Notes on Logistic Regression

Gabriele Tolomei

April 16, 2020

1 Probabilistic Interpretation of Logistic Function

Let Y be a binary random variable, i.e., a random variable whose possible outcomes are $\{0, 1\}$. We can assume this variable being distributed according to a *Bernoulli* distribution with (unknown) parameter p, that is $Y \sim Bernoulli(p)$. Now, suppose we want to estimate the unknown parameter p, from a set of m i.i.d. observations of $\{(X_i, Y_i)\}_{i=1}^{m}$, where each $X_i = (x_{i,1}, \ldots, x_{i,d})$ is a d-dimensional random vector of features.

The goal of logistic regression is to find the best estimate of p, namely to learn the probability of any given example X to be labelled as $Y = 1$. More formally, logistic regression tries to estimate the *posterior probability*:

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

Estimating "directly" p using standard linear regression – i.e., fitting a regression line using input features like $p = \theta_0 + \theta_1 x_1 + \ldots \theta_d x_d = \boldsymbol{\theta}^T \cdot X$ – may not be the best solution here. Indeed, the output of any linear regression is generally ranging in $(-\infty, +\infty)$, whilst probabilities must range in $[0, 1]$. Enforcing the output of a standard linear regression to be "squashed" inside that range might require very complicated constraints on the parameters of the model $\boldsymbol{\theta}$ to be learned. Therefore, let us rewrite the definition above of p using Bayes' rule of conditional probability:

$$p = P(Y = 1|X) = \frac{\overbrace{P(X|Y = 1)}^{\text{likelihood}}\overbrace{P(Y = 1)}^{\text{prior}}}{\underbrace{P(X|Y = 1)P(Y = 1) + P(X|Y = 0)P(Y = 0)}_{\text{marginal}=P(X)}}$$

Similarly:

$$1 - p = P(Y = 0|X) = \frac{P(X|Y = 0)P(Y = 0)}{P(X|Y = 1)P(Y = 1) + P(X|Y = 0)P(Y = 0)}$$

We define the *odds* (of success) as the ratio between the probability of success (i.e., $Y = 1$) and the probability of failure (i.e., $Y = 0$).

$$\text{odds}(p) = \frac{p}{1-p} = \frac{P(X|Y=1)P(Y=1)}{P(X)} * \frac{P(X)}{P(X|Y=0)P(Y=0)} =$$

$$= \frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)}$$

Now, assuming a *uniform* prior over all the possible values of Y (i.e., the 2 events "success" and "failure" are considered equally likely), then $P(Y=1) = P(Y=0) = 0.5$ and the equation above can be simplified to:

$$\text{odds}(p) = \frac{p}{1-p} = \frac{P(X|Y=1)}{P(X|Y=0)}$$

Let us now go back to the initial definition of p:

$$p = P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1) + P(X|Y=0)P(Y=0)}$$

Assuming again that $P(Y=1) = P(Y=0) = 0.5$, we can rewrite the above equation as follows:

$$p = P(Y=1|X) = \frac{P(X|Y=1)}{P(X|Y=1) + P(X|Y=0)}$$

We can divide both the numerator and the denominator by $P(X|Y=1) \neq 0$:

$$p = P(Y=1|X) = \frac{1}{1 + \frac{P(X|Y=0)}{P(X|Y=1)}}$$

Note that it always holds that $a = e^{\log_e(a)}$, therefore:

$$p = P(Y=1|X) = \frac{1}{1 + e^{\log_e\left[\frac{P(X|Y=0)}{P(X|Y=1)}\right]}}$$

Moreover, $\frac{P(X|Y=0)}{P(X|Y=1)} = \frac{1-p}{p} = \frac{1}{\text{odds}(p)}$. As such:

$$p = P(Y=1|X) = \frac{1}{1 + e^{\log_e\left[\frac{1}{\text{odds}(p)}\right]}} = \frac{1}{1 + e^{-\log_e\left[\text{odds}(p)\right]}}$$

Because odds range between 0 and $+\infty$ (i.e., when $p = 0$ and $p = 1$, respectively), applying the natural logarithm make them taking values on the *whole* spectrum of real numbers (i.e., from $-\infty$ to $+\infty$). The natural logarithm of odds is called *logit*, and it can be used as the (continuous) response variable we want to predict from our input features using standard linear regression. In other words:

$$\text{logit}(p) = \log_e\left[\text{odds}(p)\right] = \log_e\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x1 + \ldots + \theta_d x_d = \boldsymbol{\theta}^T \cdot X$$

By substituting the above expression into the latest equation, we will obtain the following:

$$p = P(Y = 1|X) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \cdot X}}$$

The equation above is exactly the definition of the logistic sigmoid function $\ell(z) = \frac{1}{1+e^{-z}}$ whose input is the linear signal, namely $z = \boldsymbol{\theta}^T \cdot X$

2 Interpretation of Logistic Regression Coefficients

As we already said, logistic regression coefficients have a nice, natural interpretation since the output of the linear signal ($\boldsymbol{\theta}^T \cdot X$) is expressed in terms of the natural logarithm of the odds. This means that the effect of a change on one input feature x_j ($j = \{1, \ldots, d\}$) is measured as the change in the natural logarithm of odds. We already proved that such a change does not depend on the actual value of the feature we plug in; in fact, the effect is constant since the *odds ratio* is constant.
To clarify this better, suppose we have 2 input data points: $X = (x_1, \ldots, x_i, \ldots, x_d)$ and $X' = (x_1, \ldots, x_i + 1, \ldots, x_d)$, where X' is the same as X, except for the value of the i-th feature, which has been increased by 1 unit. First of all, let us work out what are the log-odds associated with X and X' respectively:

$$\log_e \left[\text{odds}(p)\right] = \log_e \left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x_1 + \ldots \theta_i x_i + \ldots + \theta_d x_d = \boldsymbol{\theta}^T \cdot X$$

or, analogously:

$$\text{odds}(p) = \frac{p}{1-p} = e^{\theta_0 + \theta_1 x_1 + \ldots \theta_i x_i + \ldots + \theta_d x_d} = e^{\boldsymbol{\theta}^T \cdot X}$$

Similarly, for X':

$$\log_e \left[\text{odds}(p')\right] = \log_e \left(\frac{p'}{1-p'}\right) = \theta_0 + \theta_1 x_1 + \ldots \theta_i (x_i + 1) + \ldots + \theta_d x_d = \boldsymbol{\theta}^T \cdot X'$$

or, analogously:

$$\text{odds}(p') = \frac{p'}{1-p'} = e^{\theta_0 + \theta_1 x_1 + \ldots \theta_i (x_i + 1) + \ldots + \theta_d x_d} = e^{\boldsymbol{\theta}^T \cdot X'}$$

Let us take the ratio of the two odds above, indeed the *odds ratio*.

$$\frac{\text{odds}(p')}{\text{odds}(p)} = \frac{e^{\theta_0 + \theta_1 x_1 + \ldots \theta_i (x_i + 1) + \ldots + \theta_d x_d}}{e^{\theta_0 + \theta_1 x_1 + \ldots \theta_i x_i + \ldots + \theta_d x_d}} = \frac{e^{\theta_0 + \theta_1 x_1 + \ldots \theta_i x_i + \ldots + \theta_d x_d} * e^{\theta_i}}{e^{\theta_0 + \theta_1 x_1 + \ldots \theta_i x_i + \ldots + \theta_d x_d}} = e^{\theta_i}$$

The first thing to notice is that the odds ratio **does not** depend on the value of x_i: no matter whether $x_i = 10$ or $x_i = 10^6$, the effect of adding to it 1 unit (i.e., $x_i = 10 + 1$ or $x_i = 10^6 + 1$) on the odds ratio will be the same.

Of course, odds themselves are not constant at all! Indeed, if we compute the odds for X where $x_i = 10$ and then for X where $x_i = 10^6$, those will be clearly different.

Example. Suppose we apply logistic regression to predict the probability a company X will default; the event "default" can be represented by a binary random variable Y, which evaluates to 1 if the default occurs, 0 otherwise. Eventually, we want to use logistic regression to estimate $p = P(Y = 1|X)$.
Let us assume X is represented by just 2 features, i.e., $X = (x_1, x_2)$, where $x_1 =$ `profile` and $x_2 =$ `annual revenue` (in millions of dollars). For the sake of simplicity, we assume `profile` is a binary feature taking on 2 values: `startup` (denoting a company which has just got to the market, indicated by 1) or `consolidated` (denoting a company which has been active on the market since a long time, and indicated by 0). Among the output of our logistic regression model, there is also the value of the coefficients θ_1 and θ_2 associated with x_1 and x_2, respectively. Suppose that the odds ratio associated with θ_1 is equal to 1.15 ($e^{\theta_1} = 1.15$, or analogously, $\theta_1 = 0.14$), whilst the odds ratio associated with θ_2 is equal to 0.64 ($e^{\theta_2} = 0.64$, or analogously, $\theta_1 = -0.45$). When the odds ratio is greater than 1, it describes a *positive relationship* (such as for θ_1), and it can be interpreted as follows: by increasing feature x_1 by 1 unit the odds of default (i.e., the odds of our target event happens) increase by 1.15 times (or +15%). Since feature x_1 is binary, increasing it by 1 unit means switching it from 0 (`consolidated`) to 1 (`startup`). In other words, switching from a consolidated company to a startup increases the odds of a default by 1.15 times (+15%). Conversely, when the odds ratio is smaller than 1, it describes a *negative relationship* (such as for θ_2), and it can be interpreted as follows: by increasing feature x_2 by 1 unit the odds of default increase by 0.64 times (actually, reducing them by 36%). Differently from feature x_1, feature x_2 is continuous: increasing it by 1 unit means adding 1 million of dollars to the `annual revenue`. Therefore, adding 1 million of dollars of annual revenue to a company "increases" its odds of default by a factor of 0.64, namely it decreases the odds of default by 36%.
Notice, again, that odds themselves are not constant! For instance, suppose $x_1 = 1$ and $x_2 = 5$ and let $\theta_0 = 0$, then:

$$\text{odds}(p) = e^{\theta_0 + \theta_1 x_1 + \theta_2 x_2} = e^{0 + 0.14 * 1 - 0.45 * 5} = 0.12$$

On the other hand, if we fix $x_1' = x_1 = 1$ and let $x_2' = 20$, we obtain the following:

$$\text{odds}(p') = e^{\theta_0 + \theta_1 x_1' + \theta_2 x_2'} = e^{0 + 0.14 * 1 - 0.45 * 20} = 0.00014$$

The two odds above are clearly different. However, if we compute the odds ratio between the two quantities above we get:

$$\frac{\text{odds}(p')}{\text{odds}(p)} = \frac{0.00014}{0.12} \approx 0.0012$$

Since we know that for each unit increase of x_2 the odds ratio increase by 0.64, and $x_2' = x_2 + 15$, the odds ratio will increase by $(0.64)^{15} \approx 0.0012$.