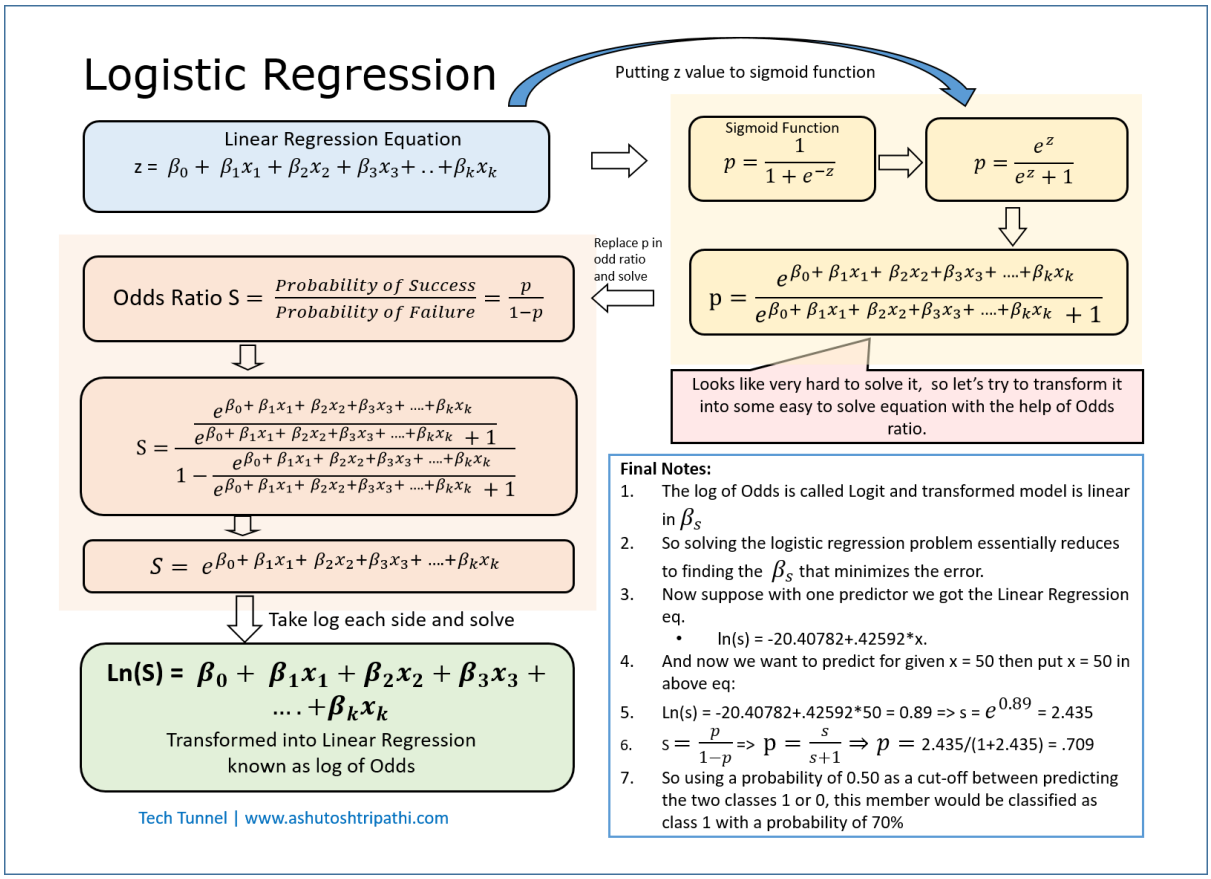
Mathematics behind Logistic Regression Model



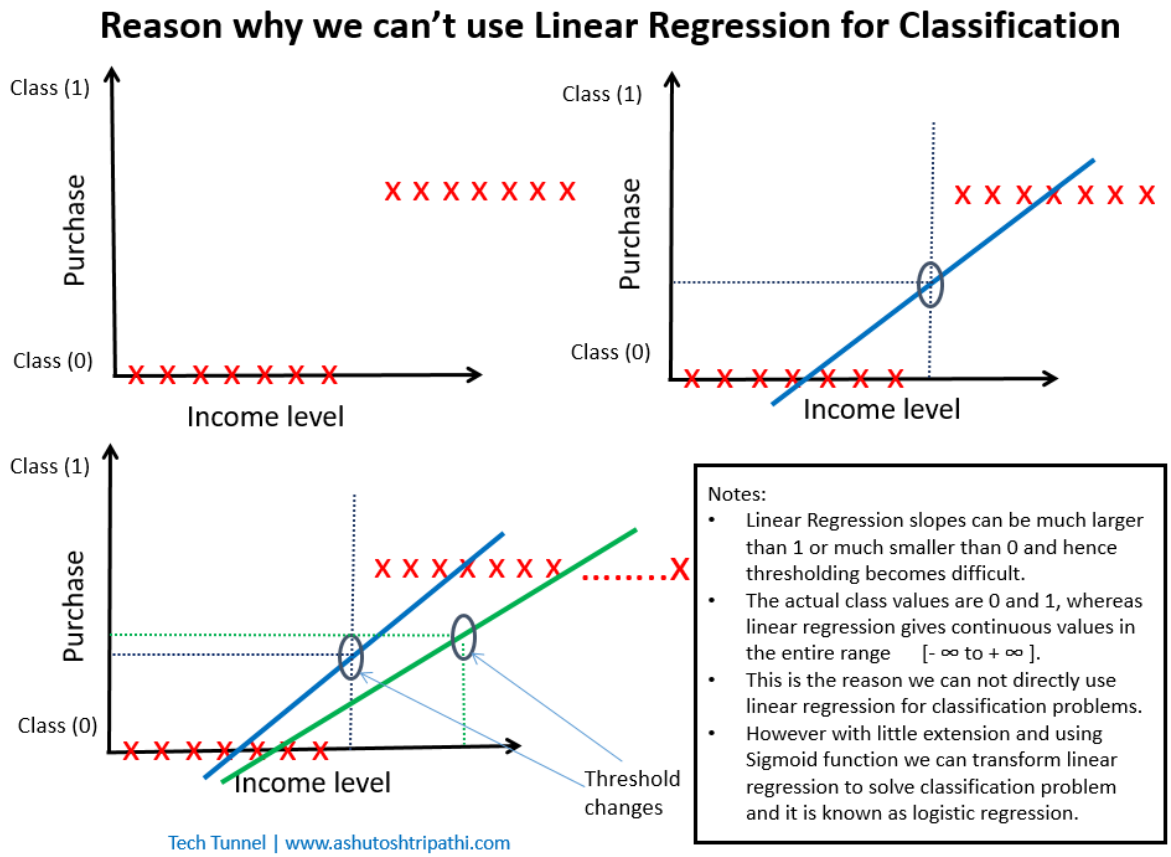
Logistic regression is the most widely used machine learning algorithm for classification problems. In its original form it is used for binary classification problem which has only two classes to predict. However with little extension and some human brain, logistic regression can easily be used for multi class classification problem. In this post I will be explaining about binary classification. I will also explain about the reason behind maximizing log likelihood function.

To understand logistic regression, it is required to have good understanding of linear regression [https://ashutoshtripathi.com/2019/01/06/what-is-linear-regression-part2/] concepts and cost function that is nothing but the minimization of sum of squared errors. I have explained this in detail in my earlier articles and I would recommend you to refresh linear regression before going deep into logistic regression. Assuming you have good understanding of linear regression let's start deep diving to logistic regression. However there arises one more question why can't we use linear regression for classification problems. Let's understand this first as this will be very good foundation for understanding the logistic regression.

Why can’t we use linear regression for classification problems?

Linear regression produces continuous values between [- ∞ to + ∞] as output for the prediction problem. So if we have a threshold defined so that we can say that above the threshold it belongs to one class and below the threshold it is another class and in this way we can intuitively say that we can use linear regression to solve a classification problem. However story does not ends here. Question arises, how to set the threshold and what about adding new records won’t change the threshold? Threshold we calculate by looking into the best fit line and by adding new record sets, best fit line changes which further changes the threshold value also. Hence we cannot confidently say that a particular record belongs to which class as we don’t have a certain threshold defined. And this is the main reason we cannot directly use linear regression for classification problems.

Below image describes this whole concept with an example.

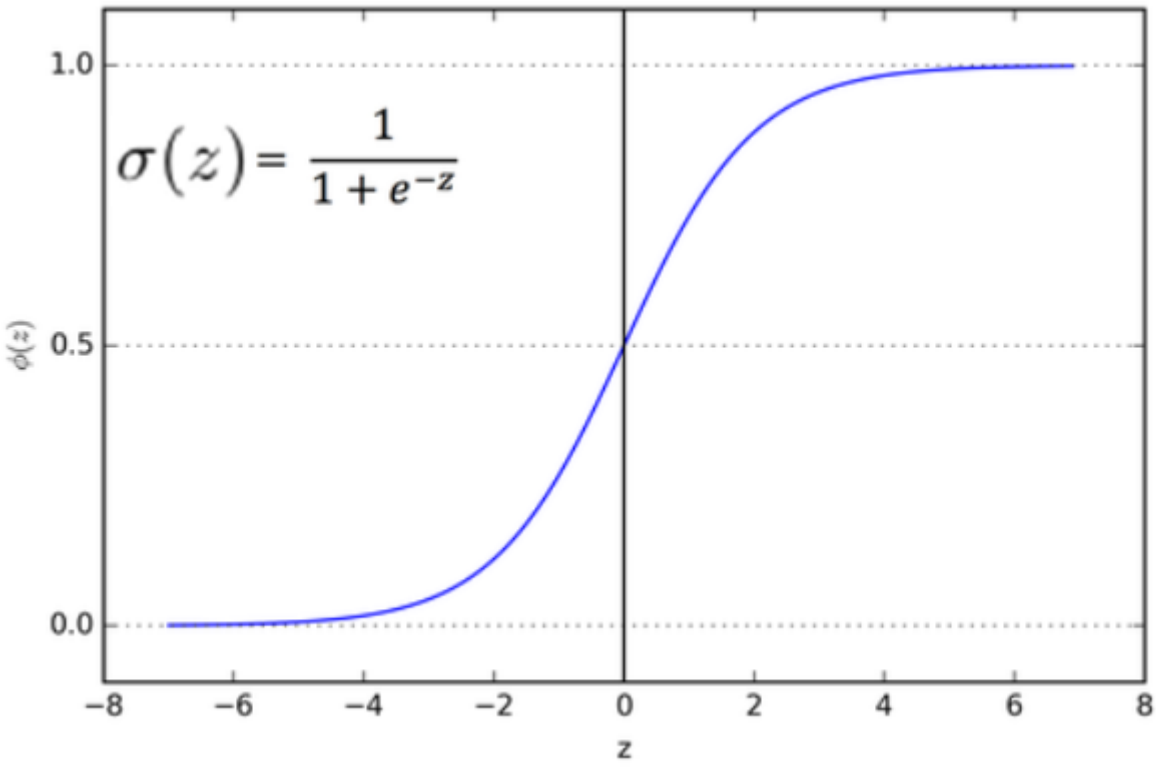


If we extend the concept of linear regression and limit the range of continuous values output [- ∞ to + ∞] to [0 to 1] and have function which calculates the probability [0 to 1] of belonging to a particular class then our job will be done. And fortunately **Sigmoid** or **Logistic** function do the job for us. Hence we also say that logistic regression is the transformation of linear regression using sigmoid function.

Sigmoid Function

A **Sigmoid function** is a mathematical function having a characteristic "S"-shaped curve or **sigmoid curve**. Often, *sigmoid function* refers to the special case of the [logistic function](#) shown in the first figure and defined by the formula (source: Wikipedia):

$$\sigma(z) = p = \frac{1}{1+e^{-z}} = \frac{e^{z}}{e^{z}+1}.$$



So Sigmoid function gives us the probability of being into the class 1 or class 0. So generally we take the threshold as .5 and say that if p >.5 then it belongs to class 1 and if p<.5 then it belongs to class 0. However this is not the fixed threshold. This vary based on the business problem. And what threshold value

should be, we can decide it with the help of AIC and ROC curves. Which I will be explaining later, in this post I will target mostly on how logistic regression works.

How Logistic Regression works:

As I have already written above that logistic regression uses Sigmoid function to transform linear regression into the logit function. Logit is nothing but log of Odds. And then using log of Odds it calculate the required probability. So let's understand first what the log of Odds is.

Log of Odds:

Odds ratio is obtained by the probability of an event occurring divided by the probability that it will not occur. And taking the log of Odds ratio will give the log of Odds. So what is the significance log of Odds here?

Logistic function or sigmoid function can be transformed into an Odds ratio:

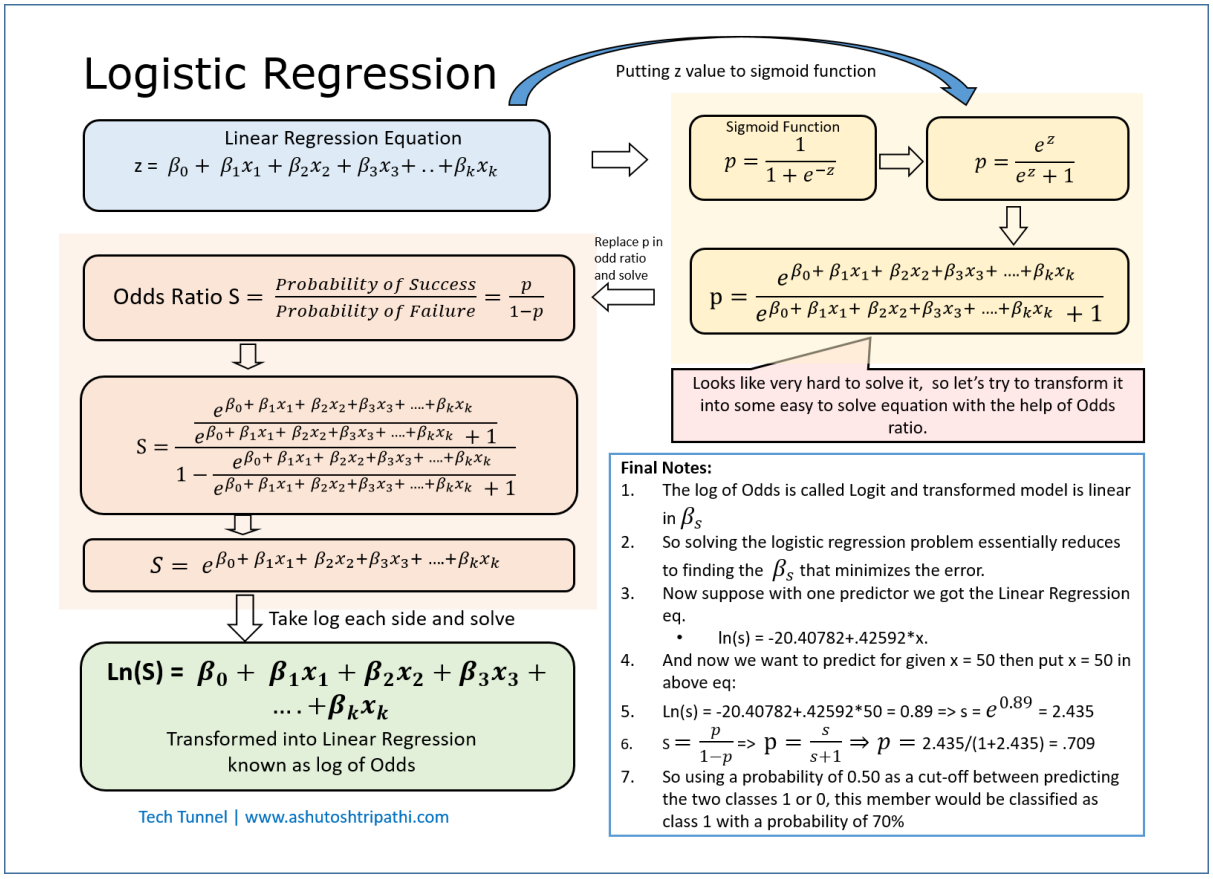
$$\text{Odds Ratio S} = \frac{p}{1-p}$$

Let's do some examples to understand probability and odds:

- Odds s = p/q, p is probability of winning, q is probability of losing that is 1-p. then if s is given then probability of winning p = numerator/ (numerator + denominator) and probability of losing q = denominator/ (numerator + denominator). Now let's solve some examples.
- If the probability of winning is 5/10 then what are the odds of winning? p = 5/10, => q = 1-p => q = 5/10 hence s = p/q => s = 1:1
- If the odds of winning are 13:2, what is the probability of winning? Probability of winning p = numerator/ (numerator + denominator) => p = 13/ (13+2) = 13/15.
- If the odds of winning are 3:8, what is the probability of losing? probability of losing q = denominator/(numerator + denominator) => q = 8/(3+8) => q = 8/11
- If the probability of losing q is 6/8 then what are the odds of winning? s= p/q, (1-q)/q => s = 2/6 or 1/3.
-

Logistic Model

In the below info graphics I have explained complete working of logistic model.



One more thing to note here is that logistic regression uses maximum likelihood estimation (MLE) instead of least squares method of minimizing the error which is used in linear models.

Least Squares vs Maximum likelihood Estimation

In Linear regression we minimized SSE.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

In Logistic Regression we maximize log likelihood instead. The main reason behind this is that SSE is not a convex function hence finding single minima won't be easy, there could be more than one minima. However Log likelihood is a convex function and hence finding optimal parameters is easier. Optimal could be max or min and here in case of log likelihood it is max.

$$\text{Log likelihood} = \sum[y_i ln\hat{y}_i + (1 - y_i)\ln(1 - \hat{y}_i)]$$

Now let's understand how log likelihood function behaves for two classes 1 and 0 of target variable.

Case 1: when Actual target class is 1 then we would like to have predicted target y hat value as close to 1. Let's understand how log likelihood function achieve this.

Putting y_i =1 will make the second part (after the +) of the equation 0 and only remaining is ln(y_i hat). And y_i hat will be between 0 to 1. ln(1) is 0 and ln(less than 1) will be less than 0 means negative. Hence max value of Log likelihood will be 0 and this will be only when y_i hat will be as close to 1. So maximizing the log likelihood is equivalent to getting y_i hat as close to 1 which means it will clearly identify predicted target as 1 which is same as actual target.

Case 2: when actual target class is 0 then we would like to have predicted target y hat as close to 0. Let's again understand this how maximizing log likelihood in this case will produce y_i hat closer to 0.

Putting y_i = 0 will make the first part (before + sign) of the equation 0 and only (1-y_i)ln(1-y_i hat) will remain. 1-y_i will again be 1 as y_i is 0 hence after reducing further equation will remain ln(1 – y_i hat). So now again 1 – y_i hat will be less than 1 as y_i hat will be between 0 to 1. So maximum value of ln(1 – y_i hat) can be 0. Means 1 – y_i hat should be close to 1 which implies y_i hat should be close to 0. that is as expected as actual value y_i is also 0.This is the reason we maximize the log likelihood.

Thank You

For more Articles please visit -> https://ashutoshtripathi.com