Logistic Regression and Newton-Raphson

Rahul Ramachandran
cs21btech11049

Rishit D
cs21btech11053

October 8, 2023

Contents

1 Gradient, Hessian and Newton-Raphson

1.1 Error Function

The error function is given by:

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \ln(y_n) + (1-t_n)\ln(1-y_n)\right)$$

This can alternatively be written as:

$$E(y_1, y_2, \ldots, y_N) = -\sum_{n=1}^{N} \left(t_n \ln(y_n) + (1-t_n)\ln(1-y_n)\right)$$

to show the dependance on y_is. To derive the gradient and the hessian, we will use the numerator layout. Note that the design matrix Φ is given by:

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

We represent the ith row of Φ as ϕ_i.

1.2 Gradient

We find the gradient of E with respect to $\mathbf{w}$ by using the total-derivative chain rule:

$$\nabla_{\mathbf{w}} E = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}} \tag{1}$$

First, we find $\frac{\partial E}{\partial y_n}$:

$$\frac{\partial E}{\partial y_n} = -\left(\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right) \tag{2}$$

$$= \frac{y_n - t_n}{y_n(1-y_n)} \tag{3}$$

To find $\frac{\partial y_n}{\partial \mathbf{w}}$, we note the following:

$$a_n = \mathbf{w}^T \phi_n \tag{4}$$

$$y_n = \sigma(a_n) \tag{5}$$

where σ is the sigmoid function. Therefore, we have:

$$\frac{\partial y_n}{\partial \mathbf{w}} = \frac{\partial y_n}{\partial a_n} \frac{\partial a_n}{\partial \mathbf{w}} \tag{6}$$

$$= y_n(1-y_n)\phi_n^T \tag{7}$$

where we've used the fact that $\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1-\sigma(x))$. Combining (3) and (7), we get:

$$\frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}} = (y_n - t_n)\phi_n^T \tag{8}$$

Using (8) in (1), we get:

$$\nabla_{\mathbf{w}} E = \sum_{n=1}^{N} (y_n - t_n)\phi_n^T \tag{9}$$

$$= (\mathbf{y} - \mathbf{t})^T \Phi \tag{10}$$

1.3 Hessian

From the last section, we obtained the gradient of E with respect to $\mathbf{w}$ as:

$$\nabla_{\mathbf{w}} E = \sum_{n=1}^{N} (y_n - t_n)\phi_n^T$$

The Hessian is given by the transpose of the Jacobian of the gradient:

$$\nabla_{\mathbf{w}}^T \nabla_{\mathbf{w}} E = \nabla_{\mathbf{w}}^T \left(\sum_{n=1}^{N} (y_n - t_n)\phi_n^T\right) \tag{11}$$

$$= \sum_{n=1}^{N} \nabla_{\mathbf{w}}^T (y_n - t_n)\phi_n^T \tag{12}$$

$$= \sum_{n=1}^{N} (\nabla_{\mathbf{w}} (y_n - t_n)\phi_n^T)^T \tag{13}$$

Now, we find $\nabla_{\mathbf{w}}(y_n - t_n)\phi_n^T$:

$$\left(\frac{\partial(y_n - t_n)\phi_n^T}{\partial \mathbf{w}}\right)_i = \frac{\partial(y_n - t_n)\phi_{i-1}(x_n)}{\partial \mathbf{w}} \tag{14}$$

$$= \frac{\partial(y_n - t_n)}{\partial \mathbf{w}}\phi_{i-1}(x_n) \tag{15}$$

$$= \frac{\partial y_n}{\partial \mathbf{w}}\phi_{i-1}(x_n) \tag{16}$$

$$= y_n(1 - y_n)\phi_{i-1}(x_n)\phi_n^T \tag{17}$$

Therefore, we have:

$$\nabla_{\mathbf{w}}(y_n - t_n)\phi_n^T = y_n(1 - y_n)\phi_n\phi_n^T$$

Using this in the expression for the Hessian, we get:

$$\nabla_{\mathbf{w}}^T\nabla_{\mathbf{w}}E = \sum_{n=1}^{N}(y_n(1 - y_n)\phi_n\phi_n^T)^T \tag{18}$$

$$= \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^T \tag{19}$$

$$= \Phi^T R\Phi \tag{20}$$

where R is the diagonal matrix with $y_n(1 - y_n)$ on the diagonal.

1.4 Newton-Raphson

The Newton-Raphson update is given by:

$$\mathbf{w}^{new} = \mathbf{w}^{old} - (H)^{-1}\nabla_{\mathbf{w}}^T E$$

where H is the Hessian. Using (10) and (20), we get the following update equation:

$$\mathbf{w}^{new} = \mathbf{w}^{old} - (\Phi^T R\Phi)^{-1}\Phi^T(\mathbf{y} - \mathbf{t}) \tag{21}$$

Please refer to Algorithm 1 for the algorithm to determine $\mathbf{x}$ to maximize likelihood.

2 Relation to Weighted Least Squares

The update equation for the Newton-Raphson method can be rewritten as:

$$\mathbf{w}^{new} = \mathbf{w}^{old} - (\Phi^T R\Phi)^{-1}\Phi^T(\mathbf{y} - \mathbf{t}) \tag{22}$$

$$= (\Phi^T R\Phi)^{-1}(\Phi^T R\Phi)\mathbf{w}^{old} - (\Phi^T R\Phi)^{-1}\Phi^T(\mathbf{y} - \mathbf{t}) \tag{23}$$

$$= (\Phi^T R\Phi)^{-1}(\Phi^T R\Phi\mathbf{w}^{old} - \Phi^T(\mathbf{y} - \mathbf{t})) \tag{24}$$

$$= (\Phi^T R\Phi)^{-1}(\Phi^T R(\Phi\mathbf{w}^{old} - R^{-1}(\mathbf{y} - \mathbf{t}))) \tag{25}$$

$$= (\Phi^T R\Phi)^{-1}\Phi^T R\mathbf{z} \tag{26}$$

$$\tag{27}$$

where $\mathbf{z} = \Phi\mathbf{w}^{old} - R^{-1}(\mathbf{y} - \mathbf{t}) \in \mathbf{R}^N$. This matches the form of the solution we obtained for weighted least squares. Here, the matrix R is not constant, and depends on the changing vector $\mathbf{w}$. Because of this, and since the update equation is iteratively applied, the Newton-Raphson method is also called *Iterative Reweighted Least Squares Method.*

Algorithm 1: Newton-Raphson Update Algorithm

Data: $\mathbf{\Phi}$: Design Matrix
$\mathbf{t}$: Output Vector
$\epsilon > 0$: Terminator
Result: $\mathbf{w}$ which maximizes log-likelihood.

$\mathbf{w} \leftarrow \mathbf{w_0}$; // w_0 is preferably close to the root
$\mathbf{y} = (\sigma(w^T\phi_1), \sigma(w^T\phi_2), ..., \sigma(w^T\phi_n))^T$; // σ refers to the sigmoid function
$grad \leftarrow (\mathbf{y} - \mathbf{t})^T\mathbf{\Phi}$;
while $|grad| \geq \epsilon$ **do**
  $R \leftarrow diag(y_1(1-y_1), ..., y_2(1-y_2))$;
  $\mathbf{H} \leftarrow \mathbf{\Phi}^T R \mathbf{\Phi}$;
  $\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}^{-1}\mathbf{\Phi}^T(\mathbf{y} - \mathbf{t})$;
  $\mathbf{y} = (\sigma(w^T\phi_1), \sigma(w^T\phi_2), ..., \sigma(w^T\phi_n))^T$;
  $grad \leftarrow (\mathbf{y} - \mathbf{t})^T\mathbf{\Phi}$;
end
return $\mathbf{w}$

3 Convexity

To show that the error function is convex, we will show that the Hessian is positive semi-definite, i.e., $\forall \mathbf{v} \in \mathbf{R}^M, \mathbf{v}^T H \mathbf{v} \geq 0$. Let $\mathbf{v} \in \mathbf{R}^M$. Therefore:

$$\mathbf{v}^T H \mathbf{v} = \mathbf{v}^T \Phi^T R \Phi \mathbf{v} \tag{28}$$
$$= (\Phi \mathbf{v})^T R (\Phi \mathbf{v}) \tag{29}$$
$$= u^T R u \tag{30}$$

where $u = \Phi \mathbf{v}$. Further note that the diagonal elements of R are $y_n(1-y_n) > 0$, since y_n is a sigmoid function. Therefore,

$$u^T R u = \sum_{n=1}^{N} u_n^2 y_n (1-y_n) \tag{31}$$
$$\geq 0 \tag{32}$$

showing that H is positive semi-definite and that E is a convex function of $\mathbf{w}$.