Stockholms universitet

Gradients and Hessians for log-likelihood in logistic regression

Frank Miller, Department of Statistics

Spring 2021

Minimisation of negative log-likelihood

The maximum likelihood estimate (MLE) is the solution $\hat{\boldsymbol{\beta}}$ of $g(\hat{\boldsymbol{\beta}}) = \min g(\mathbf{b})$ with

$$g(\mathbf{b}) = -\text{log-likelihood}(\mathbf{b}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{n} \text{log-likelihood}(\mathbf{b}, \mathbf{x}_i, y_i),$$

where $\mathbf{x}_i$ is the vector of explanatory variables and y_i is the dependent variable for observation i. $\mathbf{X}$ is the design matrix having rows $\mathbf{x}_i^\top$ and $\mathbf{y}$ is the n-dimensional vector of dependent variables. Further, in machine learning it is common to scale the negative log-likelihood with factor $\frac{1}{n}$. Of course, this does nor change the minimisation problem.

For simple logistic regression with a single explanatory variable w_i, we have $\mathbf{x}_i = (1, w_i)^\top$.

Logistic regression

For logistic regression, $y_i \in \{0, 1\}$ and

$$p(\mathbf{x}_i) = P(Y = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \mathbf{b})} = \frac{\exp(\mathbf{x}_i^\top \mathbf{b})}{1 + \exp(\mathbf{x}_i^\top \mathbf{b})}.$$

The scaled negative log-likelihood for logistic regression is

$$\begin{aligned} g(\mathbf{b}) &= -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log\{p(\mathbf{x}_i)\} + (1 - y_i) \log\{1 - p(\mathbf{x}_i)\} \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[-\log\{1 - p(\mathbf{x}_i)\} - y_i \log\left\{ \frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)} \right\} \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\log\left\{ 1 + \exp(\mathbf{x}_i^\top \mathbf{b}) \right\} - \mathbf{x}_i^\top \mathbf{b} y_i \right]. \end{aligned}$$

The gradient is

$$\mathbf{g}'(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\exp(\mathbf{x}_i^\top \mathbf{b})}{1 + \exp(\mathbf{x}_i^\top \mathbf{b})} - y_i \right\} \mathbf{x}_i = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{1 + \exp(-\mathbf{x}_i^\top \mathbf{b})} - y_i \right\} \mathbf{x}_i = \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_i'(\mathbf{b})$$

with

$$\mathbf{g}_i'(\mathbf{b}) = \left\{ \frac{1}{1 + \exp(-\mathbf{x}_i^\top \mathbf{b})} - y_i \right\} \mathbf{x}_i.$$

The Hessian matrix of the scaled negative log-likelihood is then

$$\mathbf{g}''(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} p(\mathbf{x}_i)\{1 - p(\mathbf{x}_i)\}\mathbf{x}_i\mathbf{x}_i^\top.$$

(Note that instead of writing $\mathbf{g}'(\mathbf{b})$ for the gradient and $\mathbf{g}''(\mathbf{b})$ for the Hessian, the notation $\nabla g(\mathbf{b})$ and $\mathbf{H}(\mathbf{b})$ is often used in literature.)
For simple logistic regression with a single explanatory variable w_i,

$$\mathbf{x}_i\mathbf{x}_i^\top = \begin{pmatrix} 1 & w_i \\ w_i & w_i^2 \end{pmatrix}.$$