IRWLS algorithm for MLE in logistic regression

Xiao (Cosmo) Zhang

March 15, 2015

Unregularized and regularized logistic regression

In a Generalized Linear Model, we can express the distribution for y in the canonical form, which is

$$f(y;\theta) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\},$$

where θ is called canonical parameter and ϕ is called the dispersion parameter. And the log-likelihood is

$$l(y;\theta) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi).$$

We can obtain the first derivative or score of the log-likelihood w.r.t θ, which is

$$l'(\theta;y) = \frac{y - b'(\theta)}{a(\phi)},$$

and the second derivative, which is

$$l''(\theta;y) = -\frac{b''(\theta)}{a(\phi)}.$$

By the known property, $E[l'(\theta;y)] = 0$, it follows that $\mu = b'(\theta)$. The information equality shows

$$\text{var}[l'(\theta;y)] = E[l'^2(\theta;y)] - E^2[l'(\theta;y)] = -E[l''(\theta;y)]$$

$$\Rightarrow E[l'^2(\theta;y)] = -E[l''(\theta;y)],$$

therefore it follows $E[\frac{(y-\mu)^2}{a^2(\phi)}] = \frac{b''(\theta)}{a(\phi)}$. Thus, we have $\text{var}(y) = E[(y-\mathbf{u})^2] = b''(\theta)a(\phi)$.

For the canonical form of the GLM, we have the linear predictor

$$\eta = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x},$$

link function

$$g(\mu) = \eta,$$

and the obtained fact

$$\mu = b'(\theta).$$

Now we can establish the Fisher Scoring algorithm for the GLM model:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (-E[l''(\boldsymbol{\beta}^{(t)})])^{-1} l'(\boldsymbol{\beta}^{(t)}),$$

where $l'(\boldsymbol{\beta}^{(t)})$ is the score and $-E[l''(\boldsymbol{\beta}^{(t)})]$ is the expected information. $\forall \beta_j$, we have the chain equation:

$$\frac{\partial l}{\partial \beta_j} = (\frac{\partial l}{\partial \theta})(\frac{\partial \theta}{\partial \mu})(\frac{\partial \mu}{\partial \eta})(\frac{\partial \eta}{\partial \beta_j}).$$

And we can derive the expression of them with ease:

$$\begin{aligned}
\frac{\partial l}{\partial \theta} &= l'(\theta; y) = \frac{y - b'(\theta)}{a(\phi)} \\
\frac{\partial \theta}{\partial \mu} &= \frac{1}{\frac{\partial \mu}{\partial \theta}} = \frac{1}{b''(\theta)} = \frac{a(\phi)}{\mathrm{var}(y)} \\
\frac{\partial \eta}{\partial \beta_j} &= x_{ij}, \forall \mathbf{x}_i
\end{aligned}$$

Combining the above results, we have

$$\frac{\partial l}{\partial \beta_j} = \frac{(y - \mu)}{\mathrm{var}(y)} (\frac{\partial \mu}{\partial \eta}) x_{ij}, \tag{1}$$

and by using the property of information equality and the derived results above, we have

$$\begin{aligned}
-E[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}] &= E[(\frac{\partial l}{\partial \beta_j})(\frac{\partial l}{\partial \beta_k})] \\
&= E[(\frac{y - \mu}{\mathrm{var}(y)})^2 (\frac{\partial \mu}{\partial \eta})^2 x_{ij} x_{ik}] \\
&= \frac{1}{\mathrm{var}(y)} (\frac{\partial \mu}{\partial \eta})^2 x_{ij} x_{ik}.
\end{aligned} \tag{2}$$

With equations 1 and 2, we can construct the Fisher scoring algorithm. By rewriting equation 1 in the vector form, we obtain

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}} \mathbf{A} (\mathbf{y} - \boldsymbol{\mu}),$$

where

$$\mathbf{X} = (\mathbf{x}_1, \ldots, x_N)^{\mathrm{T}},$$

$$\mathbf{A} = \text{diag}\{[\text{var}(y_i)]^{-1}(\frac{\partial \mu_i}{\partial \eta_i})\},$$
$$\mathbf{y} = (y_1, \ldots, y_N),$$

and

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N).$$

Similarly, we have

$$-E[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}] = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X},$$

where

$$\begin{aligned}\mathbf{W} &= \text{diag}\{w_i\} \\ &= \text{diag}\{[\text{var}(y_i)]^{-1}(\frac{\partial \mu_i}{\partial \eta_i})^2\} \\ &= \text{diag}\{[\text{var}(y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2]^{-1}\}.\end{aligned}$$

Hence, we can construct the Fisher scoring as

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \{-E[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}]\}^{-1}\frac{\partial l}{\partial \beta_j} \Rightarrow \\ \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A}(\mathbf{y} - \boldsymbol{\mu}). \qquad (3)\end{aligned}$$

By rewriting equation 3, we have

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}[\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{X}^{\mathrm{T}}\mathbf{A}(\mathbf{y} - \boldsymbol{\mu})]. \qquad (4)$$

Since $\eta = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, we can write

$$\mathbf{X}\boldsymbol{\beta} = (\eta_1, \ldots, \eta_N)^{\mathrm{T}} = \boldsymbol{\eta}.$$

And

$$\mathbf{A} = \mathbf{W}(\frac{\partial \eta}{\partial \mu}),$$

where $\frac{\partial \eta}{\partial \mu} = \text{diag}(\frac{\partial \eta_i}{\partial \mu_i})$. Hence, we can write equation 4 as

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z}, \qquad (5)$$

where

$$\mathbf{z} = \boldsymbol{\eta} + (\frac{\partial \eta}{\partial \mu})(\mathbf{y} - \boldsymbol{\mu}) = (z_1, \ldots, z_N)^{\mathrm{T}},$$

and elementwisely,

$$z_i = \eta_i + (\frac{\partial \eta_i}{\partial \mu_i})(y_i - \mu_i), \qquad (6)$$

and

$$w_i = [\text{var}(y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2]^{-1}. \tag{7}$$

In the logistic regression, letting $x_0 = 1$, then we can write $\beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ as $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$. For each individual trial, we have the canonical form

$$\begin{aligned} f(y) &= \exp\{\log\{(\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{1+\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})})^y(\frac{1}{1+\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})})^{1-y}\} \\ &= \exp\{y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - \log[1+\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})]\}, \end{aligned}$$

and the log-likelihood is

$$l = y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - \log[1+\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})].$$

Hence, we can write the corresponding conponents of the canonical form:

$$\begin{aligned} &y = y \\ &\theta = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \\ &a(\phi) = 1 \\ &b(\theta) = \log\{1+\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} \\ &c(y,\phi) = 0. \end{aligned}$$

By easy derivation, we can also obtain the following relations:

$$\begin{aligned} &E[y] = \mu = b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \pi \\ &\text{var}(y) = a(\phi)b''(\theta) = b''(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}\frac{1}{1+\exp(\theta)} = \pi(1-\pi) = \mu(1-\mu) \\ &\eta = g(\mu) = logit(\pi) = \log(\frac{\pi}{1-\pi}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \text{ where } \pi = Pr(G=1|\mathbf{x}) = \frac{\exp(\theta)}{1+\exp(\theta)} \end{aligned}$$

Since $\mu = \pi$,

$$\eta = \log(\frac{\mu}{1-\mu}) = -\log(\mu^{-1}-1).$$

Therefore, we can obtain

$$\frac{\partial \eta}{\partial \mu} = \frac{\mu^{-1}}{1-\mu} = \frac{1}{(1-\mu)\mu} = \frac{1}{p(\mathbf{x})(1-p(\mathbf{x}))},$$

where $p(\mathbf{x}) = \pi = Pr(G=1|\mathbf{x})$. Plugging back to equations 6 and 7, we get

$$z_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \frac{y - p(\mathbf{x}_i)}{p(\mathbf{x}_i)(1-p(\mathbf{x}_i))}$$

and

$$\begin{aligned} w_i &= [\mathrm{var}(y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2]^{-1} \\ &= [\pi(1-\pi)(\frac{1}{\pi(1-\pi)})^2]^{-1} \\ &= \pi(1-\pi) \\ &= p(\mathbf{x}_i)(1-p(\mathbf{x}_i)). \end{aligned}$$

Evaluating at the current parameter $\tilde{\boldsymbol{\beta}}$, we can obtain

$$z_i = \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i + \frac{y_i - \tilde{p}(\mathbf{x}_i)}{\tilde{p}(\mathbf{x}_i)(1-\tilde{p}(\mathbf{x}_i))}$$

as the working variate, and

$$w_i = [\mathrm{var}(y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2]^{-1} = \tilde{p}(\mathbf{x}_i)(1-\tilde{p}(\mathbf{x}_i))$$

as the weight, which are corresponding to the content in the paper.

From the Equation 15 in the paper, which is the form of l_Q, we can have the same algorithm by constructing the Fisher Scoring, by using the Newton–Raphson method first. In this case, the iteration can be formed in this way:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{l_Q^{'}}{l_Q^{''}},$$

and assyptotically, this is equavalent to

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{l_Q^{'}}{E[l_Q^{''}]},$$

which is the Fisher Scoring. A easy rearrange of the terms will lead this equivalanet to equation 5, by using the given z_i and w_i in the paper.

Putting all the above together, we have the IRWLS update algorithm to find the MLE in a logistic regression model:

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z},$$

where $\mathbf{z} = (z_1, \ldots, z_i, \ldots, z_N)$, and $\mathbf{W} = \mathrm{diag}(w_i)$. Also, theoratically, a cholesky decomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}$ that is s.p.d to $\mathbf{L}\mathbf{L}^{\mathrm{T}}$, instead of computing the inverse, can accelerate the speed, by forming $\mathbf{L}\mathbf{L}^{\mathrm{T}}\boldsymbol{\beta}^{(t+1)} = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z}$.