Logistic Regression

Prerequisite-

-Descriptive statistics

-Linear Regression

Objectives-

- To understand the prerequisite terms such as odds, probability, log odds, logit function and sigmoid function.
- What is logistic regression and the corresponding cost function?
- Learn how to optimize weights using gradient descent.

Odd and Probability-

The theory of chances for success or failure of the events are often expressed as odds and probabilities. Probability can be defined as to how likely an event can occur. Odds is ratio of the probability of success to failure of an event.

Probability- Probability is defined as the ratio of the number of way that are favourable to the occurrence of an event A to the total number of outcomes of the experiment. Probabilities are expressed either as percentage or decimal and value lies between 0 and 1.

$$Proability\ P(A) = \frac{chances\ in\ favor(A)}{total\ trials}$$

Example: The probability of getting Ace in a deck of 52 cards is given by:

$$P(A) = \frac{4}{52} = 0.077\ or\ 7.7\%$$

Odd- Odds is defined as the ratio to chances in the favor of event to the chances against it. The value of odd may lie between 0 to ∞.

$$oddratio(A) = \frac{chances\ in\ favor(A)}{chances\ against(A)}$$

Mathematically Odds Ratio = $\frac{P}{1-P}$, where P denotes the probability of success of the desired event.

Example: The odds of getting Ace in a deck of 52 cards is given by:

$$odds\ ratio(A) = \frac{(4/52)}{(48/52\ )} = \frac{4}{48}$$

Relationship Between the Probability and the Odds Ratio-

$$oddratio = \frac{Probability\ of\ event(A)\ occuring}{Probability\ of\ event(A)\ not\ occuring}$$

$$oddratio\ (A) = \frac{P(A)}{1 - P(A)}$$

$$Probability\ P(A) = \frac{odd\ ratio}{1 + odd\ ratio}$$

Log odds and logit-

We now know that the odds ratio is the probability of an event occurring to the probability of that event not occurring. Taking log of odd ratio is called log odds and defined as:

$$\log(A) = \log(\frac{P(A)}{1 - P(A)})$$

The logit function is the log odds function modelled for a probability 'p'.

$$\log(odds) = logit(P) = \log(\frac{P}{1 - P})$$

Logit Function- Logit function is mainly used in working with probabilities. The logit function is the log of the odds that Y equals one of the categories. The value of logit function varies between $(-\infty\ , \infty)$. The value approaches towards ∞ when the probability value approaches to 1 and it goes to $-\infty$ when the probability value approaches to 0. The logit function is very important in the field of statistics as it can map the probability values ranges (0, 1) to a full range value of real numbers.

$$logit\big(y(z)\big) = \log(\frac{z}{1 - z})$$

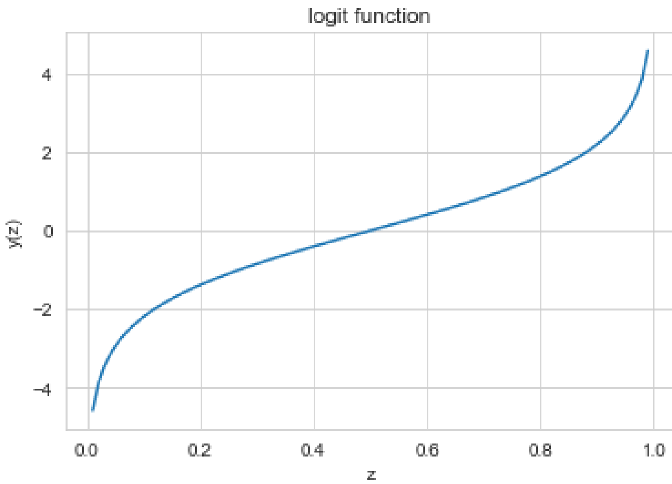


Figure 1: Logit Function

Sigmoid function- Sigmoid function can be thought of as an inverse to the logit function. This means for a probability value P we have:

$$P = sigmoid\big(logit(P)\big)$$

Sigmoid (sometimes also called the S-curve) performs the inverse of logit which means it maps any arbitrary real number into the range of (0, 1). The function is defined as:

$$\boldsymbol{sigmoid(z) = \frac{1}{1 + e^{-z}}}$$

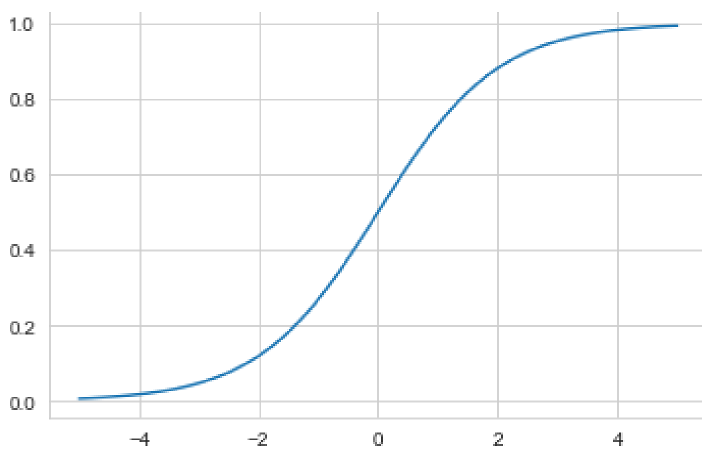


Figure 2 Sigmoid function

Logistic Regression-

In the case of linear regression, the target variable 'y' is a continuous variable but let us assume that the 'y' is a categorical variable which has two classes then linear regression should not be used to predict the value of target variable. The output of the Linear Regression is not bound within [0,1] as it can take any real value from $(-\infty, \infty)$. Logistic regression is used to solve such problem which gives us the corresponding probability outputs and then we can decide the appropriate cut-off points to get the target class outputs.

Precisely Logistic Regression is defined as a statistical approach, for calculating the probability outputs for the target labels. In its basic form it is used to classify binary data. Logistic regression is very much similar to linear regression where the explanatory variables(X) are combined with weights to predict a target variable of binary class(y). The main difference between linear regression and logistic regression is of the type of the target variable. Logistic regression model can be expressed as:

$$y = \frac{1}{1 + e^{-(w_0 + w_1 x)}} = \frac{e^{(w_0 + w_1 x)}}{e^{(w_0 + w_1 x)} + 1}$$

Examples-

1. **Churn prediction-** Churn is the probability of a client to abandon a service or stop paying to that particular service provider. The ratio of clients that abandon the service during a particular time interval is called churn rate. Churn prediction is considered as a problem of binary classification that whether a client or customer is going to churn. For example a particular client churn the on the basis of monthly charge of service.

$$P(churn = 1|monthly\ charge) = \frac{1}{1 + e^{-(w_0 + w_1 monthly\ charge)}}$$

$$churn = \begin{cases} 1 \ \ if \ \ P(churn = 1|monthly\ charge) > \ 0.5 \\ 0 \ \ if \ \ P(churn = 1|monthly\ charge) \leq 0.5 \end{cases}$$

Here, we have selected the probability cut-off to 0.5 for predicting the target classes.

2. **Spam Detection-** Problem to identify that an email is spam or not.
3. **Banking-** Problem to predict a particular customer default a loan or not.

Cost Function-

Linear regression uses mean squared error as its cost function but unfortunately this cannot be used with logistic regression. Logistic regression uses Cross-Entropy or Log-Loss function as its cost function defined for two class classification problem.

$$cost(w_0, w_1) = -\frac{1}{m}\sum_{i=1}^{m}\{y_i log(a_i) + (1 - y_i) log(1 - a_i)\}$$

Where

$$a_i = sigm(yhat_i)$$

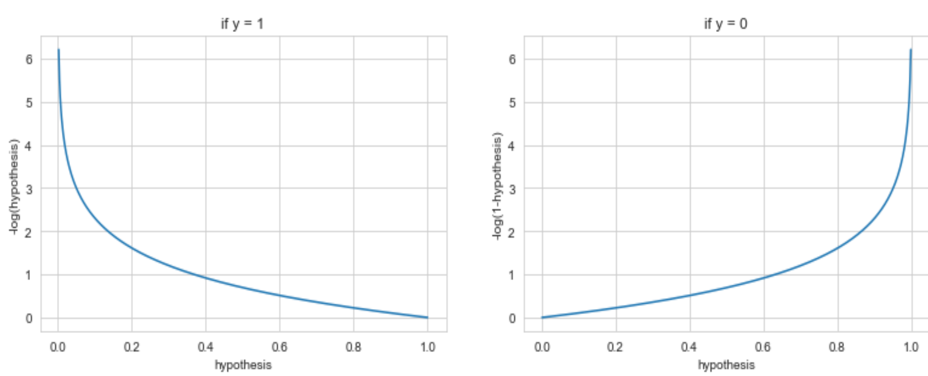
$$sigm(yhat_i) = \frac{1}{1 + e^{-yhat_i}}$$

The cost function can be divided into two separate functions as:

$$cost(w_0, w_1) = -\frac{1}{m}\sum_{i=1}^{m} log(a_i) \;\; \boldsymbol{if} \; y = 1$$

and

$$cost(w_0, w_1) = -\frac{1}{m}\sum_{i=1}^{m} log(1 - a_i) \;\; \boldsymbol{if} \; y = 0$$



Optimization of coefficient or weight parameter- Again gradient descent is used to optimize the value of weight parameters. The derivative of Loss function is defined as:

$$Cost = \frac{1}{m}\sum_{i=1}^{m}\{-y_i \log a_i - (1-y_i)\log(1-a_i)\}$$

where

$$a_i = \frac{1}{1+ e^{-yhat_i}}$$

and

$$yhat_i = w_0 + w_1 x_i$$

Now let find the derivative of cost function (By chain rule of partial derivative):

$$\frac{\partial Cost}{\partial w_i} = \frac{\partial Cost}{\partial a_i} * \frac{\partial a_i}{\partial yhat_i} * \frac{\partial yhat_i}{\partial w_i}$$

So,

$$\frac{\partial Cost}{\partial a_i} = \frac{a_i - y_i}{a_i(1-y_i)}$$

$$\frac{\partial a_i}{\partial yhat_i} = a_i(1-a_i)$$

and

$$\frac{\partial yhat_i}{\partial w_i} = x_i$$

By using all above the generalize formula expressed as:

$$\frac{\partial Cost}{\partial w_i} = \frac{1}{m}\sum_{i=1}^{m}(a_i - y_i)x_i \;\; with\; x_0 = 1\; for\; w_0$$

Parameter update:

$$\boldsymbol{w_i = w_i - lrate}\frac{\partial Cost}{\partial w_i}$$

Example- Consider an example where we are interested to find the effect of studying hours per day over the result in examination and predict that a student will pass or fail for given study hours. We have sample data about six students for their grades and total study hours per day.

Table-1	
Study Hours per day	**Grades**
6	1(Pass)
5	0(Fail)
4	0(Fail)
7	1(Pass)
8	1(Pass)
2	0(Fail)

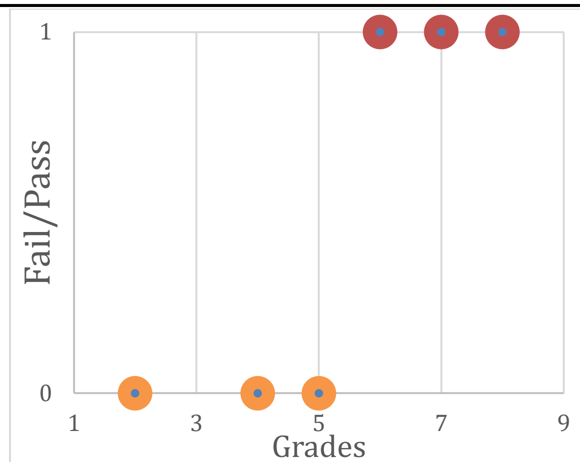


Figure-1: Scatter Graph

To solve the problem using logistic regression let us model the linear equation as:

$$y(Grades) = w_0 + w_1\, x(Study\ Hours\ per\ day)$$

and predict the result using:

$$P(result = 1|studyhours) = \frac{1}{1 + e^{-(w_0 + w_1\, x(Study\ Hours\ per\ day))}}$$

$$yhat_i = w_0 + w_1 x_i$$

and

$$a_i = sigm(yhat_i) = \frac{1}{1 + e^{-yhat_i}}$$

$$yhat_i = w_0 + w_1 x_i$$

Cost Function:

$$Cost(w_0, w_1) = \frac{1}{m} \sum_{i=1}^{m} \{-y_i \log a_i - (1 - y_i)\log(1 - a_i)\ \}$$

Gradients:

$$\frac{\partial Cost(w_0, w_1)}{\partial w_0} = \frac{1}{m} \sum_{i=1}^{m} (a_i - y_i)$$

and

$$\frac{\partial Cost(w_0, w_1)}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m}(a_i - y_i)x_i$$

Parameter updates:

$$w_0 = w_0 - lrate\frac{\partial Cost(w_0, w_1)}{\partial w_0}$$

and

$$w_1 = w_1 - lrate\frac{\partial Cost(w_0, w_1)}{\partial w_1}$$

We have,

X:	6	5	4	7	8	2
y:	1	0	0	1	1	0

Iteration #1:

Let $w_0 = 1$ and $w_1 = 1, with\ lrate = 0.01$

$yhat_i = w_0 + w_1 x_i \quad and \;\; a_i = sigm(yhat_i) = \frac{1}{1 + e^{-yhat_i}}$

yhat:	7	6	5	8	9	3
a:	0.999	0.997	0.993	0.999	0.999	0.995

So,

$$\frac{\partial Cost(w_0, w_1)}{\partial w_0} = \frac{\begin{array}{c}(0.999 - 1) + (0.997 - 0) + (0.993 - 0) + \\ (0.999 - 1) + (0.999 - 1) + (0.995 - 0)\end{array}}{6}$$

$$\frac{\partial Cost(w_0, w_1)}{\partial w_0} = 0.497$$

and

$$\frac{\partial Cost(w_0, w_1)}{\partial w_1} = \frac{\begin{array}{c}(0.999 - 1) * 6 + (0.997 - 0) * 5 + (0.993 - 0) * 4 + \\ (0.999 - 1) * 7 + (0.999 - 1) * 8 + (0.995 - 0) * 2\end{array}}{6}$$

$$\frac{\partial Cost(w_0, w_1)}{\partial w_1} = 1.821$$

Parameter update:

$$w_0 = w_0 - lrate * \frac{\partial Cost(w_0, w_1)}{\partial w_0} = 1 - 0.01 * (0.497) = 0.995$$

$$w_1 = w_1 - lrate * \frac{\partial Cost(w_0, w_1)}{\partial w_1} = 1 - 0.01 * (1.821) = 0.982$$

Iteration #1:

Let $w_0 = 0.995 \; and \; w_1 = 0.982, \; with \; lrate = 0.01$

yhat:	6.887	5.905	4.923	7.869	8.851	2.959
***a*:**	0.999	0.997	0.993	0.999	0.999	0.950

$$\frac{\partial Cost(w_0, w_1)}{\partial w_0} = \frac{\begin{array}{c}(0.999-1)+(0.997-0)+(0.993-0)+\\(0.999-1)+(0.999-1)+(0.950-0)\end{array}}{6} = 0.489$$

and

$$\frac{\partial Cost(w_0, w_1)}{\partial w_1} = \frac{\begin{array}{c}(0.999-1)*6+(0.997-0)*5+(0.993-0)*4+\\(0.999-1)*7+(0.999-1)*8+(0.950-0)*2\end{array}}{6} = 1.806$$

Parameter update:

$$w_0 = w_0 - lrate * \frac{\partial Cost(w_0, w_1)}{\partial w_0} = 0.995 - 0.01 * (0.489) = 0.990$$

$$w_1 = w_1 - lrate * \frac{\partial Cost(w_0, w_1)}{\partial w_1} = 0.982 - 0.01 * (1.806) = 0.964$$

and so on……….

Evaluation of Logistic regression model- Performance measurement of classification algorithms are judge by confusion matrix which comprise the classification count values of actual and predicted labels. The confusion matrix for binary classification is given by:

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3 Confusion Matrix

Confusion matrix cells are populated by the terms:

True Positive(TP)- The values which are predicted as True and are actually True.

True Negative(TN)- The values which are predicted as False and are actually False.

False Positive(FP)- The values which are predicted as True but are actually False.

False Negative(FN)- The values which are predicted as False but are actually True.

Classification performance metrics are based on confusion matrix values. The most popularly used metrics are;

Precision- measure of correctness achieved in prediction.

$$precision = \frac{TP}{TP + FP}$$

Recall (sensitivity)- measure of completeness, actual observations which are predicted correctly.

$$recall = \frac{TP}{TP + FN}$$

Specificity- measure of how many observations of false category predicted correctly.

$$specificity = \frac{TN}{TN + FP}$$

F1-Score- a way to combine precision and recall metric in a single term. F1score is defined as harmonic mean of precision and recall.

$$F1score = \frac{2 * precision * recall}{precision + recall}$$

ROC Curve- Receiver Operating Characteristic(ROC) measures the performance of models by evaluating the trade-offs between sensitivity (true positive rate) and false (1-specificity) or false positive rate.

AUC - The area under curve (AUC) is another measure for classification models is based on ROC. It is the measure of accuracy judged by the area under the curve for ROC.

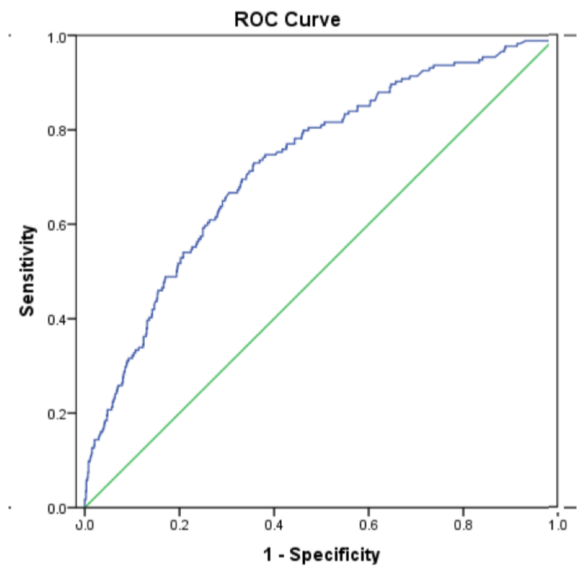


Figure 4 ROC Curve

Pros and cons of Linear Regression:

Pros- Logistic regression classification model is simple and easily scalable for multiple classes.

Cons- Classifier constructs linear boundaries and the interpretation of coefficients value is difficult.

****