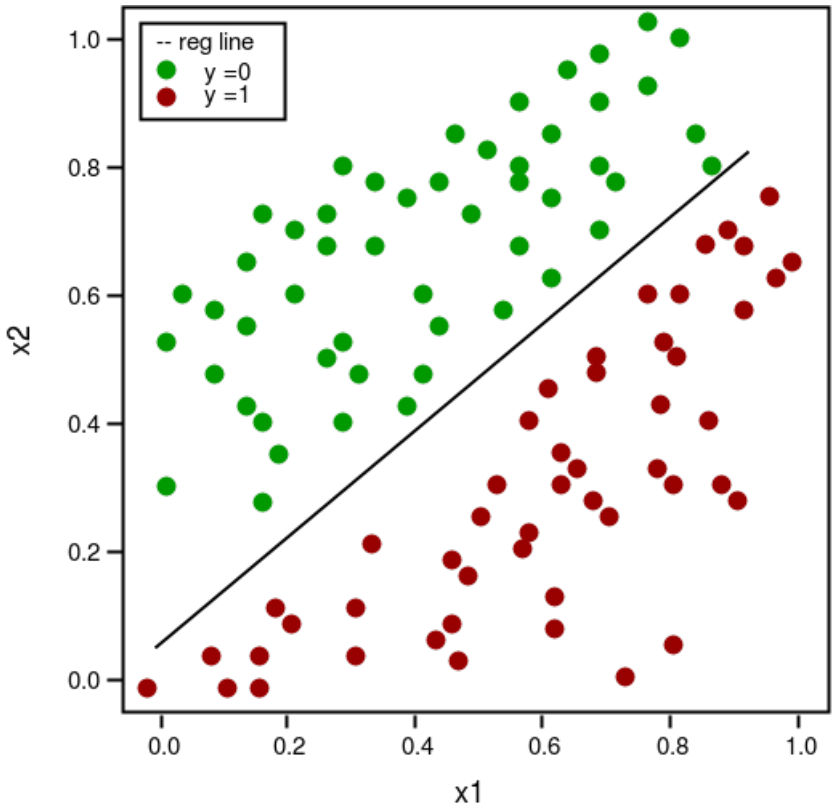
Logistic Regression

Definition

Logistics Regression is fundamentally classification technique. We can understand intuition behind logistic regression using Geometric, probabilistics, and Loss function.

Objective

We have to find the plane(2-D)/hyperplane(n-D) which classify most of data points correctly and avoid misclassified points



We can see that all the green and red points are correctly classified

Assumption of Logistic Regression

Logistic regression assumes the data is linearly separable or almost linearly separable using linear surface.

Plane representation

plane can represent by two part (w,b) where w is normal to plane and b is intercept term. If plane passes through origin then b will be zero.

The general formula of plane is : (W^T*X+b)= 0 where W and X belongs to Vector R^d and b belongs to R^1 which is scalar.

Logistic regression finds the distance of each points from plane using formula : W^T*Xi/||W|| but W is normal to plane so ||W||=1

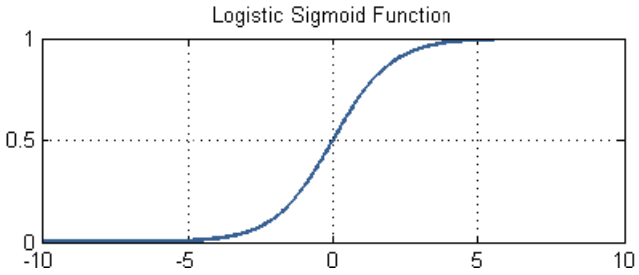
If Yi *W^TX*i>0 then Yi is correctly classfied. If Yi*W^TX*i<0 then Yi is incorrectly classified.

Classifier to be very good if minimum points are misclassified and maximum points are classified.

Plane(PI) defines by W and to find the optimal W we have to maximize the sum of signed distance. But If there is any one outlier , in that case , classifier will be get affected.

To resolve the outlier problem, Squashing techique came into the picture. The idea behind squashing technigue, If signed distance is large we have to make it small. If it is small, we need to use as it is. Using the sigmoid function, we are fullfilling the squashing technique

Sigmoid Function



Formula of sigmoid = (1/1+e^-x). the horizontal line defines the signed distance . If the signed distance will increase then sigmoid function will do tappering off and it dont allow to take large value. that is how, we can overcome the outlier problem

Formula1 : argmax(W)Summation(i=1 to n)(1/1+exp(-YiW^TXi)) which is very hard to compute so to solve the optimization problem we have used the monotonic function log(x).

Formula2 :

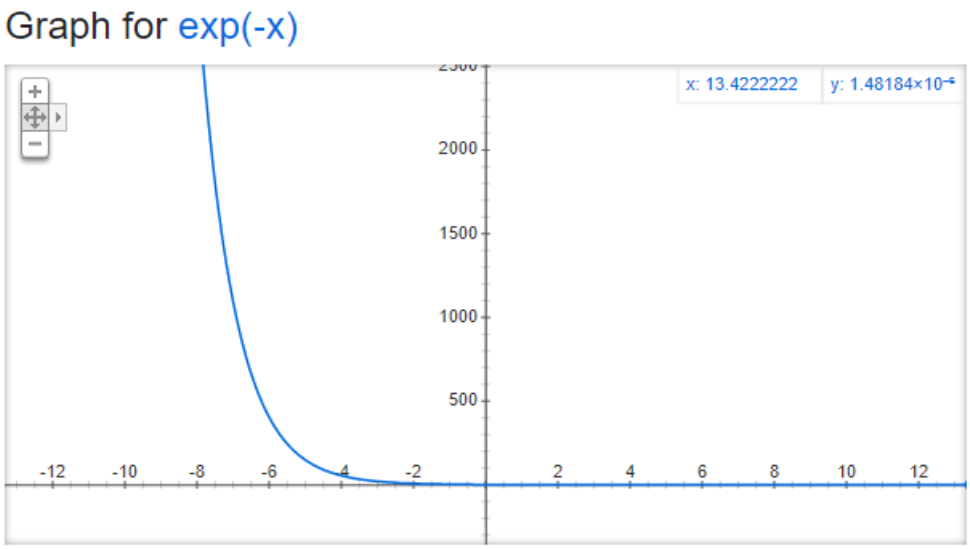
argmax(W)Summation(i=1 to n)log(1/1+exp(-YiW^TXi))

argmax(W)Summation(i=1 to n)log(-(1+exp(-YiW^TXi))

argmax(W)Summation(i=1 to n)log(-(1+exp(-YiW^TXi)) which is equale to argmin(W)Summation(i=1 to n)log(1+exp(-Yi W^TXi))

Overfitting and Underfitting

graph for function exp(-x)



We can see that for every point of x, we are getting the value of exp(-x)>0. But if x tends to infinity then the result will go to minima point

Similarly, we are getting the minimum value eg. 0 for argmin(W)Summation(i=1 to n)log(1+exp(-YiW^TXi)) when zi tends to infinity for

all of i value only then the minimum value occurs.

Xi and Yi is fix value but W is varible if we pick w such that all training points are correctly classified and zi tends to infinity. But any points are outlier then occurs the overfitting problem. it means we are doing the great job for training data but no gaurentee for test data.

To get Zi tends to be infinity, wi should be +infinity or -infinity

To control this problem , regularization comes into then picture then the formula will be argmin(W)Summation(i=1 to n)(log(1+exp(-Yi*W^TXi)) + lambda*W^TW). then L2 regularization don't allow to loss-term(log(1+exp(-Yi *W^TXi)) to be an 0. lambda is hyper-parameter which can get using cross-validation or k-fold cross validation.

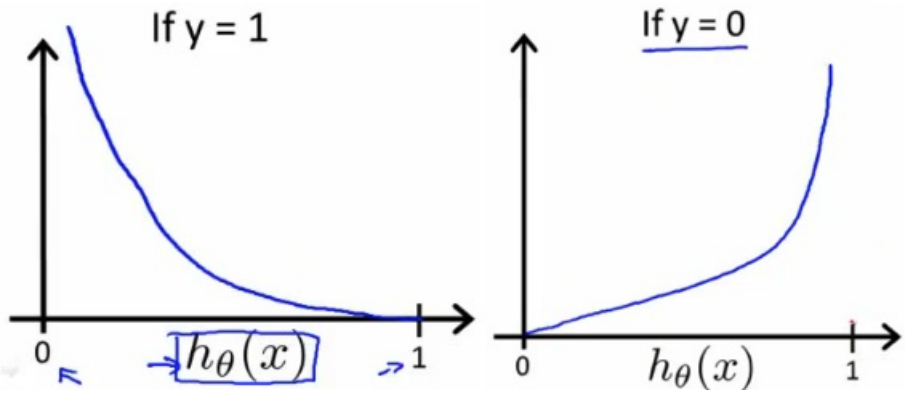
Some time , if lambda will be very large then to train the model, only regularization term comes into that picture. in that case underfit can be happen.

Hyperparmater should be correct

L1 regualrization is other alternative for L2 regularization. The formula of L1 is abs(W)(i =0 to n). It create sparsity . It will make Zero for those fearures which are unimportant

Loss Interpretation

Using logistic loss, we can get logistic regression. Loss Minimization Interpretation is working very nicely to get machine learning algorithm. Logistic loss is one of the approximation of 0-1 loss. The 0-1 loss can not be differentaite.



if y tends to 1 then loss function is decreasing . On the other hand , if y=0 loss function increasing

Hyper-parameter

lambda is hyperparmeter in logistic regression. if lambda = 0 then overfitting will happen. if lambda= infinity then underfitting will happen. lambda belongs to real value There are two technique to get optimal hyperparameter.

1. Grid-Search CV
2. Random Search CV

1. Grid Serach CV

Take the lambda value which belonga to real number like = [0.0001, 0.001, 0.01,0.1,1,10,100,1000] and plot the graph between lambda(which are x axis value) and Cross_Validation error. The value, for which we get low corss validation error, that value will be optimal value.

There is one drawback of Grid Search CV that it is more complex

2. Random Seach CV

We will provide the interval like = [0.000001,100000] and this technique pick the random value and try to minimize the cross validation error. That value will be optimal and that will be best hyperparameter

Column/Feature Standardization

Column standardization is very useful technique where we need to calculate the distance. In logistic regression, we are calculating the signed distance. Feature can be different scale. To make same scale, means centering and standardization is usefule.

We are calculating the mean and standardization for each feature and subtracting the each value in column like (xij)'= (xij-meanj)/sigmaj.

Feature Importance

For each features, we have corresponding weights. Feature importance works in logistic regression using |wj|. If corresponding weight is large that feature will be important . Features are independent then only |wj| as feature values.If Features are collinear in that case also |wj| does not work. We will use forward feature technique.

Feature Engineering/Feature Transformation

FE and FT tenchniqu works gor Non-Linearable data set.

In [ ]: