Logistic Regression

Logistic regression is a technique used to solve classification problems. Classification types have binary classification and multi-class classification. linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

It seeks to model the probability of an event occurring depending on the values of the *independent variables,* which can be categorical or numerical. It estimates the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur at all. It mainly predicts the effect of a series of variables on a binary response variable. It can also classify observations by estimating the probability that an observation is in a particular category (such as approved or not approved in the problem). Logistic regression has become an important tool in the discipline of [machine learning](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML).

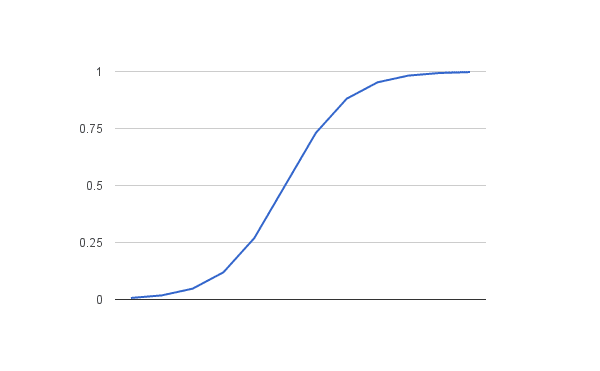
Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a [data set](https://whatis.techtarget.com/definition/data-set). As more relevant data comes in, the algorithm should get better at predicting classifications within data sets.

A logistic regression model predicts a [dependent data variable](https://whatis.techtarget.com/definition/dependent-variable) by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college. The resulting analytical model can take into consideration multiple input criteria. In the case of college acceptance, the model could consider factors such as the student’s grade point average, SAT score and number of extracurricular activities. Based on [historical data](https://whatis.techtarget.com/definition/historical-data) about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

**Logistic function**

The [logistic function](https://en.wikipedia.org/wiki/Logistic_function), is also called the sigmoid function. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

Graphical user interface

Description automatically generated

Input values (x) are combined linearly using weights or coefficient values (Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

Logistic regression models the probability of the default class (e.g. the first class).

For example, if we are modeling people’s sex as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person’s height, or more formally:

P(sex=male|height)

Written another way, we are modeling the probability that an input (X) belongs to the default class (Y=1), we can write this formally as:

P(X) = P(Y=1|X)

The probability prediction must be transformed into a binary value (0 or 1) in order to actually make a probability prediction.

The coefficients (Beta values b) of the logistic regression algorithm must be estimated from your training data. This is done using maximum-likelihood estimation.

The best coefficients would result in a model that would predict a value very close to 1 (e.g. male) for the default class and a value very close to 0 (e.g. female) for the other class.

Given a height of 150cm is the person male or female. We have learned the coefficients of b0 = -100 and b1 = 0.6. Using the equation above we can calculate the probability of male given a height of 150cm or more formally P(male|height=150). We will use EXP() for e, because that is what you can use if you type this example into your spreadsheet:

y = e^(b0 + b1\*X) / (1 + e^(b0 + b1\*X))

y = exp(-100 + 0.6\*150) / (1 + EXP(-100 + 0.6\*X))

y = 0.0000453978687

Or a probability of near zero that the person is a male.

In practice we can use the probabilities directly. Because this is classification and we want a crisp answer, we can snap the probabilities to a binary class value, for example:

0 if p(male) < 0.5

1 if p(male) >= 0.5

**Where can it be used?**

An example of binary classification is predicting if a student will pass or fail an exam when the number of hours spent studying is provided as a feature and the variables for the response has two values: pass and fail.

Organizations can use insights from logistic regression outputs to enhance their business strategies so they can achieve their business goals, including reducing expenses or losses and increasing [ROI](https://searchcio.techtarget.com/definition/ROI) in marketing campaigns, for example.

An [e-commerce](https://searchcio.techtarget.com/definition/e-commerce) company that mails expensive promotional offers to customers would like to know whether a particular customer is likely to respond to the offers or not. For example, they’ll want to know whether that consumer will be a “responder” or a “non-responder.”

Likewise, a credit card company develops a model to decide whether to issue a credit card to a customer or not will try to predict whether the customer is going to default or not on the credit card based on such characteristics as annual income, monthly credit card payments and number of defaults. In banking parlance, this is known as default propensity modeling.

Logistic regression can be used in Healthcare to identify risk factors for diseases and plan preventive measures, Weather forecasting [apps](https://searchsoftwarequality.techtarget.com/definition/application) to predict rain and weather conditions, Voting apps to determine if voters will vote for a particular candidate, Insurance to predict the chances that a policy holder will die before the term of the policy expires based on certain criteria, such as gender, age and physical examination, Banking to predict the chances that a loan applicant will default on a loan or not, based on annual income, past defaults and past debts.

**Bibliography**

* <https://searchbusinessanalytics.techtarget.com/definition/logistic-regression>
* <https://www.sciencedirect.com/topics/computer-science/logistic-regression>
* <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>
* https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc