Course: Artificial Intelligence and Machine Learning Code: 20CS51I
WEEK- 8 Machine Learning

Logistic Regression in Machine Learning

Overview

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"

Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

How Does Logistic Regression Work?

Machine learning generally involves predicting a quantitative outcome or a qualitative class. The former is commonly referred to as a regression problem. In the scenario of linear regression, the input is a continuous variable, and the prediction is a numerical value. When predicting a qualitative outcome (class), the task is considered a classification problem.

Logistic regression is part of the regression family as it involves predicting outcomes based on quantitative relationships between variables. However, unlike linear regression, it accepts both continuous and discrete variables as input and its output is qualitative. In addition, it predicts a discrete class such as “Yes/No” or “Customer/Non-customer”.

The logistic regression algorithm analyzes relationships between variables. It assigns probabilities to discrete outcomes using the sigmoid function, which converts numerical results into an expression of probability between 0 and 1.0. Probability is either 0 or 1, depending on whether the event happens or not. For binary predictions, you can divide the population into two groups with a cut-off of 0.5. Everything above 0.5 is considered to belong to group A, and everything below is considered to belong to group B

A hyperplane is used as a decision line to separate two categories (as far as possible) after data points have been assigned to a class using the Sigmoid function. The class of future data points can then be predicted using the decision boundary.

Consider the following example: An organization wants to determine an employee’s salary increase based on their performance.

For this purpose, a linear regression algorithm will help them decide. Plotting a regression line by considering the employee’s performance as the independent variable, and the salary increase as the dependent variable will make their task easier.

Now, what if the organization wants to know whether an employee would get a promotion or not based on their performance? The above linear graph won't be suitable in this case. As such, we clip the line at zero and one, and convert it into a sigmoid curve (S curve).

Based on the threshold values, the organization can decide whether an employee will get a salary increase or not.

To understand logistic regression, let's go over the odds of success.

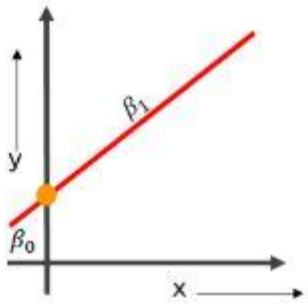
Odds (□) = Probability of an event happening / Probability of an event not happening

□ = p / 1 - p

The values of odds range from zero to ∞ and the values of probability lies between zero and one.

Consider the equation of a straight line:

□ = □0 + □1* □



Here, □0 is the y-intercept

□1 is the slope of the line

x is the value of the x coordinate

y is the value of the prediction

Now to predict the odds of success, we use the following formula:

$$\log\left(\frac{p(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x$$

Exponentiating both the sides, we have:

$$e^{ln}\left(\frac{p(x)}{1-p(x)}\right) = e^{\beta_0+\beta_1 x}$$

$$\left(\frac{p(x)}{1-p(x)}\right) = e^{\beta_0+\beta_1 x}$$

Let Y = e □0+□1 * □

Then p(x) / 1 - p(x) = Y

p(x) = Y(1 - p(x))

p(x) = Y - Y(p(x))

p(x) + Y(p(x)) = Y

p(x)(1+Y) = Y

p(x) = Y / 1+Y

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

The equation of the sigmoid function is:

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}$$

The sigmoid curve obtained from the above equation is as follows:

Conclusion

The logistic regression model is an analysis technique that helps predict the probability of an event happening in the future. Logistic regression is a supervised learning method that helps to predict events that have a binary outcome, such as whether a person will successfully pass a driving test. In order to make predictions in this scenario, you need data from past test results. The model takes this data and predicts the likelihood that the same person will pass the test in the future. The main idea behind logistic regression is to use a model based on the probability of an outcome occurring.

Assumptions for Logistic Regression:

Every statistical method has assumptions. Assumptions mean that your data must satisfy certain properties in order for statistical method results to be accurate.

The assumptions for Simple Logistic Regression include:

1. Linearity
2. No Outliers
3. Independence

Linearity

Logistic regression fits a logistic curve to binary data. This logistic curve can be interpreted as the probability associated with each outcome across independent variable values. Logistic regression assumes that the relationship between the natural log of these probabilities (when expressed as odds) and your predictor variable is linear.

No Outliers

The variables that you care about must not contain outliers. Logistic Regression is sensitive to outliers, or data points that have unusually large or small values. You can tell if your variables have outliers by plotting them and observing if any points are far from all other points.

Independence

Each of your observations (data points) should be independent. This means that each value of your variables doesn't "depend" on any of the others. For example, this assumption is usually violated when there are multiple data points over time from the same unit of observation (e.g. subject/participant/customer/store), because the data points from the same unit of observation are likely to be related or affect one another.

Understanding sigmoid function:

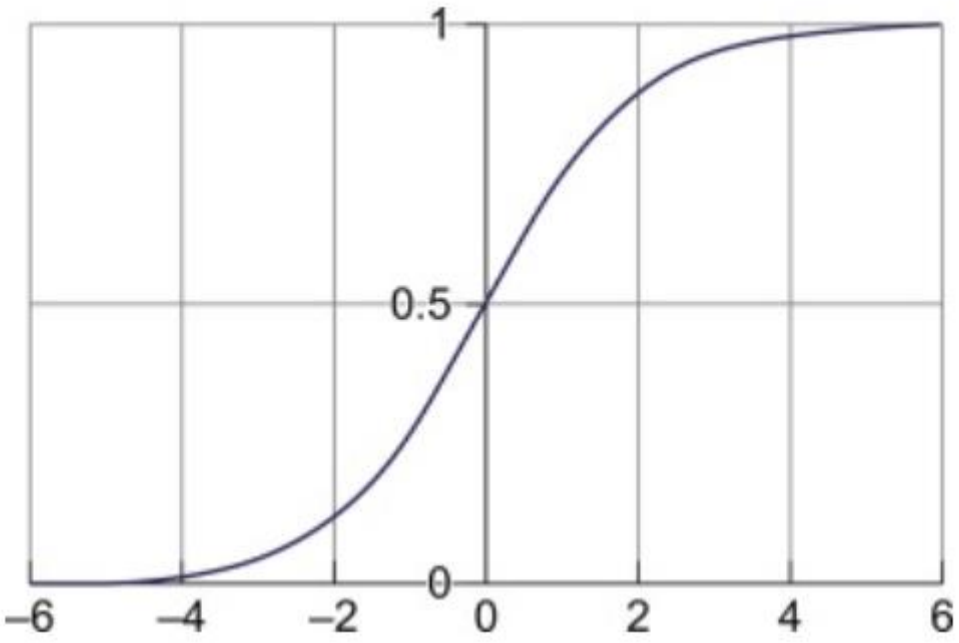
Logistic/Sigmoid Function

The sigmoid function, commonly known as the logistic function, predicts the likelihood of a binary outcome occurring. The function takes any value and converts it to a number between 0 and 1. The Sigmoid Function is a machine learning activation function that is used to introduce non-linearity to a machine learning model.

The formula of Logistic Function is:

$$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

When we plot the above equation, we get S shape curve like below.

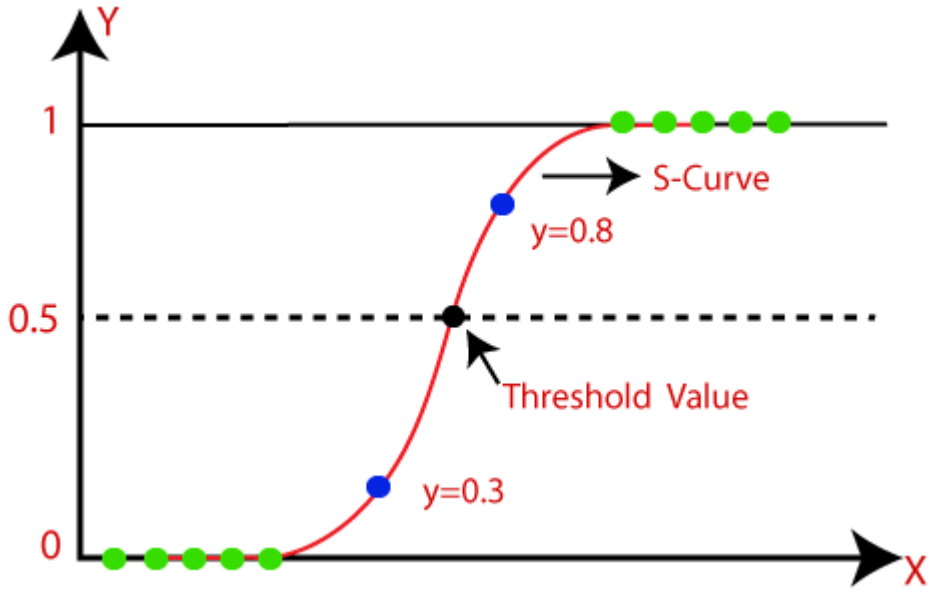


The key point from the above graph is that no matter what value of x we use in the logistic or sigmoid function, the output along the vertical axis will always be between 0 and 1.

When the result of the sigmoid function is greater than 0.5, we classify the label as class 1 or positive class; if it's less than 0.5, we can classify it as a negative class or 0.

In Logistic Regression, iterative optimization algorithms like Gradient Descent or probabilistic methods like Maximum Likelihood are used to get the "best fit" S curve

- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Note: Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Applications of Logistic Regression:

- Using the logistic regression algorithm, banks can predict whether a customer would default on loans or not
- To predict the weather conditions of a certain place (sunny, windy, rainy, humid, etc.)
- Ecommerce companies can identify buyers if they are likely to purchase a certain product
- Companies can predict whether they will gain or lose money in the next quarter, year, or month based on their current performance
- To classify objects based on their features and attributes

Build logistic regression model in python

In Python we have modules that will do the work for us. Start by importing the NumPy module.

```
import numpy
```

Store the independent variables in X.

Store the dependent variable in y.

Below is a sample dataset:

```
#X represents the size of a tumor in centimeters.

X = numpy.array([3.78, 2.44, 2.09, 0.14, 1.72, 1.65, 4.92, 4.37, 4.96, 4.52, 3.69, 5.88]).reshape(-1,1)

#Note: X has to be reshaped into a column from a row for the LogisticRegression() function to work.

#y represents whether or not the tumor is cancerous (0 for "No", 1 for "Yes").

y = numpy.array([0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1])
```

We will use a method from the sklearn module, so we will have to import that module as well:

```
from sklearn import linear_model
```

From the sklearn module we will use the LogisticRegression() method to create a logistic regression object.

This object has a method called fit() that takes the independent and dependent values as parameters and fills the regression object with data that describes the relationship:

```
logr = linear_model.LogisticRegression()

logr.fit(X,y)
```

Now we have a logistic regression object that is ready to whether a tumor is cancerous based on the tumor size:

```
#predict if tumor is cancerous where the size is 3.46mm:

predicted = logr.predict(numpy.array([3.46]).reshape(-1,1))
```

Example

See the whole example in action:

```
import numpy

from sklearn import linear_model

#Reshaped for Logistic function.

X = numpy.array([3.78, 2.44, 2.09, 0.14, 1.72, 1.65, 4.92, 4.37, 4.96, 4.52, 3.69, 5.88]).reshape(-1,1)

y = numpy.array([0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1])

logr = linear_model.LogisticRegression()

logr.fit(X,y)

#predict if tumor is cancerous where the size is 3.46mm:

predicted = logr.predict(numpy.array([3.46]).reshape(-1,1))

print(predicted)
```

Result

```
[0]
```

We have predicted that a tumor with a size of 3.46mm will not be cancerous.

Coefficient

In logistic regression the coefficient is the expected change in log-odds of having the outcome per unit change in X.

This does not have the most intuitive understanding so let's use it to create something that makes more sense, odds.

Example

See the whole example in action:

```
import numpy

from sklearn import linear_model

#Reshaped for Logistic function.

X = numpy.array([3.78, 2.44, 2.09, 0.14, 1.72, 1.65, 4.92, 4.37, 4.96, 4.52, 3.69, 5.88]).reshape(-1,1)

y = numpy.array([0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1])

logr = linear_model.LogisticRegression()

logr.fit(X,y)

log_odds = logr.coef_

odds = numpy.exp(log_odds)

print(odds)
```

Result

```
[4.03541657]
```

Run example »

MCQS Logistic Regression

Question 1: Logistic regression is used for ___?

re(B) regression

(C) clustering

(D) All of these

Question 2: Logistic Regression is a Machine Learning algorithm that is used to predict the probability of a ___?

(A) categorical independent variable

(B) categorical dependent variable.

(C) numerical dependent variable.

(D) numerical independent variable.

Question 3: You are predicting whether an email is spam or not. Based on the features, you obtained an estimated probability to be 0.75. What's the meaning of this estimated probability? (select two)

(A) there is 25% chance that the email will be spam

(B) there is 75% chance that the email will be spam

(C) there is 75% chance that the email will not be spam

(D) there is 25% chance that the email will not be spam

Question 4: In a logistic regression model, the decision boundary can be ___.

(A) linear

(B) non-linear

(C) both (A) and (B)

(D) none of these

Question 5: What's the cost function of the logistic regression?

(A) Sigmoid function

(B) Logistic Function

(C) both (A) and (B)

(D) none of these

Question 6: You are predicting whether an email is spam or not. Based on the features, you obtained an estimated probability to be 0.75. What's the meaning of this estimated probability? The threshold to differ the classes is 0.5.

(A) The email is not spam

(B) The email is spam

(C) Can't determine

(D) both (A) and (B)

Question 7: What's the the hypothesis of logistic regression?

(A) to limit the cost function between 0 and 1

(B) to limit the cost function between -1 and 1

(C) to limit the cost function between -infinity and +infinity

(D) to limit the cost function between 0 and +infinity

Question 8: In a logistic regression, if the predicted logit is 0, what's the transformed probability?

(A) 0

(B) 1

(C) 0.5

(D) 0.05

Solution are 1(A), 2(C), 3(A), 4(A), 5(C), 6(A,C), 7(A), 8(C,D)