WISE

Introduction to Logistic Regression

Tang Dexuan

2017.10.20

What is logistic regression?

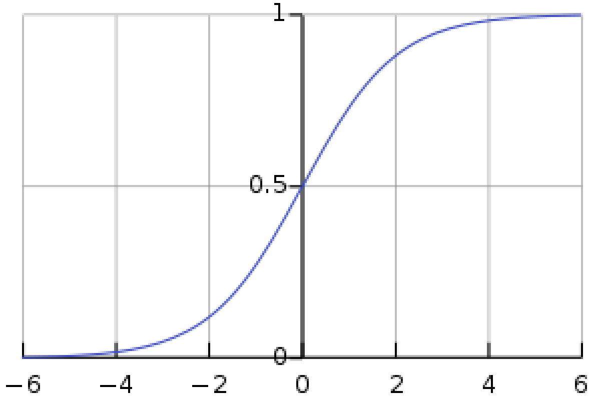
In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable—that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the dependent variable has more than two outcome categories may be analysed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model.

Logistic Function

- An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is useful because it can take any real input t,(t∈ R),whereas the output always takes values between zero and one and hence is interpretable as a probability. The logistic function $\sigma(t)$ is defined as follows:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

- A graph of the logistic function on the *t*-interval (-6,6) is shown in Figure 1.



Model fitting

Logistic regression is an important machine learning algorithm. The goal is to model the probability of a random variable Y being 0 or 1 given experimental data. Consider a generalized linear model function parameterized by θ,

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

If we attempt to model the probability that y is 0 or 1 with the function,

$$Pr(y|x;\theta) = h_\theta(x)^y (1 - h_\theta(x))^{(1-y)}$$

we take our likelihood function assuming that all the samples are independent,

$$\begin{aligned} L(\theta|x) &= Pr(Y|X;\theta) \\ &= \prod_i Pr(y_i|x_i;\theta) \\ &= \prod_i h_\theta(x_i)^{y_i} (1 - h_\theta(x_i))^{(1-y_i)} \end{aligned}$$

Model fitting

Typically, the log likelihood is maximized with a normalizing factor N^{-1},

$$N^{-1} \log L(\theta|x) = N^{-1} \sum_{i=0}^{N} \log Pr(y_i|x_i;\theta)$$

which is maximized using a something like gradient descent.
Assuming the (x,y) pairs are drawn uniformly from the underlying distribution, then in the limit of large N,

$$\begin{aligned}
\lim_{N \to +\infty} N^{-1} \sum_{i=0}^{N} \log Pr(y_i|x_i;\theta) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Pr(X=x, Y=y) \log Pr(Y=y|X=x;\theta) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Pr(X=x, Y=y) \left(-\log \frac{Pr(Y=y|X=x)}{Pr(Y=y|X=x;\theta)} + \log Pr(Y=y|X=x) \right) \\
&= -D_{KL}(Y \| Y_\theta) - H(Y|X)
\end{aligned}$$

where H(X|Y) is the conditional entropy and D_{KL} is the Kullback-Leibler divergence. This leads to the intuition that by maximizing the log-likelihood of a model, you are minimizing the KL divergence of your model from the maximal entropy distribution. Intuitively searching for the model that makes the least number of assumptions in its parameters.

Thank you

Tang Dexuan

2017.10.19