1 The Definition and Property of Binomial Logistic Regression

The first thing needs to be clarified is that although the name contains the word "regression", it is a classification method. In this method, the regression is done on the probability of being classified into some classes, and therefore it can be used for classification. Now, we are focusing on showing how to solve the model using gradient descent method.

Definition 1.1. Denote $\boldsymbol{w} = (w_1, w_2, \cdots, w_n, b)^T$, $\boldsymbol{x} = (x_1, x_2, \cdots, x_n, 1)^T$, then the logistic regression model is defined as

$$P(Y = 1|\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})} = \pi(\boldsymbol{x}) \tag{1}$$

$$P(Y = 0|\boldsymbol{x}) = \frac{\exp(-\boldsymbol{w}^T\boldsymbol{x})}{1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})} = 1 - \pi(\boldsymbol{x}) \tag{2}$$

Now we can explore the property of the logistic regression model a little bit. The odds of an event is the ratio of that event happens to does not happen. If the probability of the event's occurrence is p, then the odds of the event is $\frac{p}{1-p}$. Therefore, the log odds [1](also called logit function) of the event is

$$\text{logit}(p) = \log \frac{p}{1-p} \tag{3}$$

For logistic regression, based on formula (1) and (2), we know that

$$\log \frac{P(Y = 1|\boldsymbol{x})}{1 - P(Y = 1|\boldsymbol{x})} = \log \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \boldsymbol{w}^T\boldsymbol{x} \tag{4}$$

which means the log odds of outputting $Y = 1$ is a linear function of $\boldsymbol{x}$. Or, you can say the log odds of outputting $Y = 1$ is a model, which is the logistic regression model, represented by a linear function of $\boldsymbol{x}$.

2 The Estimation of The Parameters of The Model

For a given training data set $T = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \cdots, (\boldsymbol{x_N}, y_N)\}$ [2], where $\boldsymbol{x_i} \in R^n$, $y_i \in \{0, 1\}$, we use the maximum likelihood as shown below for estimating the parameter $\boldsymbol{w}$ in the model.

$$\prod_{i=1}^{m} [\pi(\boldsymbol{x_i})]^{y_i} [1 - \pi(\boldsymbol{x_i})]^{1-y_i} \tag{5}$$

[1] The base of all the log functions, unless specifying explicitly, is the Euler's number e.

[2] Please aware that bold face letters represent vectors, that is $\boldsymbol{x_i} = (x_1, x_2, \cdots, x_n, 1)^T$, $(i = 1, 2, \cdots, n)$.

and the log likelihood function is

$$L(\boldsymbol{w}) = \sum_{i=1}^{m} \Big[y_i \log\, \pi(\boldsymbol{x_i}) + (1 - y_i) \log\, (1 - \pi(\boldsymbol{x_i}) \Big] \tag{6}$$

For estimating w, the objective function is

$$\text{argmax } L(\boldsymbol{w}) \tag{7}$$

It is equivalent to

$$\text{argmin } -\frac{1}{m} L(\boldsymbol{w}) \tag{8}$$

Let $J(\boldsymbol{w}) = -\frac{1}{m} L(\boldsymbol{w})$, then the objective function becomes

$$\text{argmin } J(\boldsymbol{w}) \tag{9}$$

To solve problem (9), we use gradient descent method [3]. That is,

$$\boldsymbol{w^{(i+1)}} = \boldsymbol{w^{(i)}} - \alpha \nabla J(\boldsymbol{w}^{(i)}) \tag{10}$$

where α is the learning rate, and i represents the i^{th} iteration. For the j^{th} item in $\boldsymbol{w}$, the way to update w_j is

$$w_j^{(i+1)} = w_j^{(i)} - \alpha \frac{\partial J}{\partial w_j} \left(\boldsymbol{w}^{(i)} \right) \quad {}^4, \; (j = 0, 1, \cdots, n) \tag{11}$$

and

$$\frac{\partial J}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i \frac{1}{\pi(\boldsymbol{x_i})} \frac{\partial \pi(\boldsymbol{x_i})}{\partial w_j} - (1 - y_i) \frac{1}{1 - \pi(\boldsymbol{x_i})} \frac{\partial \pi(\boldsymbol{x_i})}{\partial w_j} \right] \tag{12}$$

Since

$$\begin{aligned} \frac{\partial \pi(\boldsymbol{x_i})}{\partial w_j} &= -\frac{1}{(1 + \exp(-\boldsymbol{w^T x_i}))^2} \; \frac{\partial (1 + \exp(-\boldsymbol{w^T x_i}))}{\partial w_j} \\ &= -\frac{1}{(1 + \exp(-\boldsymbol{w^T x_i}))^2} \exp(-\boldsymbol{w^T x_i}) \; \frac{\partial (-\boldsymbol{w^T x_i})}{\partial w_j} \\ &= \frac{1}{(1 + \exp(-\boldsymbol{w^T x_i}))^2} \exp(-\boldsymbol{w^T x_i}) \; x_{ij} \\ &= \frac{1}{1 + \exp(-\boldsymbol{w^T x_i})} \; \frac{\exp(-\boldsymbol{w^T x_i})}{1 + \exp(-\boldsymbol{w^T x_i})} \; x_{ij} \\ &= \pi(\boldsymbol{x_i})(1 - \pi(\boldsymbol{x_i}))x_{ij} \end{aligned} \tag{13}$$

[3] Here we are using the batch gradient descent. Other common ways are SGD (stochastic gradient descent) and mini-batch gradient descent. Please Google for more details if you are interested in them.

[4] $\frac{\partial J}{\partial w_j} \left(\boldsymbol{w}^{(i)} \right)$ is the value of the partial derivative in $\boldsymbol{w}^{(i)}$.

the result in (12) can be simplified to

$$\frac{\partial J}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m}(\pi(\boldsymbol{x_i}) - y_i)x_{ij} \tag{14}$$

Therefore, based on (11) and (14), the way to update w_j is

$$w_j^{(i+1)} = w_j^{(i)} - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{1+\exp(-(\boldsymbol{w}^{(i)})^T\boldsymbol{x_i})} - y_i\right)x_{ij}\right] \tag{15}$$

Finally, if $\left|w_j^{(i+1)} - w_j^{(i)}\right| < \textit{threshold}$, or the number of iterations exceeds the specified maximum iteration times, then stop the updating process.

3 Intuitive Analysis

In this section, we intuitively analyze why the logistic regression algorithm could fail to converge if the training data are well separated. It is because at that circumstance, term $\pi(\boldsymbol{x_i}) - y_i$ in (14) is either 0 or 1. Thus, there could be a fluctuation in the parameters updating process and therefore the algorithm fails to converge.