Logistic Regression

Tue, 24 May 2022 at 21:51

#logit #link #function #odds-ratio

Overview

DEFINITION

> *" Logistic regression is **a process of modeling the probability of a discrete outcome given an input variable**. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on." sciencedirect*

INPUT

Can be both discrete or continuous

OUTPUT

coefficients that dictate the relative impact of each variable, and a linear expression for predicting the log-odds ratio outcome as a function of drivers

Use Cases

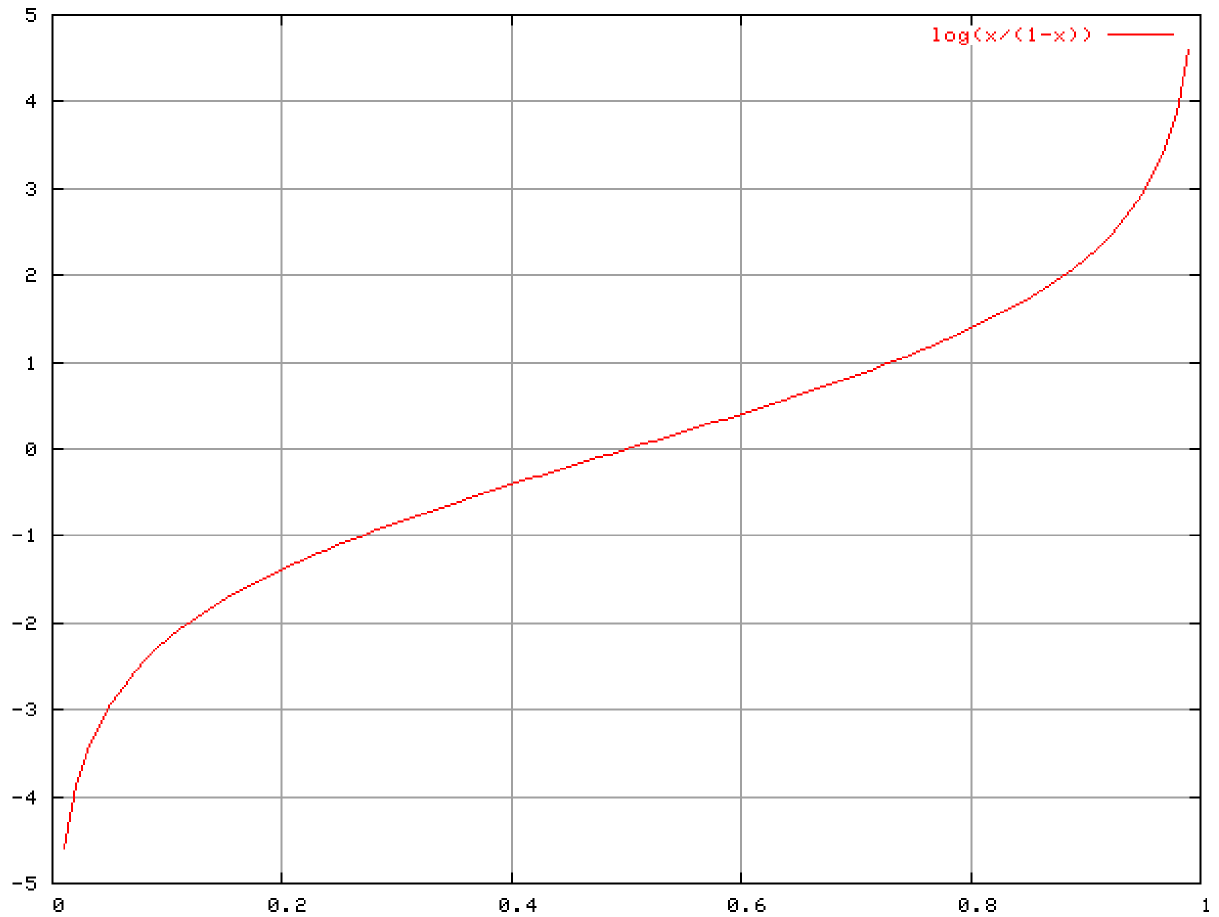
Used in classification and in the case of wanting the probability of an event happening. Examples:

- Probability that a borrower will default
- Probability that a customer will churn

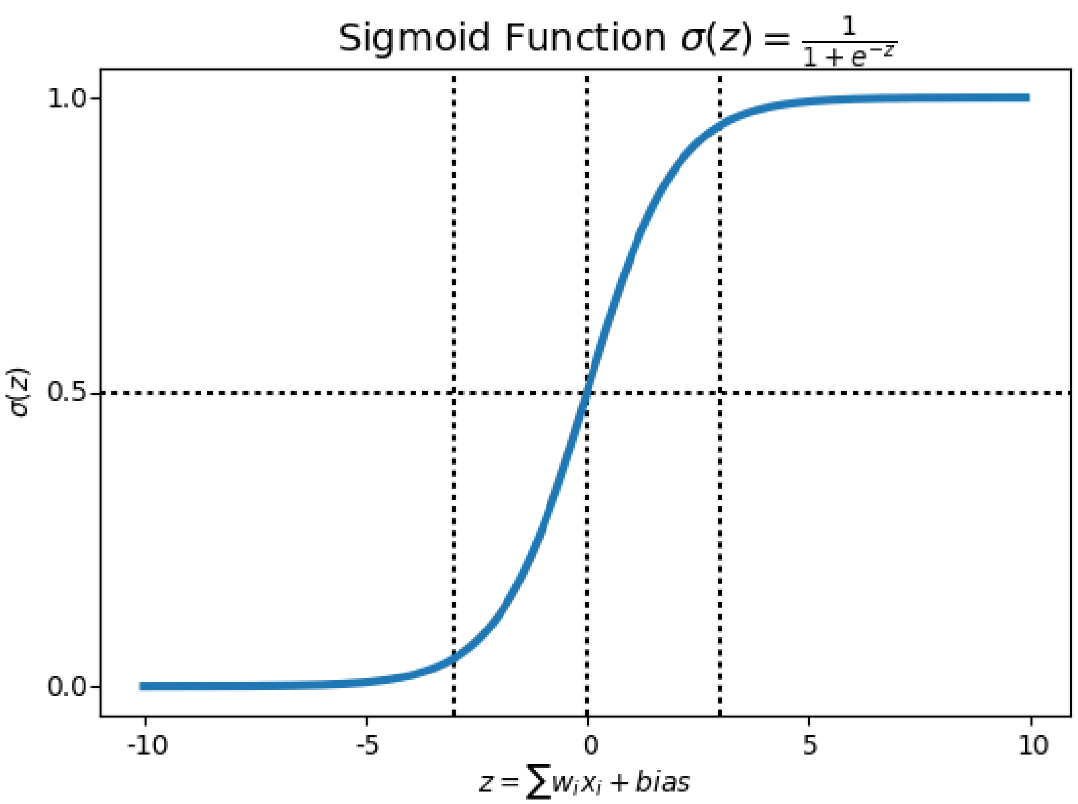
Terminology

1. *Probability* → $P(event) = \frac{outcome\, of\, interest}{all\, possible\, outcomes}$
2. *Odds* → $odds(event) = \frac{P}{1-P}$
3. *Odds Ratio* → $odds\, ratio = \frac{odds(event_1)}{odds(event_2)}$

4. ln(odds) $\rightarrow logit(p) = \ln(\frac{P}{1-P})$



5. inverse of logit $\rightarrow logit^{-1}(\alpha) = \frac{e^{\alpha}}{1+e^{\alpha}}$



6. *Maximum Likelihood Estimation* → Calculates the logistic regression coefficients
    1. Convert the classes using ln(odds) - this is the coefficients represented in a straight line
    2. Project them back using the inverse logit of the odds - the sigmoid function represents the logistic regression curve
    3. The goal is to estimate the probability. Since ln(odds)=$\beta_0 + \beta_1 x_1$ the antilog becomes the estimated probability $\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$

In logistic regression the **log odds ratio** is equal to linear additive combination of the drivers. The output of the sigmoid is the **actual probabilities**

INTERPETATION

1. e^{β_0} → the odds-ratio of the outcome in the `reference situation`
    1. all the continuous variables set to zero, and the categorical variables at their reference
2. e^{β_i} → tells us how the odds-ratio of y=1 changes for every unit change in x_i
    1. if $\beta_{credit\ score} = -0.69$
    2. then $e^{-0.69} = 0.5 = 1/2$
    3. meaning that for the same income, loan, and existing debt, the odds-ratio of default (y=1) is halved for every point increase in credit score

Pros & Cons

Reasons to Choose	Cautions
Explanatory value	Under the assumption that each variable affects the log-odds ratio linearly
Robust to redundant variables	Cannot handle variables that affect the outcome discontinuously
Easy to score	Does not handle missing values well
Concise representation with the coefficients	Does not work well with discrete drivers with a lot of values