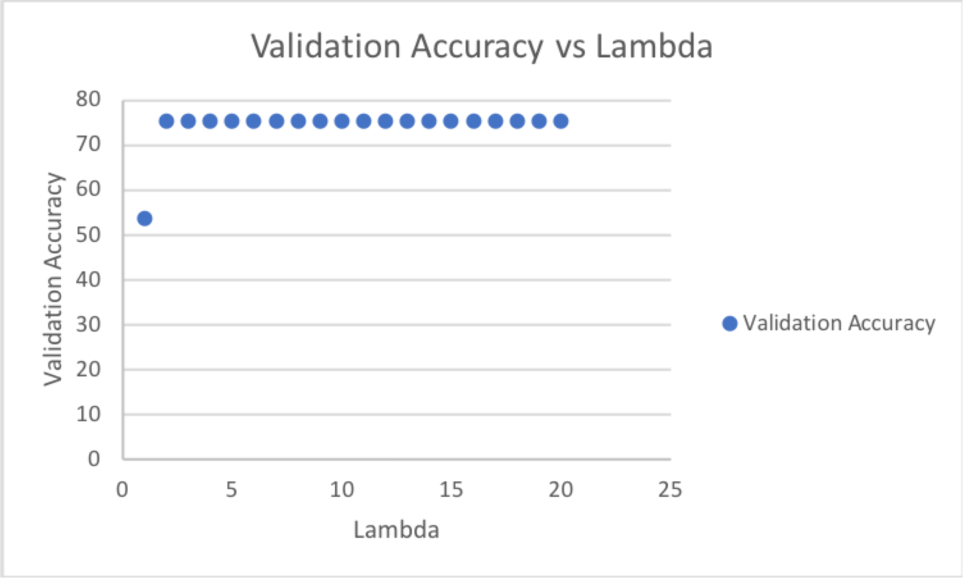


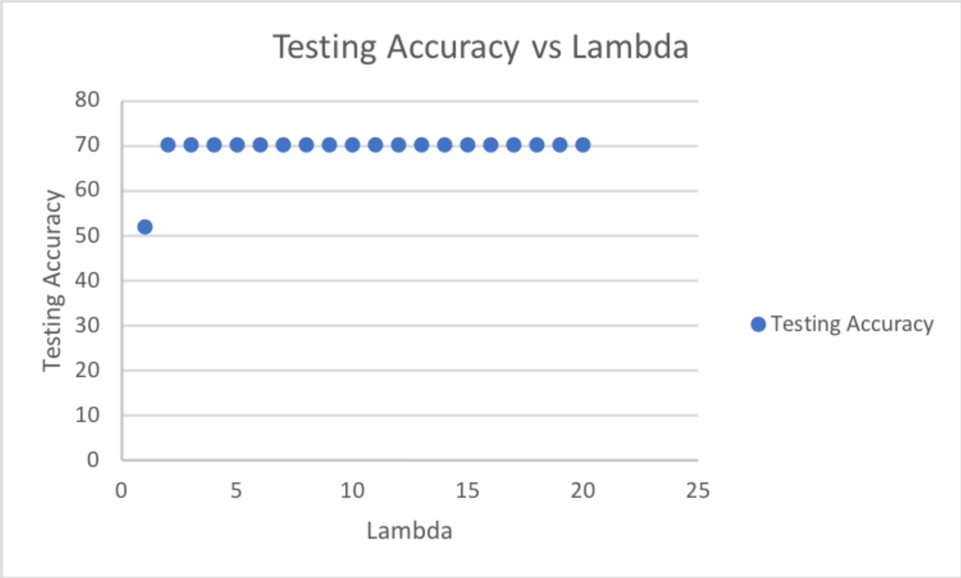
3.1.1 Conclusion

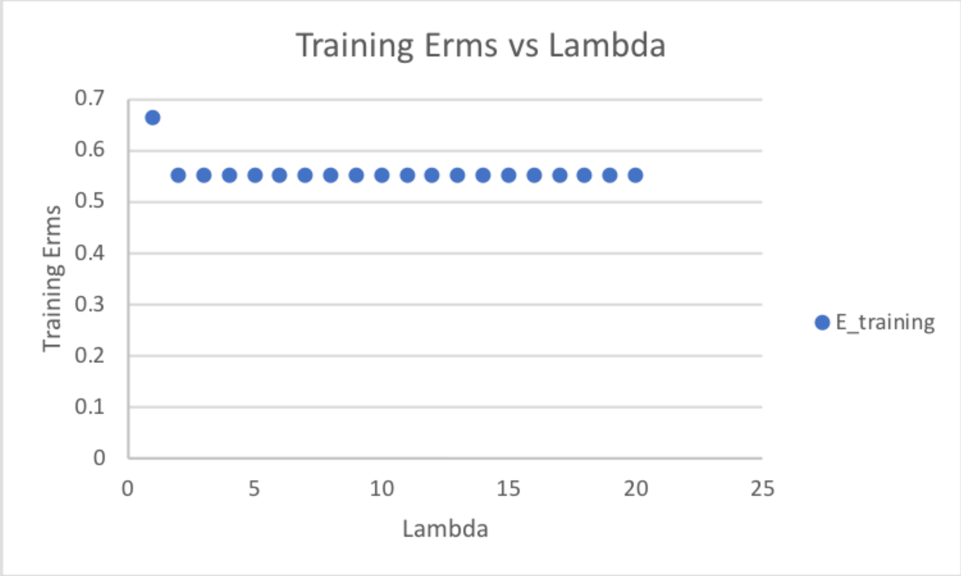
As shown in graphic, all three accuracy have constant value. During the test, I found that accuracy and Erms have really little change in general while Erm has an dramaticall change at some centain point around 5 clusters. In order to have better understanding on the influence of clusters to the linear regression model, I use an short interval in the range of 1 to 30 and a longer interval in the range of 30 to 60. It appeals that the change of clusters may not have an significant effect on the accuracy. In contrast, the Erms will framatically decrease with the clusters number increments, but they also seem to stay in constant after 30

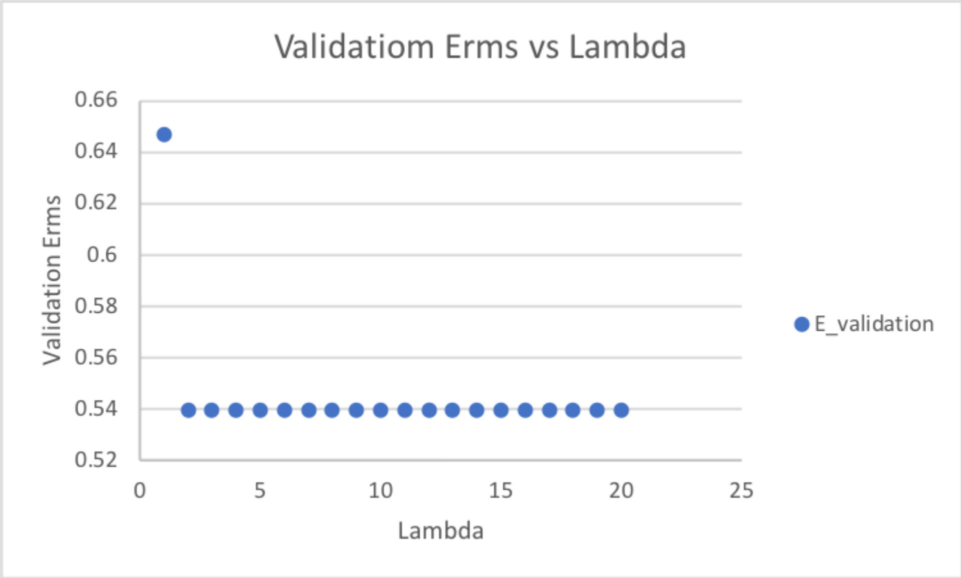
3.2 Lambda

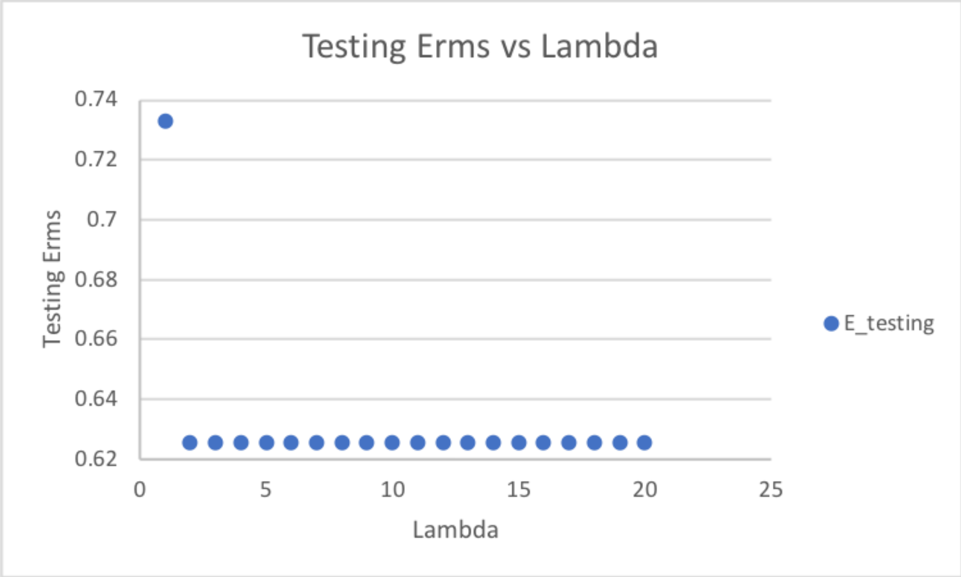

Training Accuracy vs Lambda
Training Accuracy
0
10
20
30
40
50
60
70
80
0
5
10
15
20
25
Lambda
● Training Accuracy


Validation Accuracy vs Lambda
Validation Accuracy
0
10
20
30
40
50
60
70
80
0
5
10
15
20
25
Lambda
● Validation Accuracy


Testing Accuracy vs Lambda
Testing Accuracy
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
Lambda
● Testing Accuracy


Training Erms vs Lambda
Training Erms
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
Lambda
● E_training



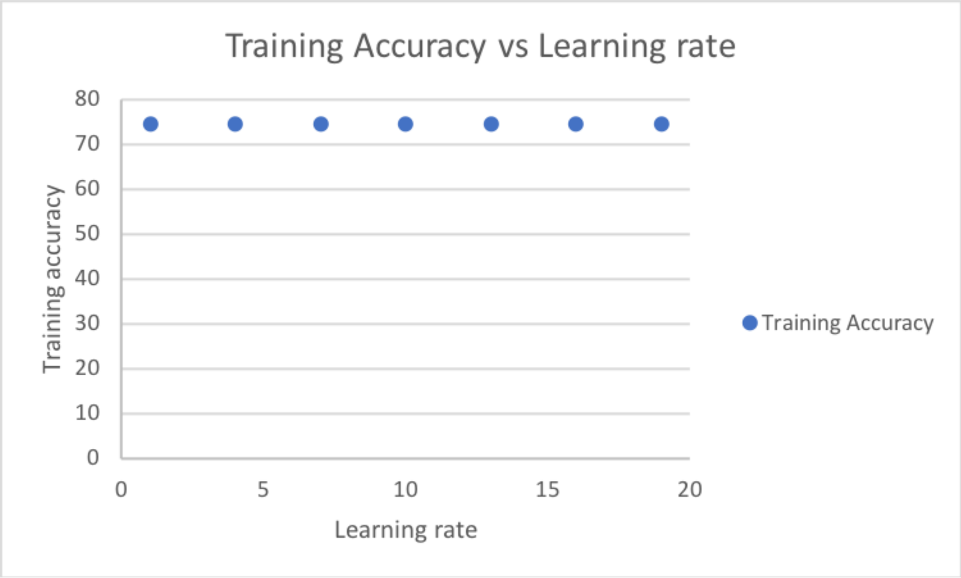


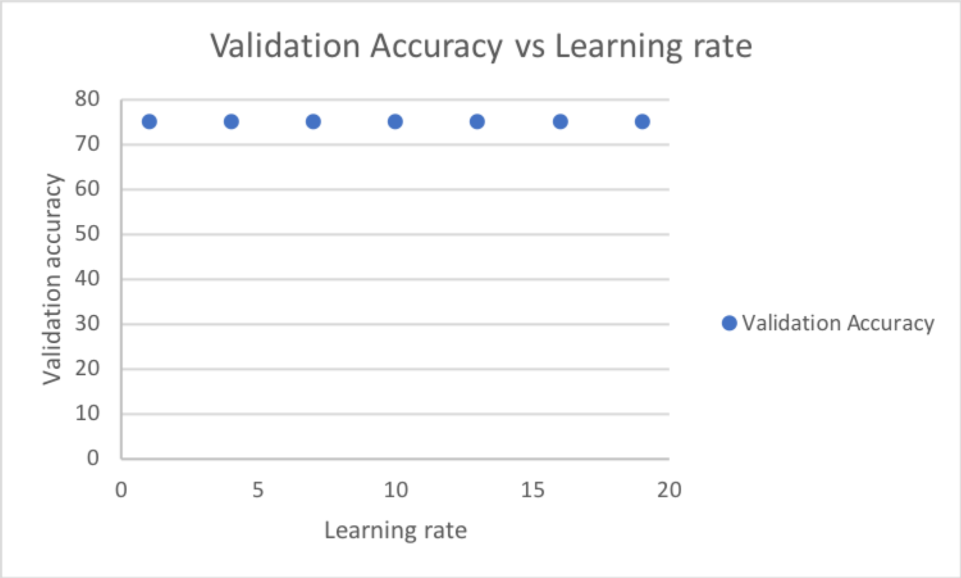
3.2.1 Conclusion

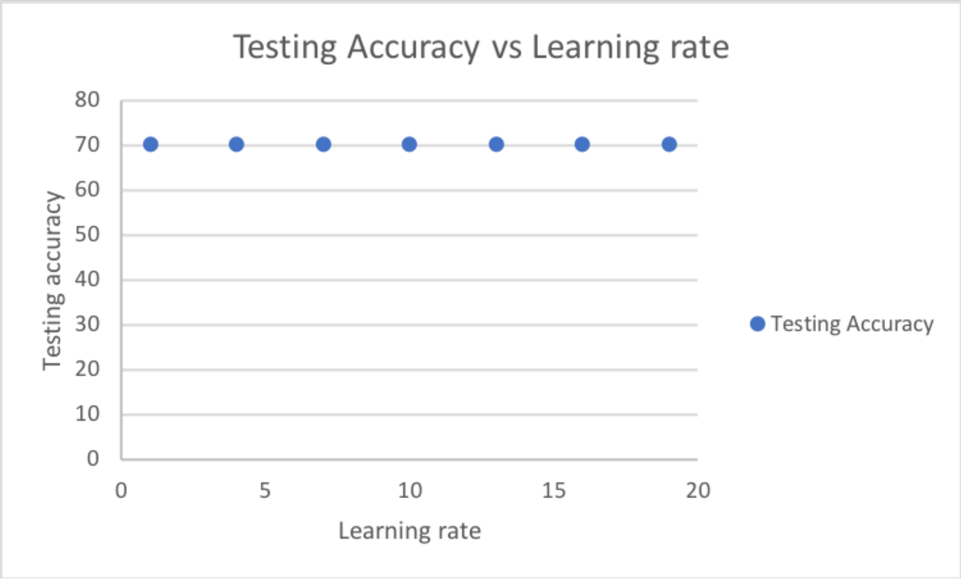
During the test, I found that the value of Lambda doesn't change the accuracy and Erms after Lambda 2. As shown above, all three accuracy and three Erms seems to stay in constant. Accuracy seems to have a lower value at lambda 1 while the Erms have a higher value. This may be an result of applying the Lambda value to the regularization.

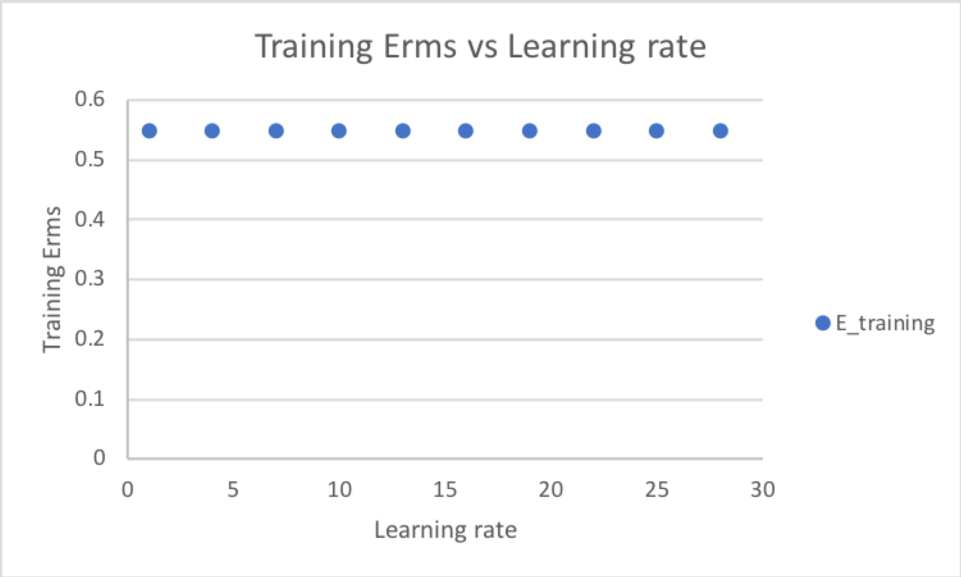
3.3 Learning Rate

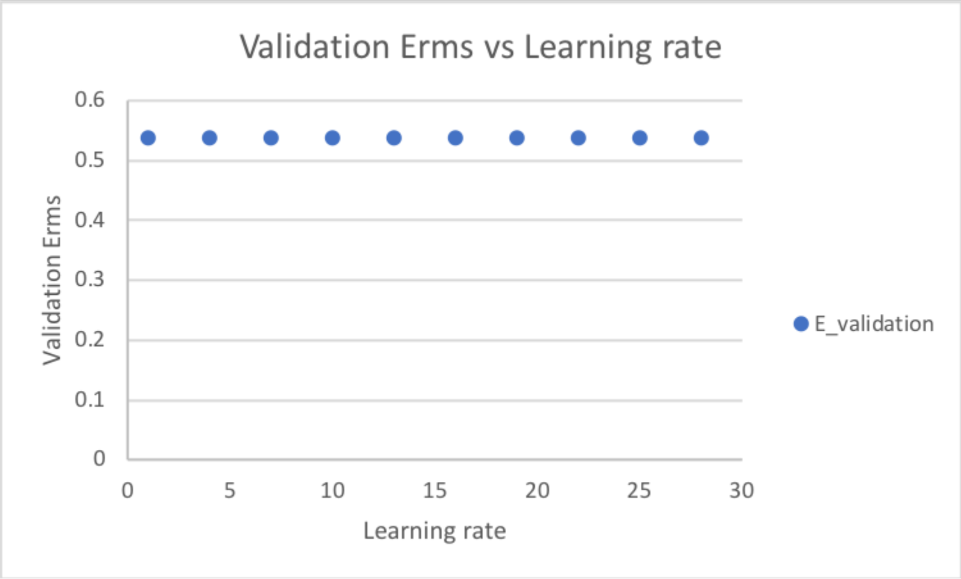
By Applying the learning rate in the Stochastic Gradient Decent, we can control how much the coeffiecients changes or learns each time it is updated.

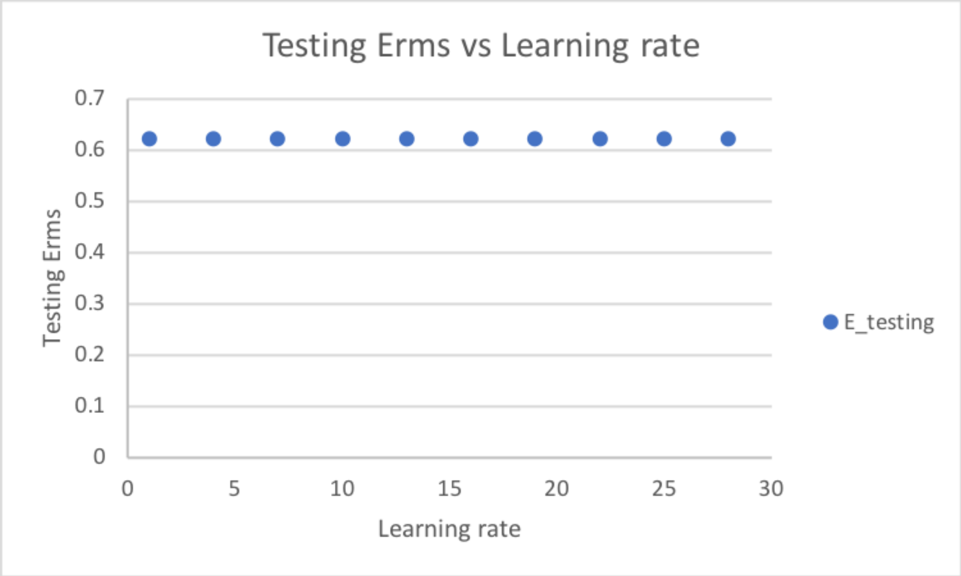





Testing Accuracy vs Learning rate
80
70
60
50
40
30
20
10
0
Testing accuracy
0
5
10
15
20
Learning rate
● Testing Accuracy


Training Erms vs Learning rate
0.6
0.5
0.4
0.3
0.2
0.1
0
Training Erms
0
5
10
15
20
25
30
Learning rate
● E_training



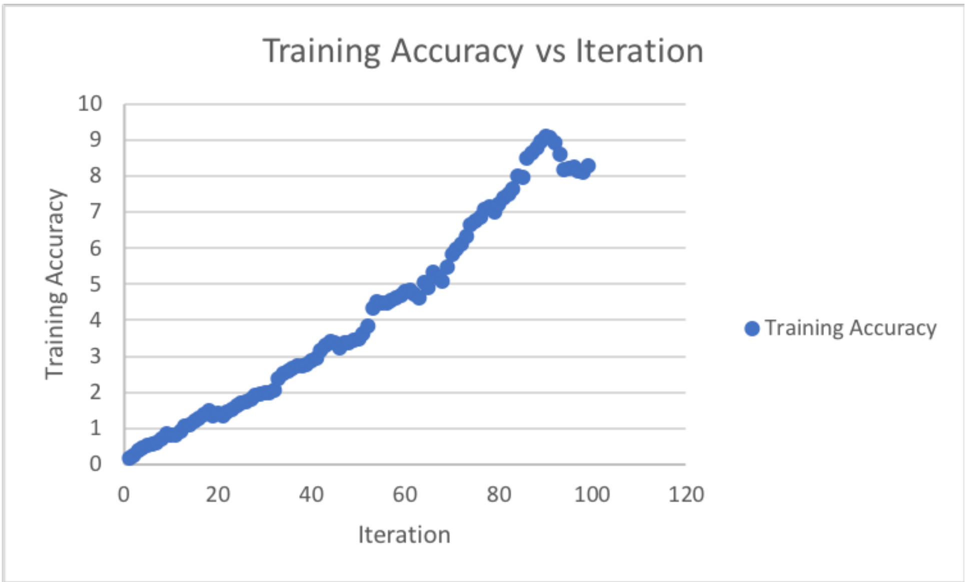


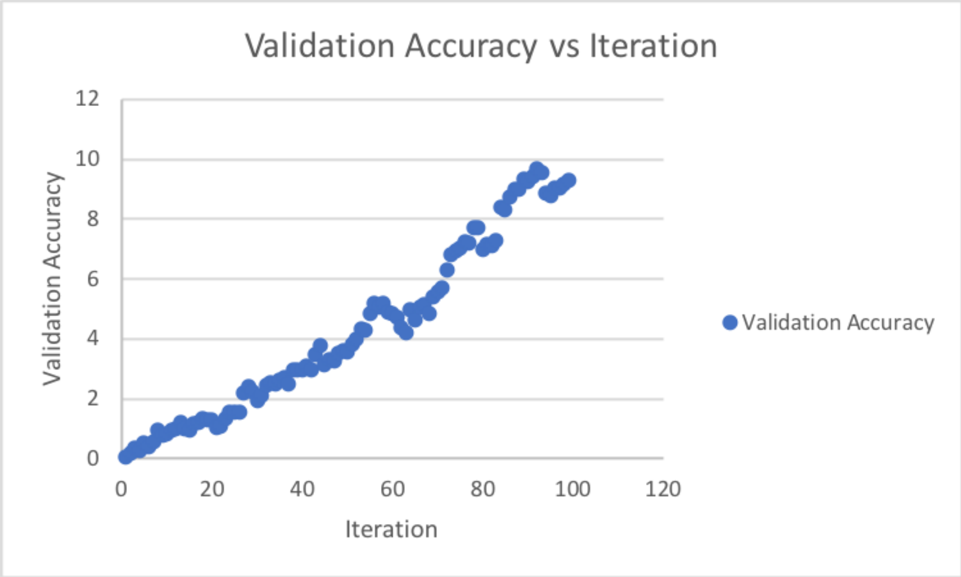
3.3.1 Conclusion

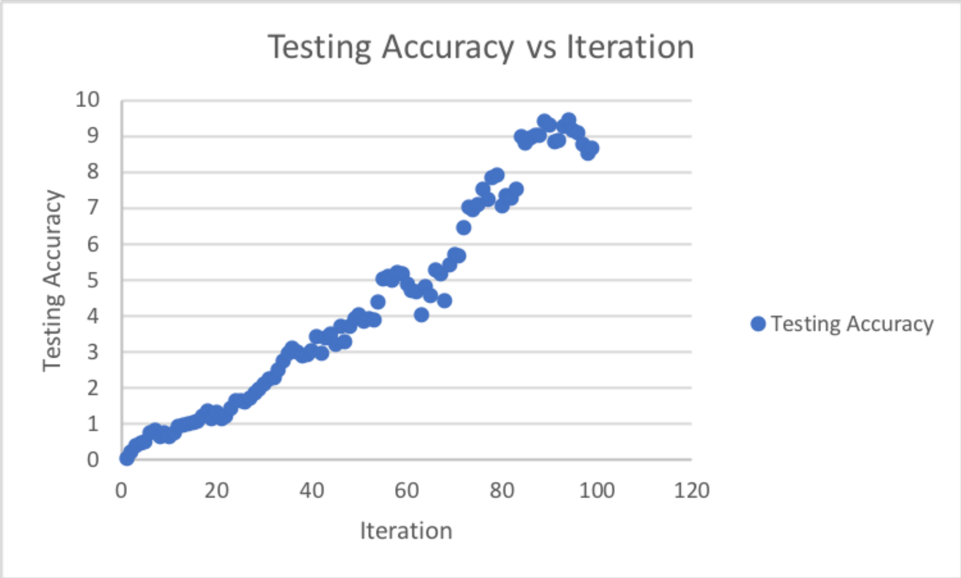
In order to set Gradient Decent work, we need to set the appropriate value for the learning rate. As shown above, all the accuracy and Erms have nearly constant value. This may be because when the learning rate is too high, we skip the optimal solution, and when the learning rate is too low, we will need more iteration to coverage to the best values. In our test, learning rate stays within the appropriate value so the accuracy and Erms stay constant.

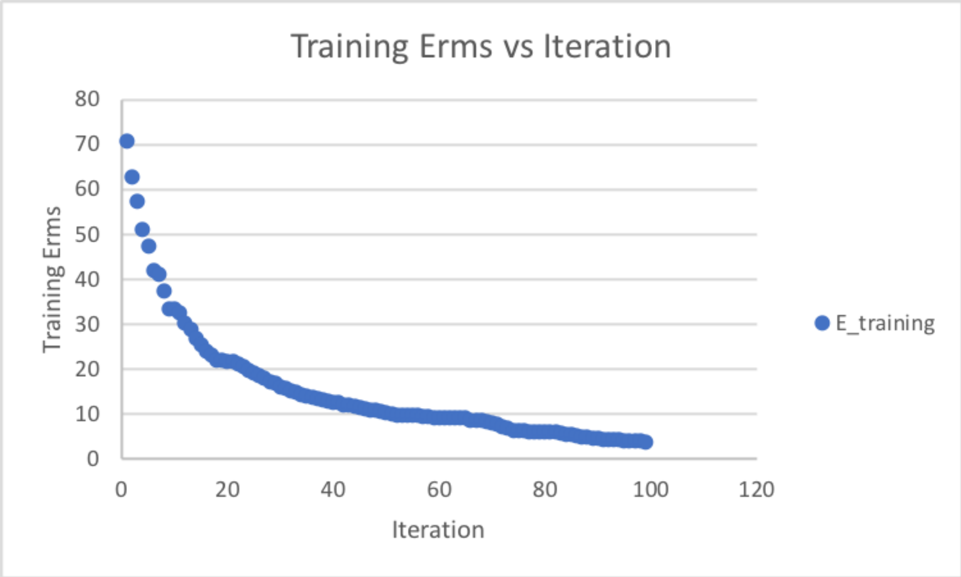
3.4 Iteration

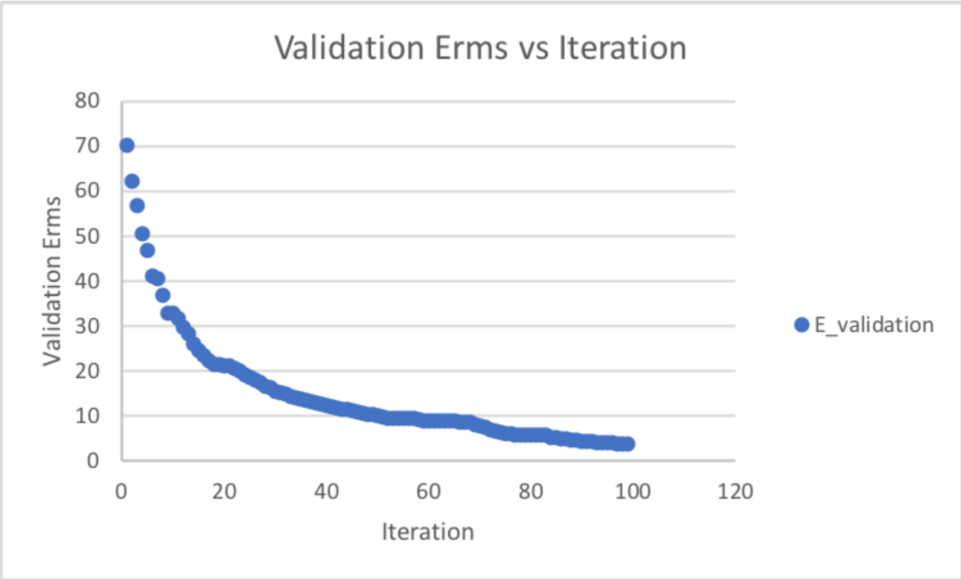
As Gradient decent is a first-order iterative optimization algorithm, the number of iteration decided how many iterations will be put into the optimization algorithm. As it needs many iterations to compute the accuracy, in our test we are running as many times as the iteration to constantly update the weights and biases. A good amount or iteration is expected from 1 to 1000. In this session, we will test the accuracy and Erms with iteration changing form 1 to 100.

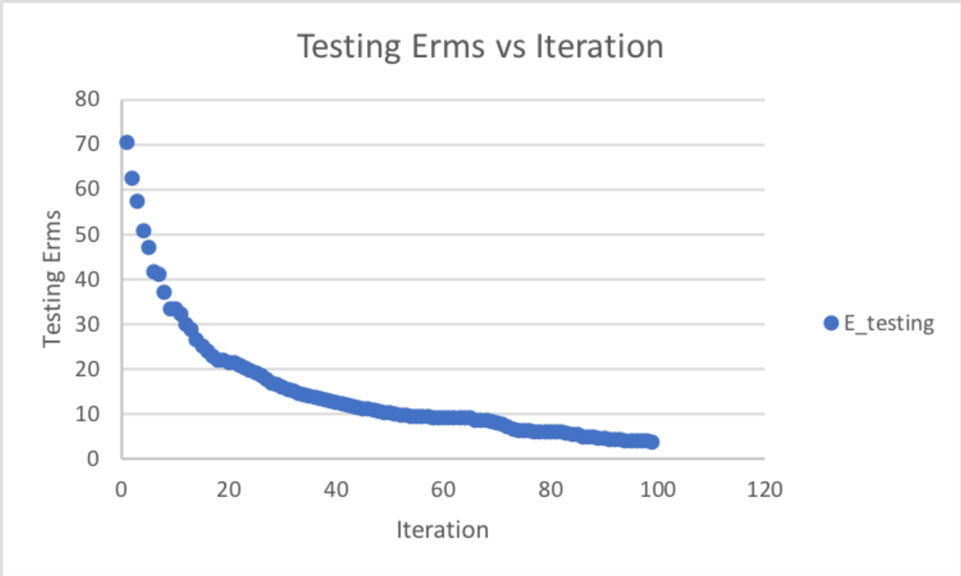





Testing Accuracy vs Iteration
Testing Accuracy
10
9
8
7
6
5
4
3
2
1
0
0
20
40
60
80
100
120
Iteration
● Testing Accuracy


Training Erms vs Iteration
Training Erms
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
Iteration
● E_training





3.4.1 Conclusion

In general, all three accuracy have an increasing rate with the increase on iteration while all three Erms have an decreasing rate. This claims that with the increase of iteration, we can train the linear regression model with more data. While it updates its weights and biases in each iteration, the accuracy will gradually increase and the Erms will decreases.

4 Summary

As a result, we learn how different value of hyper parameter will affect the performance under Closed Form Solution and Stochastic Gradient Decent. While we can increase the accuracy and lower the Erms by setting the appropriate hyper parameter. There are also other ways we could do in the iteration of the Stochastic Gradient Decent. Since we only tested 1 to 100 iteration, there may be more change when the iteration is getting much bigger. Since an ideal Gradient Decent should have an almost constant accuracy and Erms, this value may be viewed when the iteration is higher than 500. In addition, Stochastic Gradient Decent shows an powerful way to train the linear regression, as it gets similar accuracy level with only 100 iteration. Stochastic Gradient Decent needs much less dataset to calculate the overall accuracy and Erms. On the other hand, Stochastic Gradient Decent performs much faster as it just need 100 iterations to get the result. Closed Form Solution is a good approach in our problem solving. However, since Closed Form Solution needs to calculate the inverse matrix, when the dataset is much more bigger, Closed Form Solution may need much higher Computer hardware requirement to perform.

5 Reference Website

http://www.mit.edu/ 9.54/fall14/slides/Class13.pdf
https://en.wikipedia.org/wiki/Overfitting
https://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/
http://blog.datumbox.com/tuning-the-learning-rate-in-gradient-descent/