Iterativley Reweighted Least Squares

Sebastian Pölsterl

May 2, 2018

1 Weighted Least Squares

A generalisation of least squares regression is *weighted least squares regression* where instead of minimising the residual sum of squares $\mathrm{RSS}(\beta_0, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2$ the function $\mathrm{WRSS}(\beta_0, \boldsymbol{\beta})$ is minimised:

$$\mathrm{WRSS}(\beta_0, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n} w_i (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^n$ is a vector of weights, one for each sample. Least squares regression can be expressed as a weighted least squares regression with $w_i = 1$ for all i. By minimising $\mathrm{WRSS}(\beta_0, \boldsymbol{\beta})$ one obtains estimates for the unknown parameters which is analogous to setting the first derivative to zero

$$\frac{\partial}{\partial \beta_j}\mathrm{WRSS}(\beta_0, \boldsymbol{\beta}) = -\sum_{i=1}^{n} w_i x_{ij} (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i). \tag{2}$$

Let $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_m)^\top$ be a vector of all unknown parameters (including the intercept β_0), and $\mathbf{X}$ a matrix of feature vectors of size $n \times (m+1)$ where the first column contains only ones to account for the intercept. Both $\mathrm{WRSS}(\beta_0, \boldsymbol{\beta})$ and its derivative can be expressed in matrix notation as

$$\mathrm{WRSS}(\beta_0, \boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \tag{3}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathrm{WRSS}(\beta_0, \boldsymbol{\beta}) = -\mathbf{X}^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}), \tag{4}$$

where the matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal elements $w_1, \ldots, w_n$. Finally, setting the derivative to zero one can obtain an estimate $\hat{\boldsymbol{\theta}}$ of the unknown parameters:

$$\begin{aligned}
\mathbf{X}^\top \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) &= \mathbf{0} \\
\mathbf{X}^\top \mathbf{W}\mathbf{y} - \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\theta} &= \mathbf{0} \\
\mathbf{X}^\top \mathbf{W}\mathbf{y} &= \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\theta} \\
\left(\mathbf{X}^\top \mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{y} &= \hat{\boldsymbol{\theta}}.
\end{aligned} \tag{5}$$

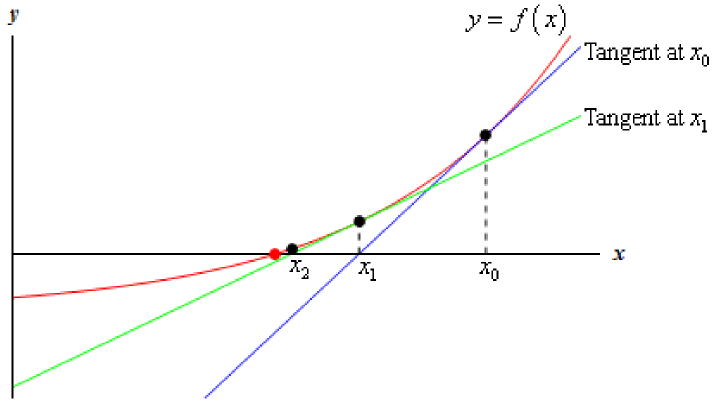

Figure 1: The blue line is the tangent to function f at x_0. We can see that this line will cross the x-axis closer to the actual solution than x_0 does. The point where the tangent crosses the x-axis is x_1 and is used as the new approximation to the solution [2].

2 Newton's method

Given a function $f(x)$ and its derivative $f'(x)$, Newton's method aims to find the value x that satisfies $f(x) = 0$ by iteratively approximating the solution by calculating

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{6}$$

Geometrically, the point $(x_{n+1}, 0)$ is the intersection with the x-axis of a line tangent to the function f at point $(x_n, f(x_n))$ (see figure 1). This process is repeated until convergence [1]. The first approximation x_0 has to be selected by the user.

3 Iteratively Reweighted Least Squares

Fitting a logistic regression model to the training data is usually accomplished by maximum likelihood, using the conditional likelihood of the class $y_i \in \{0; 1\}$ on the data $\mathbf{x}_i$ [3].

$$\begin{align} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \tag{7} \\ &= \sum_{i=1}^{n} y_i \boldsymbol{\theta}^\top \mathbf{x}_i - \log[1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i)], \tag{8} \end{align}$$

where $\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ denotes the logistic function applied to the linear model $\eta_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i$. Obtaining estimates by maximum likelihood refers to finding a solution for $\arg\max_{\beta_0, \boldsymbol{\beta}} l(\beta_0, \boldsymbol{\beta})$.

Maximisation follows the principal of least squares regression by setting the derivative of the log-likelihood to zero:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} y_i x_{ij} - \frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}_i) \cdot x_{ij}}{1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i)} \\
&= \sum_{i=1}^{n} y_i x_{ij} - \pi_i x_{ij} \\
&= \sum_{i=1}^{n} x_{ij}(y_i - \pi_i) = 0,
\end{aligned} \tag{9}
$$

where $j = 1, \ldots, m$. In contrast to least squares regression there is no closed form solution to solve equation (9).

As the goal is to find the root of the concave log-likelihood function, one can use Newton's methods for maximum likelihood estimation. Thus, $f(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and $f'(\boldsymbol{\theta})$ is given by the second partial derivative of the log-likelihood function

$$
\begin{aligned}
\frac{\partial^2}{\partial \theta_j \partial \theta_k} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} -x_{ij} x_{ik} \frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}_i)}{1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i)} + x_{ij} x_{ik} \frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}_i)^2}{(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i))^2} \\
&= \sum_{i=1}^{n} -x_{ij} x_{ik} \pi_i + x_{ij} x_{ik} \pi_i^2 \\
&= \sum_{i=1}^{n} -x_{ij} x_{ik} \pi_i (1 - \pi_i).
\end{aligned} \tag{10}
$$

The derivations above can be expressed in matrix notation as

$$
\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{X}^\top (\mathbf{y} - \mathbf{p}) \tag{11}
$$

$$
\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -\mathbf{X}^\top \mathbf{W} \mathbf{X} \tag{12}
$$

where $\mathbf{X}$ is a matrix of size $n \times (m+1)$, $\mathbf{p} = (\pi_1, \ldots, \pi_m)^\top$ the vector of fitted probabilities, and $\mathbf{W}$ a diagonal matrix of weights $w_i = \pi_i(1 - \pi_i)$ of size $n \times n$.

Substituting these values in the Newton update step from equation (6), one obtains

$$
\begin{aligned}
\boldsymbol{\theta}_{\text{new}} &= \boldsymbol{\theta}_{\text{old}} - \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right)^{-1} \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= \boldsymbol{\theta}_{\text{old}} + \left(\mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{p}) \\
&= \left(\mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{X} \boldsymbol{\theta}_{\text{old}} + \mathbf{W}^{-1} (\mathbf{y} - \mathbf{p})) \\
&= \left(\mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{z},
\end{aligned} \tag{13}
$$

where the right hand side of the equation is evaluated at $\boldsymbol{\theta}_{\text{old}}$. The update step in equation (13) is the same as the solution to a weighted least squares fit in equation (5) with the response vector $\mathbf{z} \in \mathbb{R}^n$ defined as

$$\begin{aligned}
\mathbf{z} &= \mathbf{X}\boldsymbol{\theta}_{\text{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p}) \\
z_i &= \boldsymbol{\theta}_{\text{old}}^{\top}\mathbf{x}_i + \frac{y_i - \pi_i}{\pi_i(1-\pi_i)} \\
&= \boldsymbol{\theta}_{\text{old}}^{\top}\mathbf{x}_i + \frac{y_i - \pi_i}{w_i}
\end{aligned} \tag{14}$$

These equations get solved repeatedly, since at each iteration the vector $\mathbf{p}$ changes, and hence does $\mathbf{W}$ and $\mathbf{z}$. This algorithm is referred to as *iteratively re-weighted least squares* (IRLS), since at each iteration it solves the weighted least squares problem from equation (3). The starting point for IRLS is usually chosen as $\boldsymbol{\theta} = \mathbf{0}$.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*, pages 484–496. Cambridge University Press, 2009.

[2] Paul Dawkins. Calculus I. http://tutorial.math.lamar.edu/Classes/CalcI/NewtonsMethod.aspx.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, pages 120–121. Springer, second edition, 2009.