Mathematical model for logistic regression

Professor Montaz Ali 25th January 2021

Logistic Regression (LR) incorporates two models: Linear Regression (which is a linear model) and a non-linear transformation. Here we consider a linear model (LM) as being linear in its parameters and not in its variables:

$$L_m(x) = \sum_{i=0}^{n} a_i x_i = a^T x \tag{1}$$

where, $x = (x_0, x_1, \cdots, x_n)^T$, $a_0 = b$ and $x_0 = 1$. The above LM is the linear combination of its inputs say x_i, $i = 1 \cdots n$. Inputs come from a known data set of size N. The j-th input will be denoted by x^j, $j = 1, 2, \cdots N$. The process of obtaining the optimal parameters of $L_m(x)$ is known as linear regression. Although we will not use linear classification here but it is needed to put things in perspective. Linear classification has two outcome 'yes' 'no' (1 or -1), i.e., binary classification. Two outcomes can be obtained from $L_m(x)$ by its sign i.e sign$(L_m(x))$=sign$(a^T x)$=$f(x)$. Here sign is acting as a non-linear transformation to convert $a^T x$ to $f(x)$, where $f(x) \in \{1, -1\}$.

We can consider another non-linear transformation $F(\cdot)$ such that $F(a^T x) = f(x)$ where $f(x) \in [0, 1]$. Such a transformation is known as logistic function and is given by

$$F(x) = \frac{e^x}{1 + e^x}. \tag{2}$$

It now follows that

$$f(x) = \frac{e^{a^T x}}{1 + e^{a^T x}} \tag{3}$$

is a probability which is a generalization of binary classification and has many applications but we will use this to predict the heart attacks.

When the data is collected individuals, the person may or may not have a heart attack. We use a parameter ω to denote this with $\omega = 1$ denoting the heart attack and $\omega = -1$ otherwise. When we build a model to predict ω, our noise target will be as follows

$$\Pr\left(\omega/x\right) = \begin{cases} f(x) & if \quad \omega = 1 \\ 1 - f(x) & if \quad \omega = -1 \end{cases}.$$

It follows that $\Pr\left(\omega/x\right) = F(\omega a^T x)$ since $F(-a^T x) = 1 - F(a^T x)$. We will use our data set

$$D = \{x^1, x^2, \cdots, x^N\}$$

to construct $L_m(x)$ such that $\Pr\left(\omega/x\right)$ predicts our corresponding data (ω/x) as closely as possible.

Consider the following likelihood function

$$L_f(a) = \prod_{i=1}^{N} \Pr\left(\omega^i/x^i\right) = \prod_{i=1}^{N} F(\omega^i a^T x^i) \tag{4}$$

We want to maximize (4) or equivalently to minimize some error measure. Since all are positive quantity in (4), we can write:

$$\max_a L_f(a) = \max_a \ln \prod_{i=1}^{N} F(\omega^i a^T x^i) = \sum_{i=1}^{N} \ln\left(F(\omega^i a^T x^i)\right).$$

We can then minimize the following

$$\min_a \bar{L}_f(a) = \min_a -\sum_{i=1}^{N} \ln\left(F(\omega^i a^T x^i)\right) = \min_a \sum_{i=1}^{N} \ln\left[1 + \exp\left(-\omega^i a^T x^i\right)\right] \tag{5}$$

which is a non-linear function and has to be optimized iteratively using gradient descent. The gradient of the function is given by

$$\nabla \bar{L}_f(a) = \sum_{i=1}^{N} \frac{-\omega^i x^i}{1 + e^{\omega^i a^T x^i}}. \tag{6}$$

The procedure is as follows:
1. Initialize a^k, for k=0, Calculate $\nabla \bar{L}_f(a^k)$,
2. Set $\alpha = 1$
3. Find $a^{k+1} = a^k - \alpha \times \nabla \bar{L}_f(a^k)$,
4. Compare if $L_f(a^{k+1}) < L_f(a^k)$ then set $k = k + 1$ and go to 5 else set $\alpha = \frac{1}{2} \times \alpha$ and go to 3.
5. Calculate $\nabla \bar{L}_f(a^{k+1})$. Stop if $\|\nabla \bar{L}_f(a^{k+1})\|$ is small else go to 2.