Newton's method for Logistic Regression

Wenchen Guo

Given n observations $(x_i; y_i), i = 1, ..., n, x_i \in R^p, y_i \in \{0, 1\}$, parameters $a \in R^p$ and $b \in R$.

The log-likelihood function for logistic regression is:

$$l(a,b) = \sum_{i=1}^{n} [y_i log(h(x_i;a,b)) + (1-y_i) log(1-h(x_i;a,b))]$$

Derive the Hessian Matrix.

$$\begin{aligned}
h(x_i;a,b) &= \frac{1}{1+exp(-a^T x_i - b)} \\
1 - h(x_i;a,b) &= 1 - \frac{1}{1+exp(-a^T x_i - b)} \\
&= \frac{exp(-a^T x_i - b)}{1+exp(-a^T x_i - b)}
\end{aligned}$$

Taking derivative of $h(x_i; a, b)$ with respect to a, b.

$$\begin{aligned}
\frac{\partial}{\partial a} h(x_i;a,b) &= -(1+exp(-a^T x_i - b))^{-2} * exp(-a^T x_i - b) * (-x_i) \\
&= x_i * \frac{1}{1+exp(-a^T x_i - b)} * \frac{exp(-a^T x_i - b)}{1+exp(-a^T x_i - b)} \\
&= x_i h_i (1-h_i) \\
\frac{\partial}{\partial b} h(x_i;a,b) &= -(1+exp(-a^T x_i - b))^{-2} * exp(-a^T x_i - b) * (-1) \\
&= \frac{1}{1+exp(-a^T x_i - b)} * \frac{exp(-a^T x_i - b)}{1+exp(-a^T x_i - b)} \\
&= h_i(1-h_i)
\end{aligned}$$

Thus, the derivative of the log-likelihood function for logistic regression is:

$$
\begin{aligned}
\frac{\partial}{\partial a} l(a,b) &= \sum_{i=1}^{n} \left[\frac{y_i x_i h_i (1-h_i)}{h_i} - \frac{(1-y_i)(x_i h_i (1-h_i))}{1-h_i} \right] \\
&= \sum_{i=1}^{n} [y_i x_i (1-h_i) - (1-y_i)(x_i h_i)] \\
&= \sum_{i=1}^{n} [x_i (y_i - h_i)] \\
\frac{\partial}{\partial b} l(a,b) &= \sum_{i=1}^{n} \left[\frac{y_i h_i (1-h_i)}{h_i} - \frac{(1-y_i)(h_i (1-h_i))}{1-h_i} \right] \\
&= \sum_{i=1}^{n} [y_i (1-h_i) - (1-y_i) h_i] \\
&= \sum_{i=1}^{n} [y_i - h_i]
\end{aligned}
$$

Thus, the gradient matrix is

$$
\nabla l(a,b) = \begin{bmatrix} \frac{\partial}{\partial a} l(a,b) \\ \frac{\partial}{\partial b} l(a,b) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i (y_i - h_i) \\ \sum_{i=1}^{n} (y_i - h_i) \end{bmatrix}
$$

Similarly, we can get the Hessian matrix H

$$
\begin{aligned}
\mathbf{H}_{11} &= \frac{\partial^2 l(a,b)}{\partial a \partial a} = \sum_{i=1}^{n} \left(-x_i^2 h_i (1-h_i) \right) \\
\mathbf{H}_{12} &= \frac{\partial^2 l(a,b)}{\partial a \partial b} = \sum_{i=1}^{n} \left(-x_i h_i (1-h_i) \right) \\
\mathbf{H}_{22} &= \frac{\partial^2 l(a,b)}{\partial b \partial b} = \sum_{i=1}^{n} \left(-h_i (1-h_i) \right)
\end{aligned}
$$

$$
\begin{aligned}
H = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12} & \mathbf{H}_{22} \end{bmatrix} &= \begin{bmatrix} \sum_{i=1}^{n} \left(-x_i^2 h_i (1-h_i) \right) & \sum_{i=1}^{n} \left(-x_i h_i (1-h_i) \right) \\ \sum_{i=1}^{n} \left(-x_i h_i (1-h_i) \right) & \sum_{i=1}^{n} \left(-h_i (1-h_i) \right) \end{bmatrix} \\
&= -\sum_{i=1}^{n} h_i (1-h_i) \begin{bmatrix} x_i^2 & x_i \\ x_i & 1 \end{bmatrix} \\
&= -\sum_{i=1}^{n} h_i (1-h_i) \begin{bmatrix} x_i \\ 1 \end{bmatrix} \begin{bmatrix} x_i & 1 \end{bmatrix}
\end{aligned}
$$

M is negative semi-definite $\iff$ $v^{\mathrm{T}} M v \leq 0$ for all v

$$
\begin{aligned}
v^T H v &= v^T \left(-\sum_{i=1}^{n} h_i (1-h_i) \begin{bmatrix} x_i \\ 1 \end{bmatrix} \begin{bmatrix} x_i & 1 \end{bmatrix} \right) v \\
&= -\sum_{i=1}^{n} h_i (1-h_i) \left(v^T \begin{bmatrix} x_i \\ 1 \end{bmatrix} \begin{bmatrix} x_i & 1 \end{bmatrix} v \right) \\
&= -\sum_{i=1}^{n} h_i (1-h_i) \left(\begin{bmatrix} x_i & 1 \end{bmatrix} v \right)^2
\end{aligned}
$$

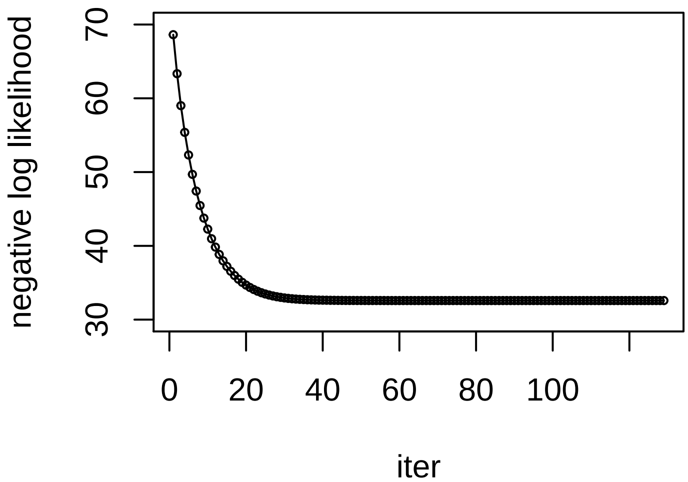
Since $0 \leq h \leq 1$,we have $h_i(1-h_i) \geq 0$
Thus, $v^T Hv \leq 0$ for all v
Thus H is negative semi-definite. this implies that $l(a, b)$ is concave and has no local maximum other than the global one.

Implement Newton's Method.

Use data logit-x.dat and logit-y.dat, where logit-x $\in R^2$, and logit-y $\in \{0, 1\}$.
Implement Newton's method for optimizing $l(a, b)$ and apply it to fit a logistic regression model to the data.

Initialize Newton's method with $a = 0, b = 0$. Set the threshold to be 10^{-10} and step-size = 0.1.

The log-likelihood function versus iterations plot:



Newton's method runs 130 iterations and the log likelihood function value converge to 32.5856.
The coefficients are $a = (0.7604, 1.1719)$ and $b = -2.6205$.

theta1	theta2	Intercept
0.7604	1.1719	-2.6205