Logistic Regression

Prerequisite:

- Descriptive statistics
- Linear Regression

Objectives:

- Understand the prerequisite terms odds, probability, log odds, logit function and sigmoid
- What is logistic regression and its cost function
- Different metrics to evaluate classification models
- Learn how to optimize weights using gradient descent

Odd and Probability:

The theory of chances for success or failure of events is often expressed as odds and probabilities. These two terms denote the same information but are different from one another in expression and defined as:

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 ∞.

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

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

$$odd(A) = \frac{4}{48}$$

Probability:

Probability is defined as the ratio to chances in the favor of events to the total trials. 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:

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

Relationship:

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

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

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

Odds Ratio:

Odd and odds ratio are often confused with each other but they are very different. The odds ratio is defined as the ratio of the odds of A in the presence of B and the odds of A in the absence of B, or equivalently (due to symmetry), the ratio of the odds of B in the presence of A and the odds of B in the absence of A. Two events are independent if and only if the OR equals 1, i.e., the odds of one event are the same in either the presence or absence of the other event. If the OR is greater than 1, then A and B are associated (correlated) in the sense that, compared to the absence of B, the presence of B raises the odds of A, and symmetrically the presence of A raises the odds of B. Conversely, if the OR is less than 1, then A and B are negatively correlated, and the presence of one event reduces the odds of the other event.

$$odds\ ratio = \frac{Odd(A)}{Odd(B)} = \frac{\frac{p_1}{1 - p_1}}{\frac{p_2}{1 - p_2}}$$

Where, p_1 is the probability of event A and p_2 is the probability of event B.

Odds ratio can have high magnitude even if the underlying probabilities are very low.

Log odds and logit:

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

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

When the function variable of log is probability p, it is called as **logit** function which means logit of probability is log of odds.

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

Logit Function: Logit function is mainly used while 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 probability value touches to 1 and it goes to $-\infty$ when probability value touches to 0. The logit function is very important in the field of statistics as it can map the probability values ranges from (0, 1) to a full range value of real numbers.

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

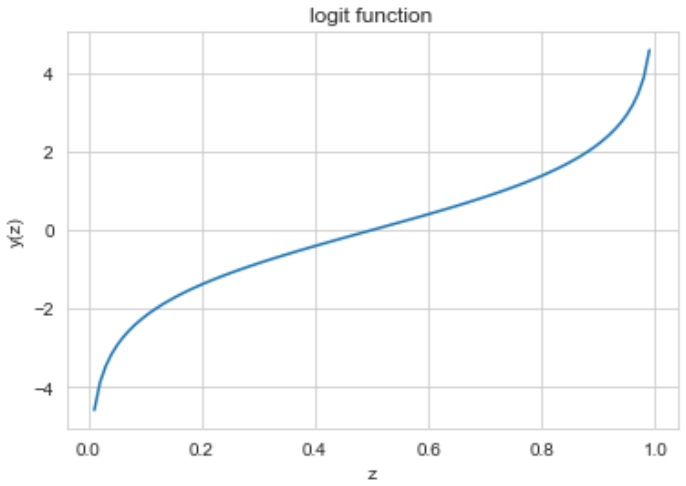


Figure 1: Logit Function

Sigmoid function:

Sigmoid function is defined as the inverse of logit function, which means for a probability value P we have:

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

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

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

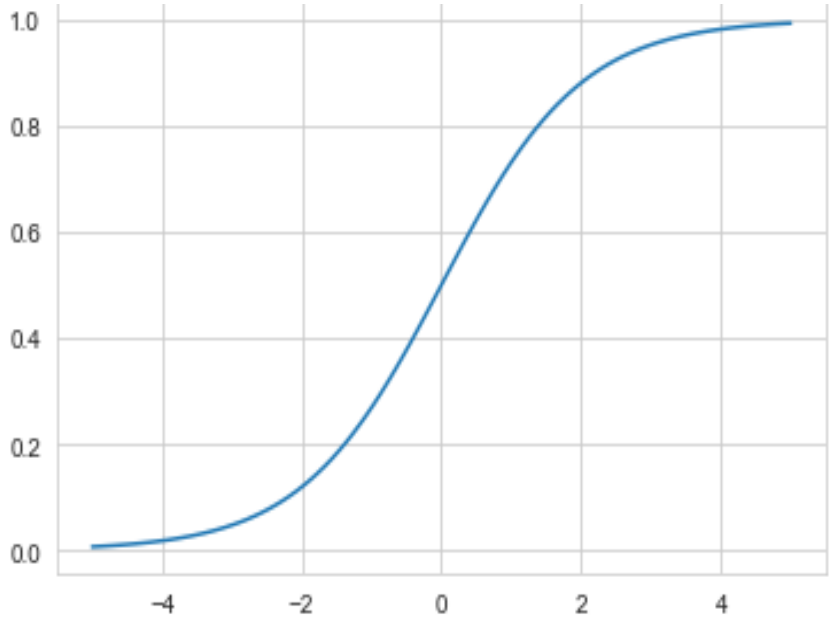


Figure 2 Sigmoid Function

Logistic Regression:

In case of linear regression, the target variable y is a continuous variable but let's suppose y is categorical variable which has two classes then linear regression cannot be used to predict the value of target variable. Logistic regression is used to solve such problem.

Precisely, logistic regression is defined as a statistical approach for classifying labels. In its basic form it is used to classify binary data. However, it is easily extended to multi-class problems. Logistic regression is very much similar to linear regression where the explanatory variables are combined with weights to predict a target variable of binary class. The main difference between linear regression and logistic regression is of target variable. The logistic regression models the target values as 0 or 1 whereas linear regression models them as numeric value. Logistic regression model is 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 client to abandon a service or stop paying as client of a 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 in future whether a client or customer churns the service based on his/her attributes. For example, a particular client churns on the basis of monthly charges of the service.

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

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

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

Cost Function:

Linear regression uses mean squared error as its cost function but unfortunately this cannot be used with classification problems. Logistic regression uses Cross-Entropy or Log-Loss function as it's 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

$$a_i = sigm(yhat_i)$$

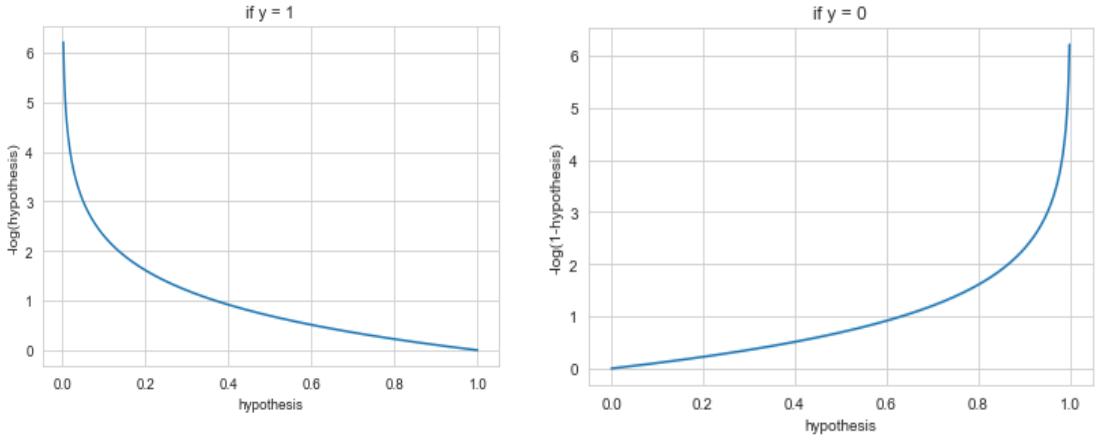
$$sigm(yhat_i) = \frac{1}{1 + e^{-yhat_i}}$$

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. Now let's find the derivative of the cost function defined above (using chain rule of partial derivatives):

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

Where,

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

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

and

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

By using all above the generalized formula is 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:

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

Where, $lrate$ is the learning rate.

Example:

Consider an example where we are interested in finding 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)

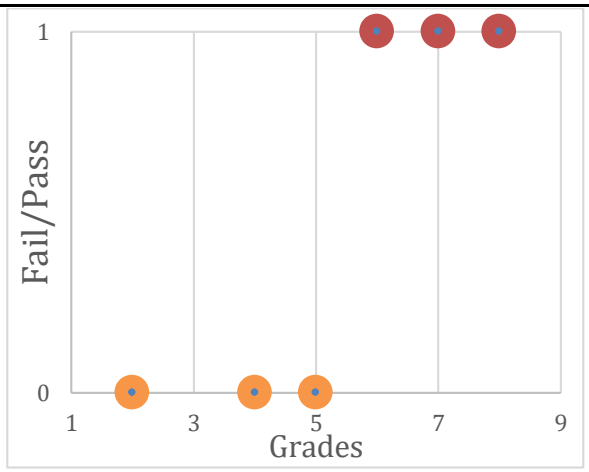


Figure-1: Scatter Graph

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:

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

Cost Function:

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

$$yhat_i = w_0 + w_1 x_i$$

and

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

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

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

Let's see this in practice. We have,

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

Iteration #1:

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

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

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

So,

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

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

and

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

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

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 #2:

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

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

and

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

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:

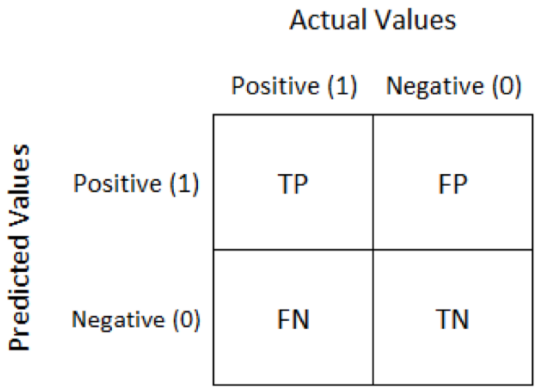


Figure 3 Confusion Matrix

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 true 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. F1-score 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 positive rate (1- specificity).

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

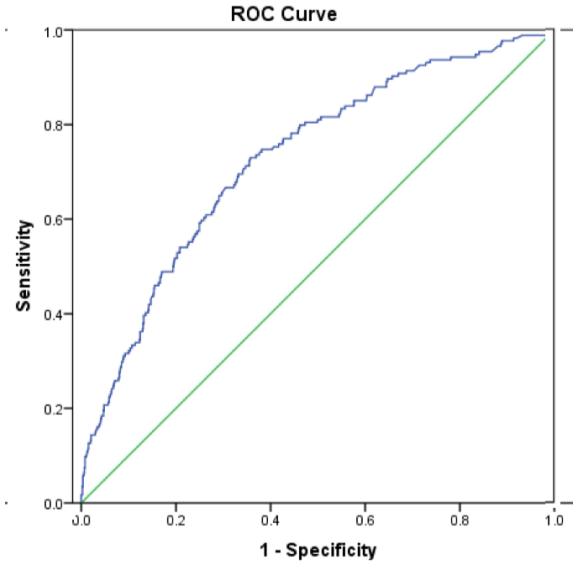


Figure 4 ROC Curve

Pros and cons of Linear Regression:

Pros:

- A classification model that does give probabilities
- Easily extended to multiple classes
- Quick to train and very fast at classifying unknown records

Cons:

- Logistic regression constructs linear boundaries
- The interpretation of coefficients is difficult
- Assumes that variables are independent

****