Logistic Regression

Data Science | CCDATSCL

Logistic Regression is a statistical algorithm often used for **classification** and predictive analytics.

Logistic regression **estimates the probability of an event occurring** based on a given dataset of **independent variables**.

Its output is **0** or **1.**

The name “logistic regression” is derived from the concept of the **logistic function** that it uses.
The logistic function is also known as the **sigmoid function**.

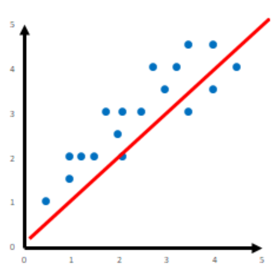
The value of this logistic function lies between **0** and **1,** which cannot go beyond this limit, so it forms a curve like an **“S” form**.

Applications

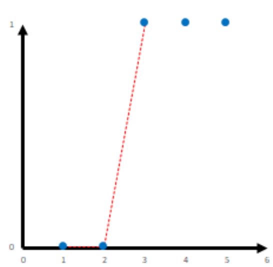
- Predicting whether a user will subscribe
- Predicting whether a patient will survive

Linear Regression vs Logistic Regression

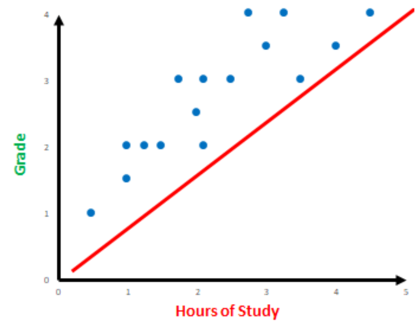
Linear Regression	Logistic Regression
It is used for solving regression problems	It is used for solving classification problems
We predict the value of continuous variables	We predict the value of categorical variables.
We find best fit line.	We find S-Curve.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.

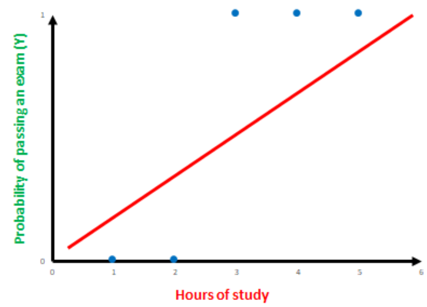


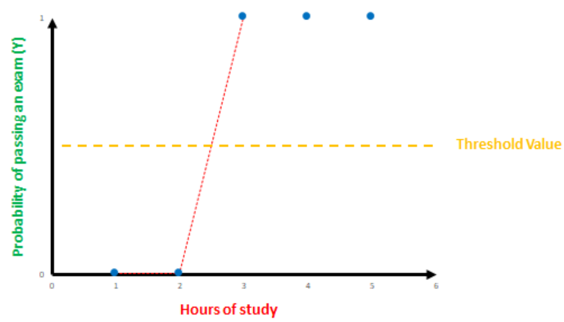
Linear Regression



Logistic Regression







Logistic/Sigmoid Function

$$y = \boldsymbol{\beta}_1(x) + \boldsymbol{\beta}_0$$

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

$$p(x) = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x)}}$$

where:
X is the value of the **independent variable**
$\boldsymbol{\beta}_1$ is the **slope/coefficient** of the line
$\boldsymbol{\beta}_0$ is the **y-intercept**

$$p(x) = \frac{e^{-(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x)}}{1 + e^{(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x)}}$$

where:
X is the value of the **independent variable**
$\boldsymbol{\beta}_1$ is the **slope/coefficient** of the line
$\boldsymbol{\beta}_0$ is the **y-intercept**

Maximum Likelihood Estimation

Maximum likelihood estimation is a method that determines values for the parameters of a model.

The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.

In Logistic Regression, MLE is used as the **Cost Function**.

$\mathcal{L}\left(y \mid x,\ \theta\right) = \mathbb{P} \left(y_1 \mid x_1\ \beta\right) * \mathbb{P}\left(y_2 \mid x_2\ \beta\right) \ldots * \mathbb{P} \left(y_n \mid x_n\ \beta\right)$

$$\mathcal{L}\left(y \mid x, \theta\right) = \prod_{i=1}^{n} \boldsymbol{P}\left(\boldsymbol{x_i}\right)^{y_i} \left(\mathbf{1} - \boldsymbol{P}\left(\boldsymbol{x_i}\right)\right)^{\mathbf{1}-y_i}$$

The likelihood function $\boldsymbol{\mathcal{L}}$ represents how plausible the model is with the parameters $\boldsymbol{\theta}$ given all the data points.

Assuming all of the data points are independent.

Log Likelihood

$$\mathcal{L}\left(\theta\right) = \prod_{i=1}^{n} \boldsymbol{P}\left(\boldsymbol{x_i}\right)^{y_i} \left(\mathbf{1} - \boldsymbol{P}\left(\boldsymbol{x_i}\right)\right)^{\mathbf{1}-y_i}$$

$$\mathcal{L}\left(\theta\right) = \sum_{i=1}^{n} y_i \ \log\left(\sigma\left(\theta^T \bar{x}_i\right)\right) + \left(1 - y_i\right) \ \log\left(1 - \sigma\left(\theta^T x_i\right)\right)$$

Gradient Ascent

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{y}{\sigma\left(\theta^T \bar{x}\right)} \frac{\partial \sigma(\theta^T \bar{x})}{\partial \theta} + \frac{(1-y)}{1 - \sigma(\theta^T \bar{x})} \frac{-\partial \sigma(\theta^T \bar{x})}{\partial \theta}$$

$$\frac{\partial L(\theta)}{\partial \theta} = \left(y - p\left(x\right)\right) \bar{x}$$

$$\frac{\partial L(\theta)}{\partial \theta} = \left(y - \sigma\left(\theta^T \bar{x}\right)\right) \bar{x}$$

Get the derivative of the log likelihood with respect to theta $\boldsymbol{\theta}$

$$\theta^{+} = \theta^{-} + \alpha \left(\frac{\partial L(\theta)}{\partial \theta}\right)$$

$$\boldsymbol{\theta}^{+} = \boldsymbol{\theta}^{-} + \boldsymbol{\alpha}\left(\boldsymbol{y} - \boldsymbol{p}\left(\boldsymbol{x}\right)\right) \overline{\boldsymbol{x}}$$

Equation to update the logistic regression model parameters $\boldsymbol{\theta}$