CSE 474/574: Introduction to Machine Learning

University at Buffalo

Keming Kuang (kemingku)

personal #: 50161776

October 31, 2018

Contents

1 Description

1.1 Objective

In this project, we will learn to solve the handwriting comparison task in forensics. By applying this into a linear regression problem, we can solve it by using stochastic gradient descent.

1.2 Task

With given data feature, we will gather dataset and target value and train a linear regression model on them by using a stochastic gradient descent.

1.3 Plan of Work

Step 1: Data Process Processing data from same pair file and different pair file into one as target value. Data processed from same pair file and different pair file will be the same amount to avoid bias, then we will shuffle the target value to keep it randomized.

Step 2: Human Observed Feature After data processing Human Observed Data, we will have the target value for Human Observed. By applying concatenation and subtraction on the feature pairs, we will have 18 pairs features row data from concatenation and 9 pairs features row data from subtraction.

Step 3: GSC Feature Similar to above, we will gather the target value for GSC Feature by data processing GSC Feature Data. By applying concatenation and subtraction on the feature pairs, we will have 1024 pairs features row data from concatenation and 512 pairs features row data from subtraction.

Step 4: Training Data with Linear Regression Using Stochastic Gradient Descent to train the regression model with a group of hyper-parameters. The input target vector and row data are gathered through Human Observed Feature Data and GSC Feature Data while applying concatenation and subtraction on each data pairs.

2 Linear Regression

In this section, I will train the gathered data set in a Stochastic Gradient Decent method. The objective is finding the similarity between the handwritten samples of the known and the unknown writer by using linear regression. We will also study the model accuracy by testing between hyper-parameters.
By default, I will use
k-means = 10
Lambda = 2
Learning Rate = 0.01
iteration = 400

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 as a problem of 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 the number of clusters vary from 1 to 10.

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.

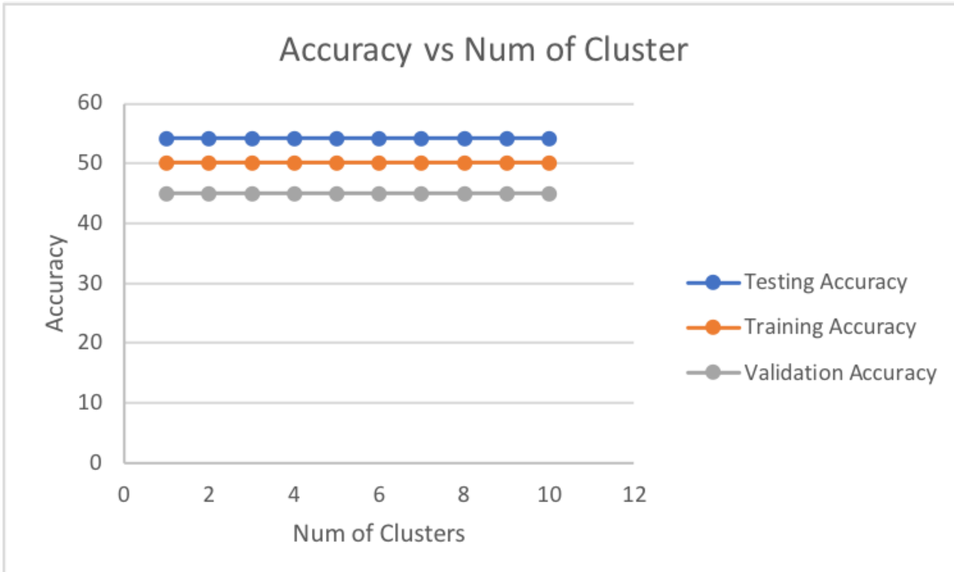
Lambda Since the criterion used for selecting the model is not the same as the criterion used to judge the suitability of the model. In our problem solving, overfitting may occur when the model capture points that particularly has significant difference within its main data group, and this usually results 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.

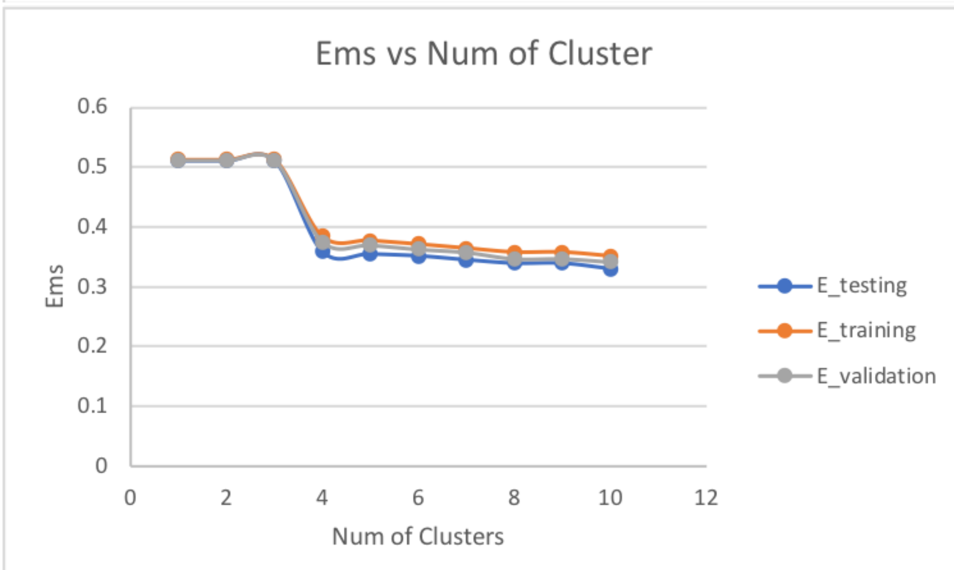
Human Observed Features Human Observed Features is collected through the features entered by human document examiner. The target data is sorted in the step of data process, in which we shuffled the same amount of same pair data set and different pair data set together to avoid bias. The row data will be trained under 2 different feature extraction: Human Observed Dataset with feature concatenation, Human Observed Dataset with feature subtraction.

GSC Features GSC Features is collected through the features generated by Gradient Structural Concavity algorithm. The target data is sorted in the step of data process, in which we shuffled the same amount of same pair data set and different pair data set together to avoid bias. The row data will be trained under 2 different feature extraction: GSC Dataset with feature concatenation, GSC Dataset with feature subtraction.

2.1 Human Observed Dataset with feature concatenation

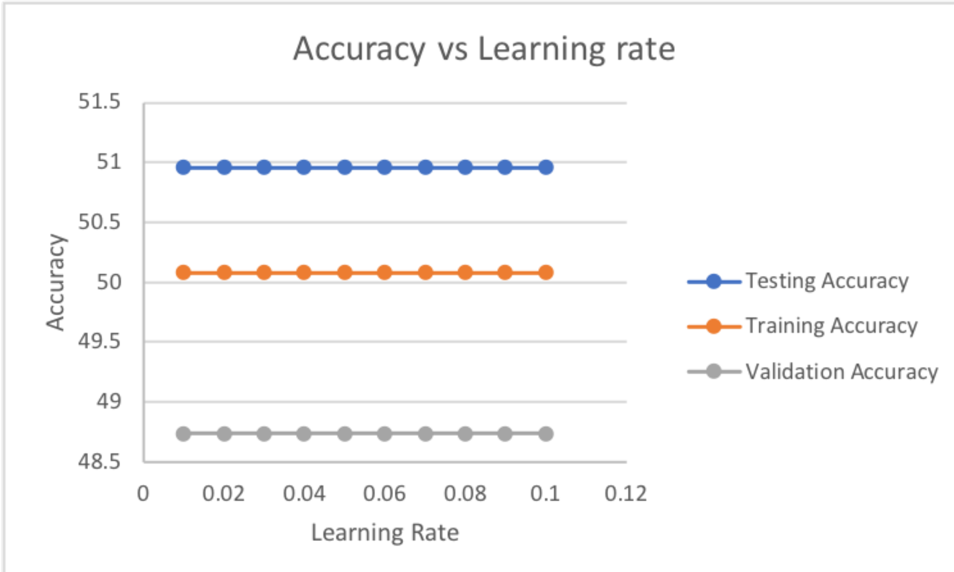
2.1.1 Number of Clusters

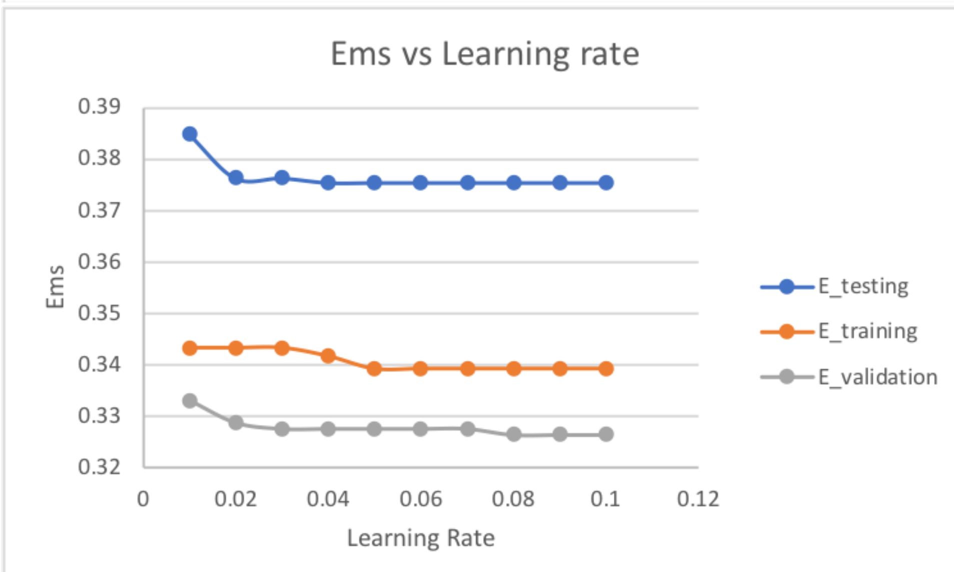




As shown above, we tested the accuracy and Erm vs Number of Cluster for Training Data, Testing data, Validation Data. The number of clusters varies from 1 to 10, and the accuracy stays the same, but the Erm decrease significantly between the number of clusters 3 to 4.

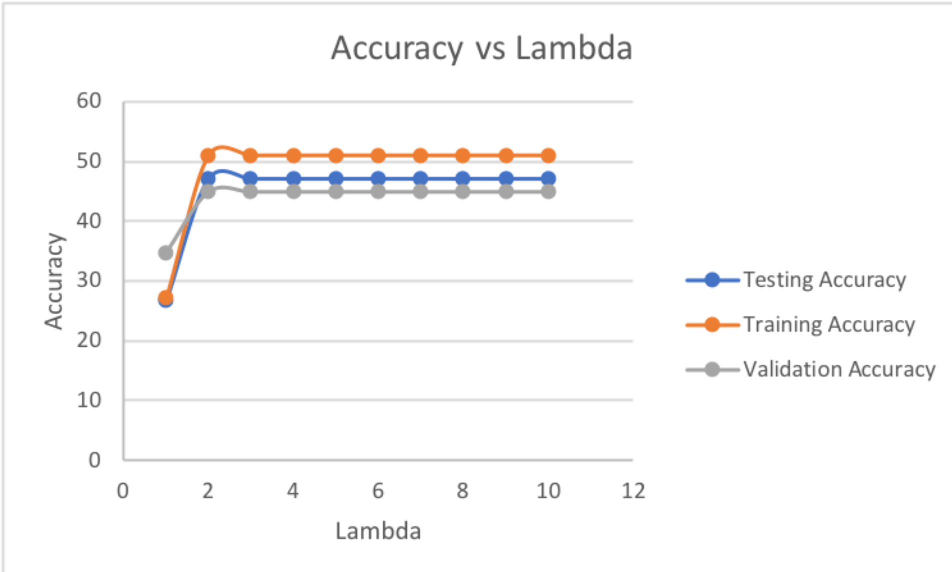
2.1.2 Learning Rate

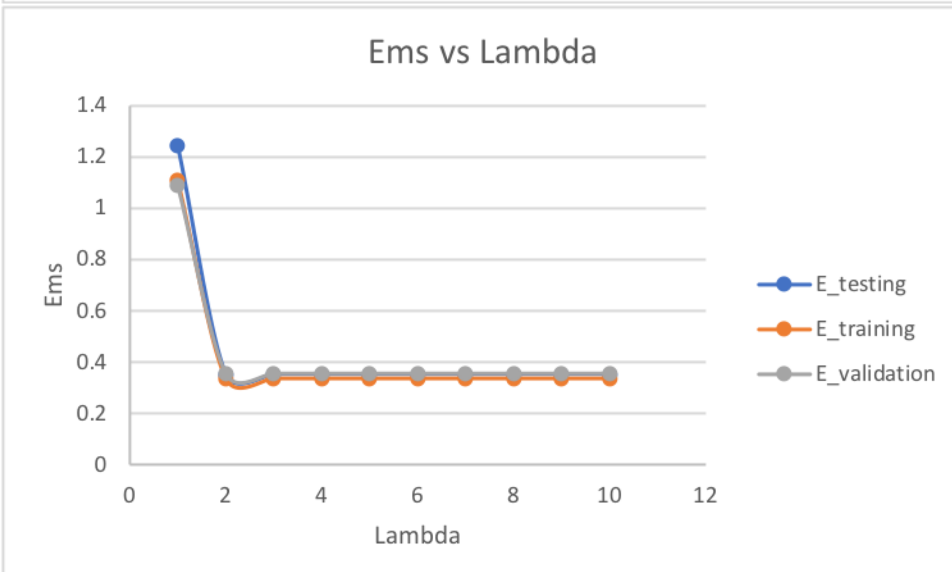




As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The learning rate varies from 0.01 to 0.1, and the accuracy stays the same. Testing data Erm drops significantly from learning rate 0.01 to 0.02, while training data Erm and validation data Erm have slight decrease over the learning rate. I believe the overall decrease trend on Erm is most likely caused by low learning rate at the beginning.

2.1.3 Lambda

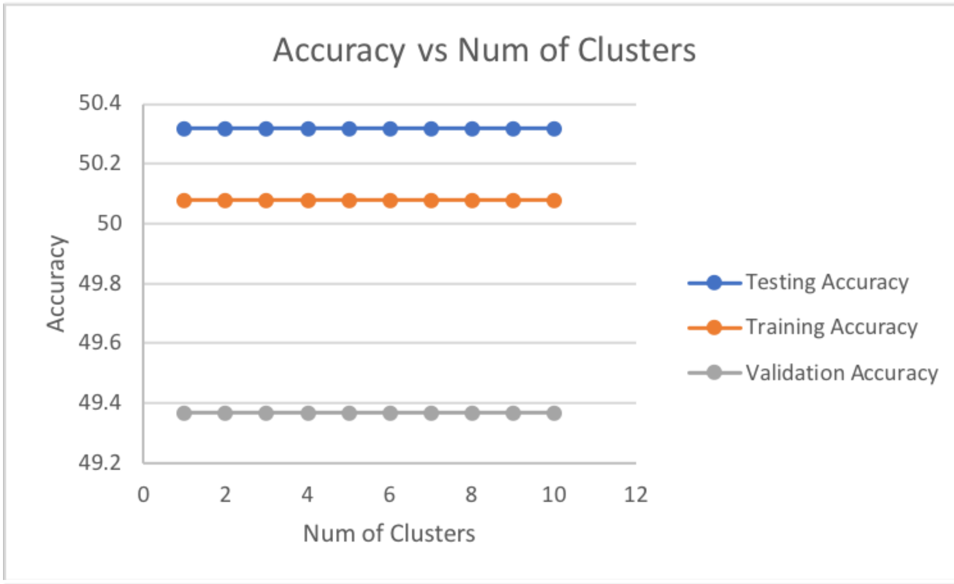


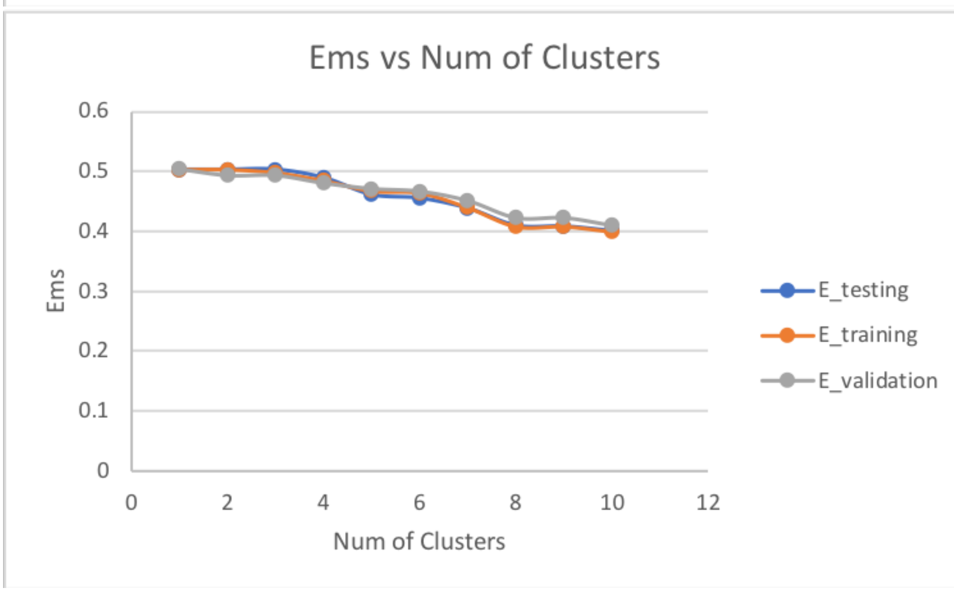


As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The Lambda varies from 1 to 10, and all three accuracy increase dramatically from lambda 1 to 2, but the Erm decrease significantly between the number of clusters 1 to 2. This change on accuracy and Erm happens at the same value boundary between 1 and 2, and this may be a result overfitting with improper Lambda value.

2.2 Human Observed Dataset with feature Subtraction

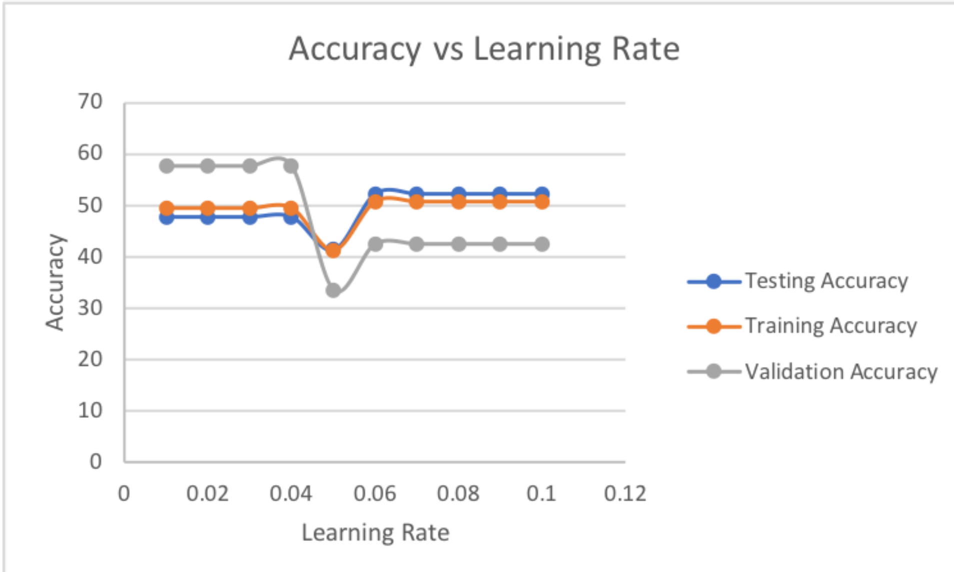
2.2.1 Number of Clusters





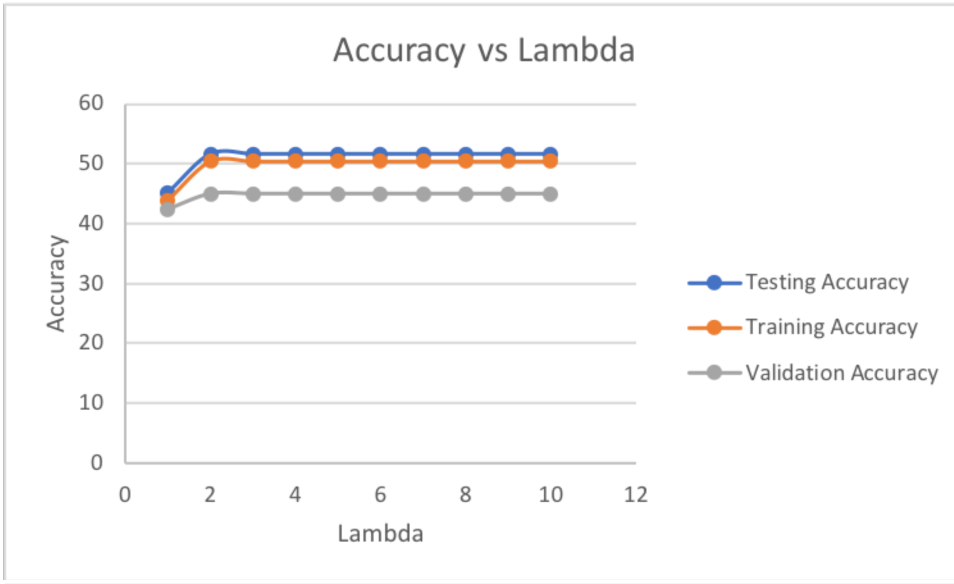
As shown above, we tested the accuracy and Erm vs Number of Cluster for Training Data, Testing data, Validation Data. The number of clusters varies from 1 to 10, and the accuracy stays the same, but the Erm decrease gradually over number of clusters.

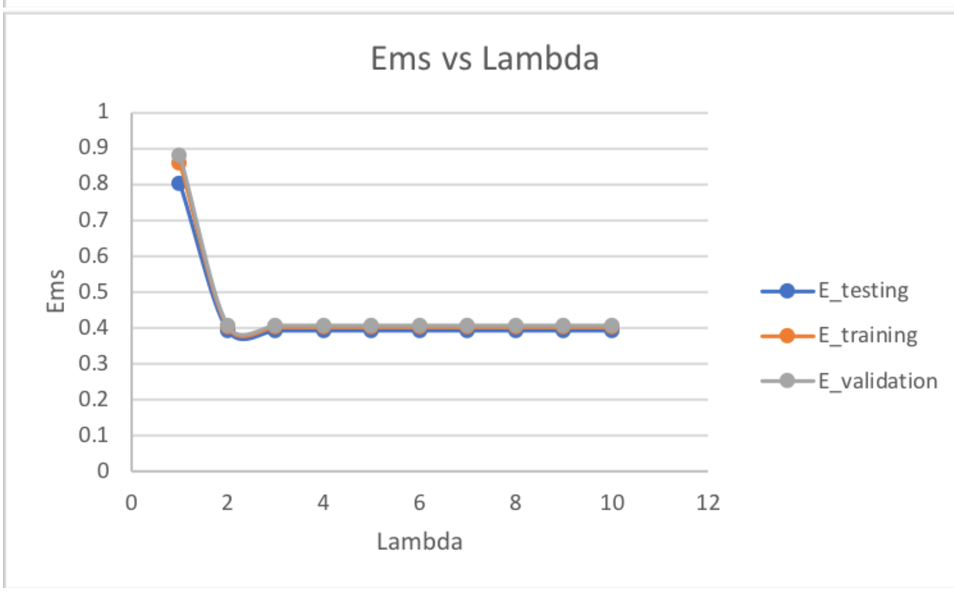
2.2.2 Learning Rate



As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The learning rate varies from 0.01 to 0.1, and the accuracy changes dramatically from learning rate 0.04 to 0.06. Testing data Erm and training data Erm drops significantly from learning rate 0.01 to 0.02, while validation data Erm have slight decrease over the learning rate. I believe the overall decrease trend on Erm is most likely caused by low learning rate at the beginning.

2.2.3 Lambda

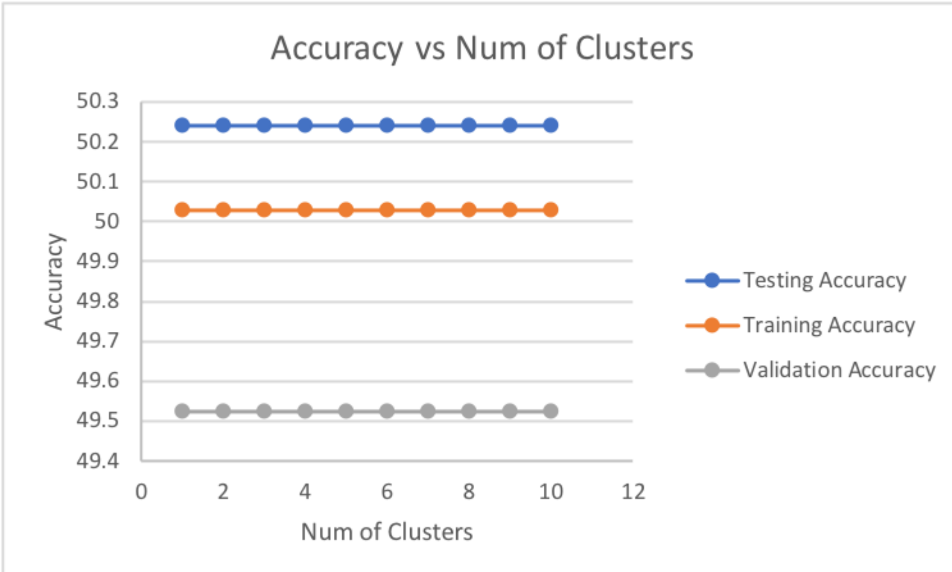


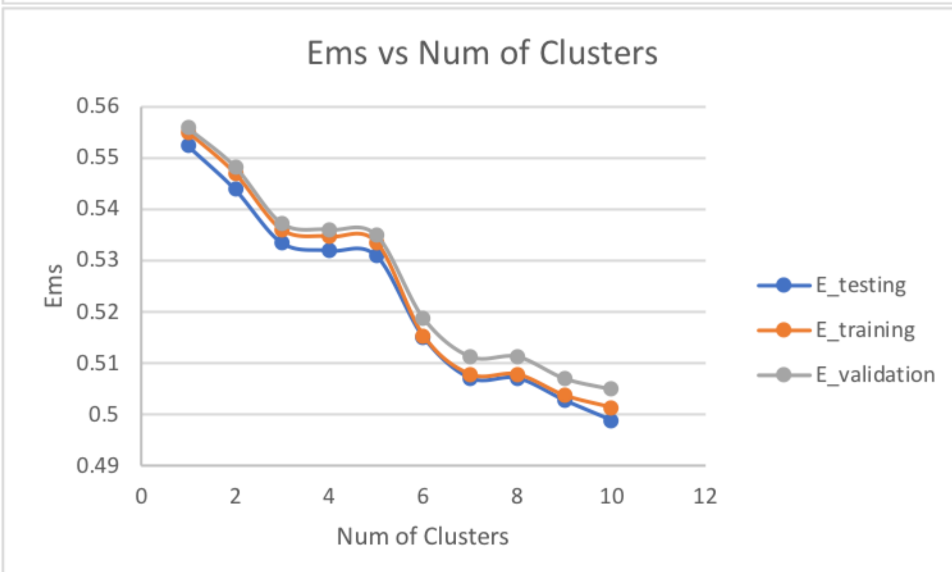


As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The Lambda varies from 1 to 10, and all three accuracy increase dramatically from lambda 1 to 2, but the Erm decrease significantly between the number of clusters 1 to 2. This change on accuracy and Erm happens at the same value boundary between 1 and 2, and this may be a result overfitting with improper Lambda value.

2.3 GSC Dataset with feature concatenation

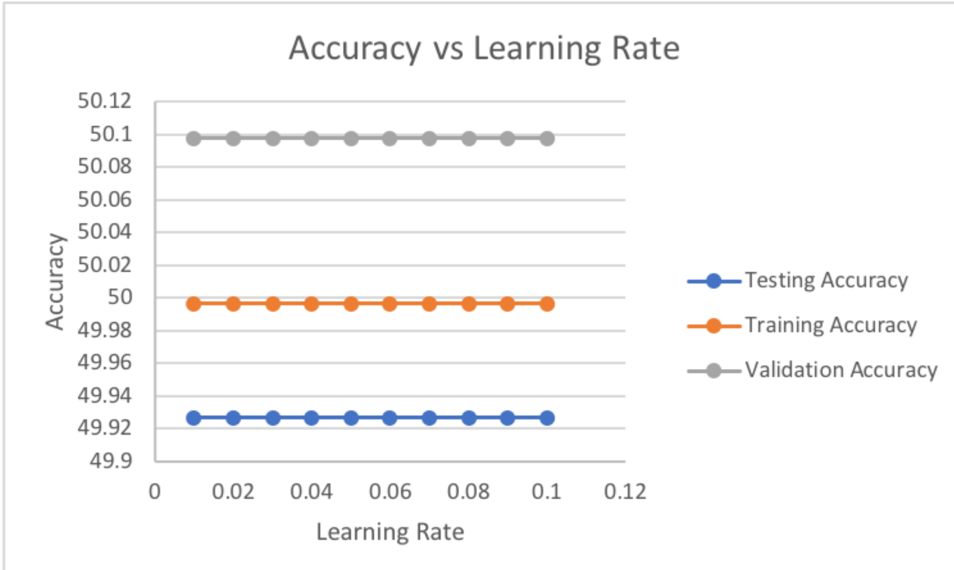
2.3.1 Number of Clusters

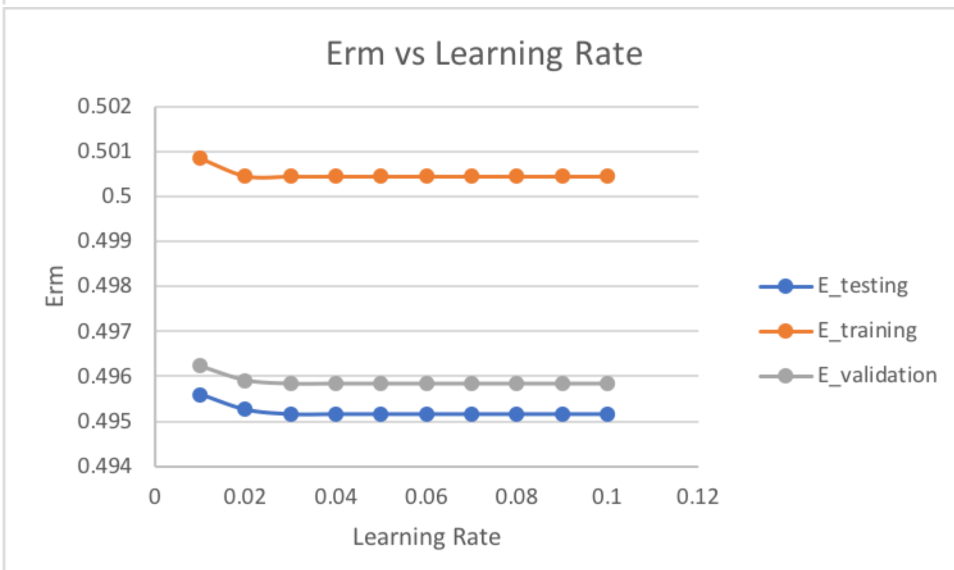




As shown above, we tested the accuracy and Erm vs Number of Cluster for Training Data, Testing data, Validation Data. The number of clusters varies from 1 to 10, and the accuracy stays the same, but the Erm decrease over number of clusters.

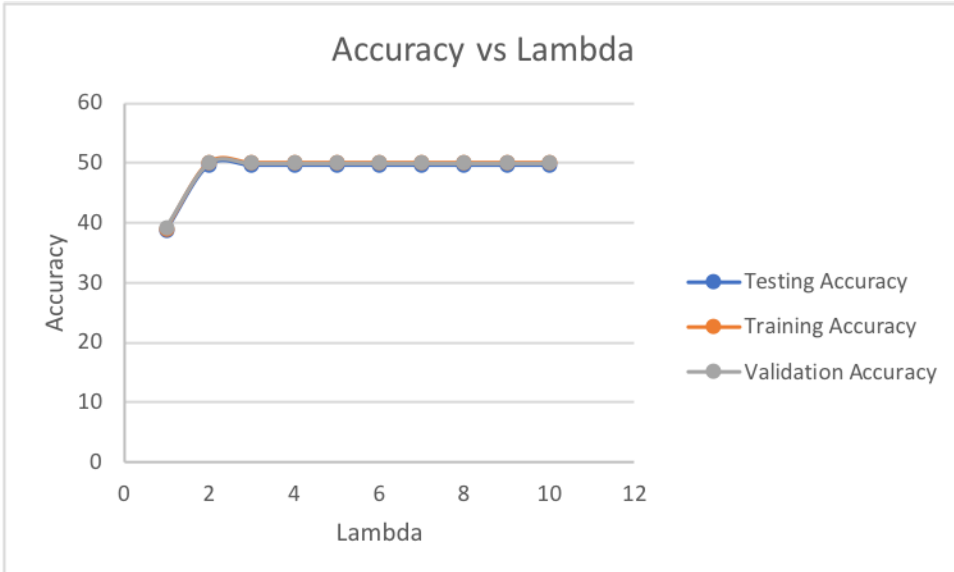
2.3.2 Learning Rate

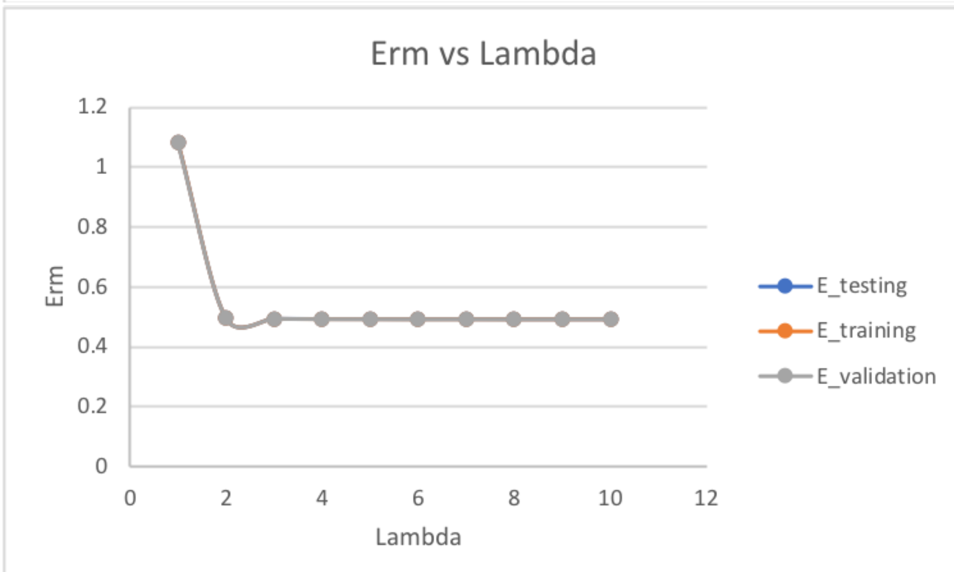




As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The learning rate varies from 0.01 to 0.1, and the accuracy stays the same. Testing data Erm drops slightly from learning rate 0.01 to 0.02, while training data Erm and validation data Erm have slight decrease over the learning rate. I believe the overall decrease trend on Erm is most likely caused by low learning rate at the beginning.

2.3.3 Lambda

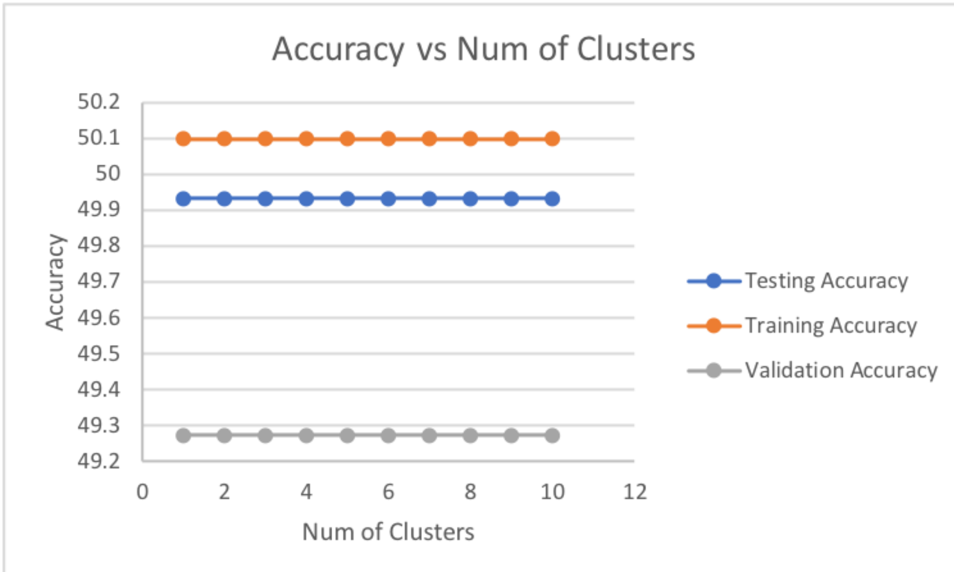


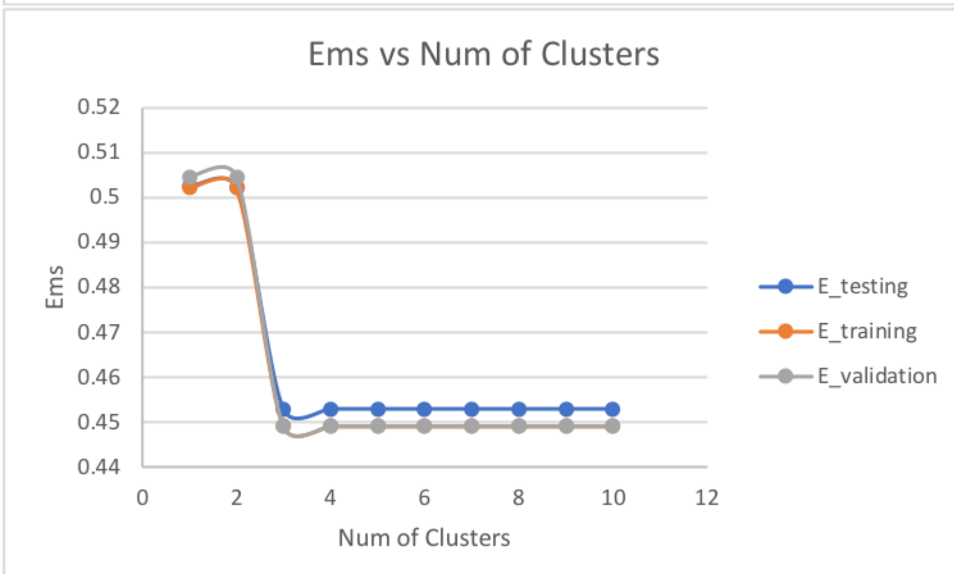


As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The Lambda varies from 1 to 10, and all three accuracy increase dramatically from lambda 1 to 2, but the Erm decrease significantly between the number of clusters 1 to 2. This change on accuracy and Erm happens at the same value boundary between 1 and 2, and this may be a result overfitting with improper Lambda value.

2.4 GSC Dataset with feature Subtraction

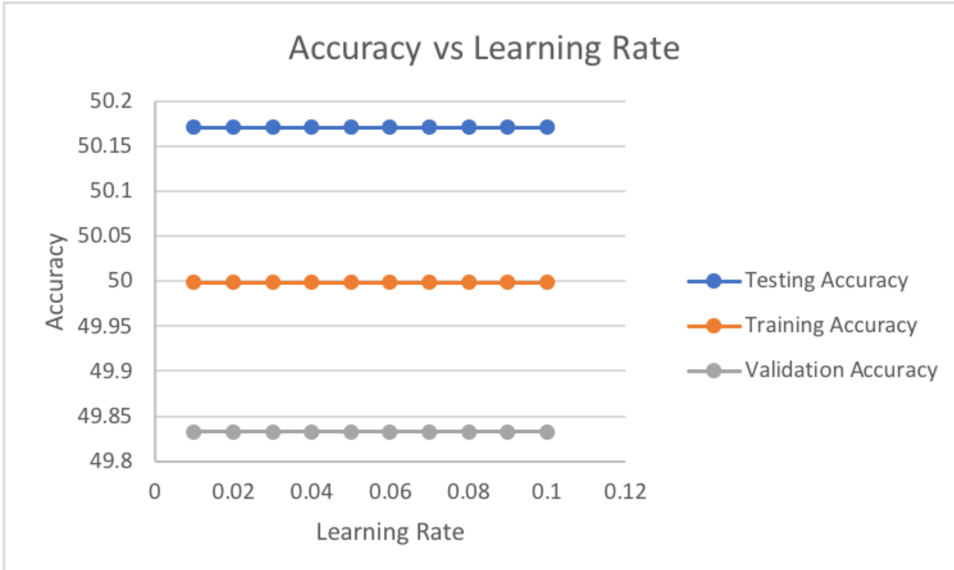
2.4.1 Number of Clusters

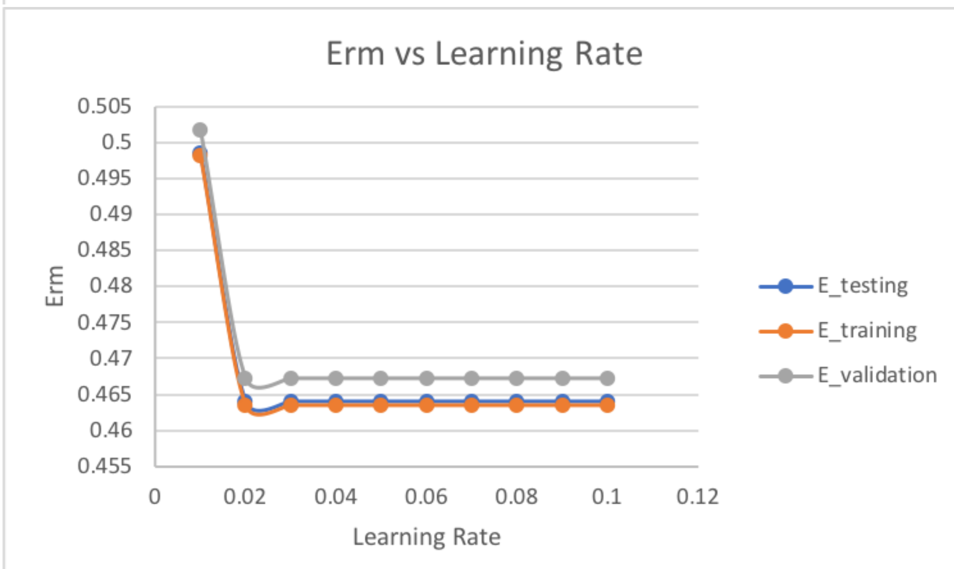




As shown above, we tested the accuracy and Erm vs Number of Cluster for Training Data, Testing data, Validation Data. The number of clusters varies from 1 to 10, and the accuracy stays the same, but the Erm decrease significantly from number of clusters 2 to 3. This great improvement on Erm may result in proper value of number of clusters.

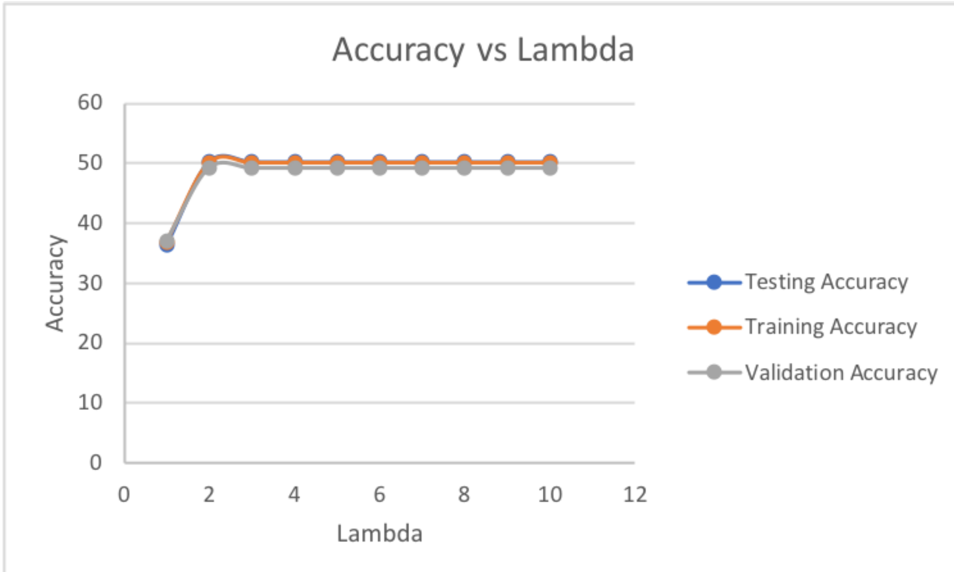
2.4.2 Learning Rate

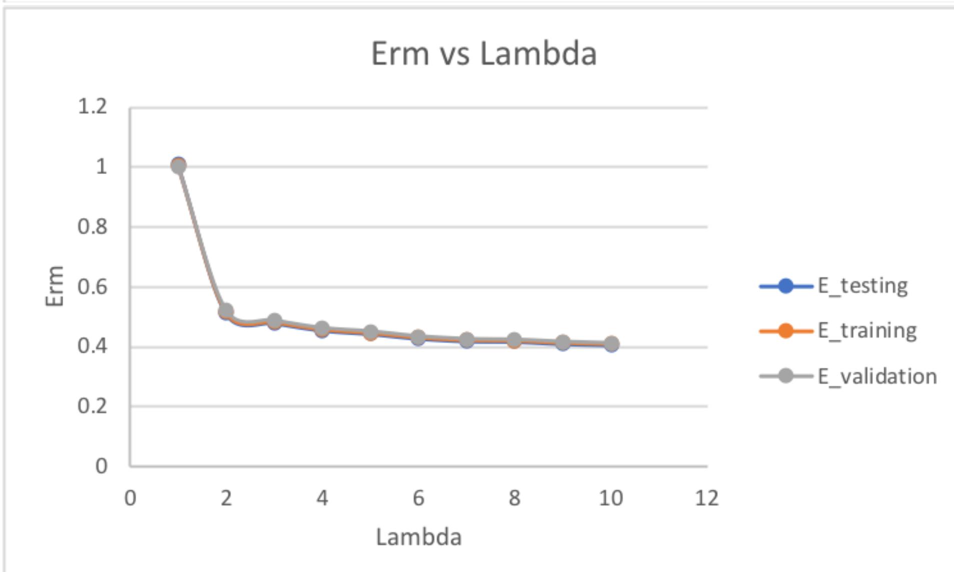




As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The learning rate varies from 0.01 to 0.1, and the accuracy stay the same. Erm drops dramatically from learning rate 0.01 to 0.02. I believe the overall decrease trend on Erm is most likely caused by low learning rate at the beginning.

2.4.3 Lambda





As shown above, we tested the accuracy and Erm vs Lambda for Training Data, Testing data, Validation Data. The Lambda varies from 1 to 10, and all three accuracy increase dramatically from lambda 1 to 2, but the Erm decrease significantly between the number of clusters 1 to 2. This change on accuracy and Erm happens at the same value boundary between 1 and 2, and this may be a result overfitting with improper Lambda value.

3 Summary

As a result, we learn how to solve the solve the handwriting comparison task in forensics. By training the linear regression model with the right data set and setting proper hyper-parameters, we can solve the problem with gradient descent. While the gradient descent has fair amount of Erm, we learn how different value of hyper parameter will affect the performance under 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. We also understand the difference between linear regression and logistic regression by solving the problem step by step. While Logistic regression is more like a classification algorithm used to assign observation to a discrete set of classes, it calculate the result as classified into 1 or 0, which is a representation of the probability.

4 Reference Website

https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html
https://machinelearningmastery.com/logistic-regression-for-machine-learning/
https://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/
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/