Math!

Jimmy Kelliher

2022-04-01

2.2 Newton-Raphson Algorithm

The Newton-Raphson method is an efficient algorithm for computing the maximum likelihood estimator (MLE) under certain conditions. However, to even consider computing an MLE, we need to choose a reasonable model for our data so that we can construct our likelihood function.

Toward that aim, let $\mathbf{X} \equiv (x_{ij})$ denote the $n \times p$ matrix corresponding to our $n = 569$ observations and our $p = 21$ features (including an intercept term). Because our outcome $\mathbf{Y} \equiv (y_1, \ldots, y_n)^t$ is binary, it is natural to model $y_i \,|\, x_{i1}, \ldots, x_{ip} \sim \text{Bernoulli}(\pi_i)$, where

$$\begin{aligned}
\pi_i &\equiv P(y_i = 1 \,|\, x_{i1}, \ldots, x_{ip}) \\
&= \left(1 + \exp\left(-\sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^{-1}, \quad \text{such that} \\
&\log\left(\frac{\pi_i}{1 - \pi_i} \right) = \sum_{j=1}^{p} \beta_j x_{ij}
\end{aligned}$$

for a vector of parameters $\boldsymbol{\beta} \in \mathbb{R}^p$. The likelihood function of $\boldsymbol{\beta}$ given our data is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta} \,|\, \mathbf{X}, \mathbf{Y}) &\equiv f(\mathbf{X}, \mathbf{Y} \,|\, \boldsymbol{\beta}) \\
&= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \\
&= \prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i).
\end{aligned}$$

To employ the Newton-Raphson procedure, we need to compute the gradient vector and hessian matrix corresponding to $\mathcal{L}$. However, as computing derivatives of the likelihood function is tedious, we instead consider the log-likelihood function

$$\begin{aligned}
l(\boldsymbol{\beta} \,|\, \mathbf{X}, \mathbf{Y}) &\equiv \log \mathcal{L}(\boldsymbol{\beta} \,|\, \mathbf{X}, \mathbf{Y}) \\
&= \log \left(\prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i) \right) \\
&= \sum_{i=1}^{n} \left(y_i \log\left(\frac{\pi_i}{1 - \pi_i} \right) - \log\left(\frac{1}{1 - \pi_i} \right) \right) \\
&= \sum_{i=1}^{n} \left(y_i \sum_{j=1}^{p} \beta_j x_{ij} - \log\left(1 + \exp\left(\sum_{j=1}^{p} \beta_j x_{ij} \right) \right) \right).
\end{aligned}$$

Toward computing the derivative of the l, we define

$$\eta_i \equiv \sum_{j=1}^{p} \beta_j x_{ij}$$

for each $i \in \{1, \ldots, n\}$. Observe that for each $k \in \{1, \ldots, p\}$,

$$\frac{\partial \eta_i}{\partial \beta_k} = x_{ik} \quad \text{and} \quad \frac{\partial \pi_i}{\partial \beta_k} = x_{ik}\pi_i(1 - \pi_i).$$

Thus, it follows that

$$\begin{aligned}
\frac{\partial l}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \left(\sum_{i=1}^{n} \left(y_i \eta_i - \log\left(1 + e^{\eta_i}\right) \right) \right) \\
&= \sum_{i=1}^{n} \left(y_i \frac{\partial \eta_i}{\partial \beta_k} - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \frac{\partial \eta_i}{\partial \beta_k} \right) \\
&= \sum_{i=1}^{n} x_{ik}(y_i - \pi_i), \quad \text{and}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l^2}{\partial \beta_k \beta_l} &= \frac{\partial}{\partial \beta_l} \left(\sum_{i=1}^{n} x_{ik}(y_i - \pi_i) \right) \\
&= -\sum_{i=1}^{n} x_{ik} \frac{\partial \pi_i}{\partial \beta_l} \\
&= -\sum_{i=1}^{n} x_{ik} x_{il} \pi_i (1 - \pi_i).
\end{aligned}$$

for $k, l \in \{1, \ldots p\}$. The above expressions completely characterize the gradient vector and hessian matrix corresponding to the log-likelihood l. However, it will be convenient to express these objects compactly via matrix notation. Toward this aim, we define $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_n)^t$ to be a vector of probabilities and

$$\mathbf{W} \equiv \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}, \quad \text{where} \quad w_i \equiv \pi_i(1 - \pi_i)$$

for each $i \in \{1, \ldots, n\}$. We can think of $\mathbf{W}$ as a pseudo-weight matrix, noting that its entries do not generally sum to unity. Given this notation, the gradient vector and hessian matrix corresponding to l are given by

$$\nabla l(\boldsymbol{\beta} \,|\, \mathbf{X}, \mathbf{Y}) = \mathbf{X}^t(\mathbf{Y} - \boldsymbol{\pi}) \quad \text{and} \quad \nabla^2 l(\boldsymbol{\beta} \,|\, \mathbf{X}, \mathbf{Y}) = -\mathbf{X}^t \mathbf{W} \mathbf{X},$$

respectively. Our Newton-Raphson algorithm is then characterized by the procedure

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t (\mathbf{Y} - \boldsymbol{\pi}).$$

To assess if this updating procedure is well-behaved, we must investigate the properties of the hessian matrix, thereby motivating the following proposition.

Proposition 1. The hessian matrix of l is negative semi-definite. If $\pi_i \in (0, 1)$ for at least one $i \in \{1, \ldots, n\}$, and if $\mathbf{X}$ is of full rank, then the hessian matrix is negative definite.

Proof. Because $\mathbf{W}$ is a diagonal matrix with non-negative elements, it is positive semi-definite. Thus, for any $u \in \mathbb{R}^p \setminus \{\mathbf{0}\}$, it follows that $\mathbf{X}u \in \mathbb{R}^p$, and hence

$$\begin{aligned} u^t \nabla^2 l(\boldsymbol{\beta} \,|\, \mathbf{X}, \mathbf{Y}) u &= -u^t \mathbf{X}^t \mathbf{W} \mathbf{X} u \\ &= -(\mathbf{X}u)^t \mathbf{W} (\mathbf{X}u) && \text{(by rules for transpose of a product)} \\ &\leq 0. && \text{(by positive semi-definiteness of } \mathbf{W}) \end{aligned}$$

That is, the hessian matrix is negative semi-definite. If we further have that $\mathbf{X}$ is of full rank, then $\mathbf{X}u \in \mathbb{R}^p \setminus \{\mathbf{0}\}$. Moreover, if $\pi_i \in (0,1)$ for at least one $i \in \{1, \ldots, n\}$, then $\mathbf{W}$ is positive definite. Together, these facts make the above inequality strict, such that the hessian matrix is negative definite. □

The above result provides us with two key insights: (1) if our data matrix is not of full rank, the hessian matrix might not be negative definite; and (2) if our fitted probabilities are all either zero or one, then our hessian matrix will not be negative definite. By culling highly correlated covariates during the exploratory data analysis phase, we have precluded scenario (1) (and indeed, we find that the Newton-Raphson algorithm does not converge if all 30 covariates are considered). To prevent scenario (2), we seed our initial guess at $\boldsymbol{\beta} = \mathbf{0}$, which corresponds to $\pi_i = \frac{1}{2}$ for all $i \in \{1, \ldots, n\}$. As added precautions, we further control for ascent direction and allow step-halving, though this is not critical for such a well-behaved optimization problem.

2.3 Logistic LASSO Algorithm

The logistic LASSO algorithm is characterized by a constrained optimization problem that is not everywhere differentiable. As such, we consider an approximation of the log-likelihood function l and a coordinate descent procedure to make the problem more tractable.

To begin, we consider the function $g : \mathbb{R}^p \to \mathbb{R}$ given by

$$g(\boldsymbol{\beta}) = -\frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \eta(\boldsymbol{\alpha}),$$

which is *proportional* to the Taylor expansion of l centered around $\boldsymbol{\alpha} \in \mathbb{R}^p$, where

$$\begin{aligned} z_i &\equiv \sum_{j=1}^{p} \alpha_j x_{ij} + \frac{y_i - \pi_i}{w_i}, && \text{(effective response)} \\ w_i &\equiv \pi_i (1 - \pi_i), && \text{(effective weights)} \\ \pi_i &\equiv \left(1 + \exp\left(-\sum_{j=1}^{p} \alpha_j x_{ij} \right) \right)^{-1}, \end{aligned}$$

and some function $\eta : \mathbb{R}^p \to \mathbb{R}$ that does not depend on $\boldsymbol{\beta} \in \mathbb{R}^p$. The details of this derivation can be found in the mathematical appendix, but the result follows largely from the generalized version of Taylor's theorem applied to the gradient vector and hessian matrix computed in the previous section.

With our quadratic approximation g of l in tow, we consider the minimization problem

$$\underset{\beta_k \in \mathbb{R}}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

This penalized optimization problem characterizes the updating procedure of our coordinate descent algorithm. However, as alleged earlier, our objective function is not differentiable at the origin. To find our desired

minimizer, we consider a series of definitions and lemmas to reduce the complexity of the problem, as inspired by Friedman et al. (2007).

Definition 1. Let $S : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ be given by

$$S(\beta, \gamma) = \operatorname{sgn}(\beta) \max \{|\beta| - \gamma, 0\}.$$

We call S the *shrinkage function* or the *soft-thresholding operator*. Observe that the map $\beta \mapsto S(\beta, 0)$ is simply the identity function. Thus, we can think of $\gamma \geq 0$ as an additive penalty that shrinks β in magnitude toward zero.

Lemma 1. Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \frac{1}{2}(x - \beta)^2 + \gamma|x|$$

for $\beta \in \mathbb{R}$ and $\gamma \in \mathbb{R}_+$. It follows that

$$\arg\min_{x \in \mathbb{R}} f(x) = S(\beta, \gamma),$$

where S is the soft-thresholding operator.

Proof. Because f is a sum of the convex functions $\frac{1}{2}(x - \beta)^2$ and $\gamma|x|$, it is also a convex function. Thus, the minimizer $x^* \in \mathbb{R}$ of f is unique. While f is not differentiable at $x = 0$, it is the case that for all $x \neq 0$,

$$\frac{df}{dx} = x - \beta + \gamma \operatorname{sgn}(x).$$

Including our boundary case at $x = 0$, we obtain three critical points that serve as candidates for our unique minimum. Namely, $x^* \in \{0, \beta - \gamma, \beta + \gamma\}$, which satisfy

$$\begin{aligned} f(0) &= \frac{1}{2}\beta^2, \\ f(\beta - \gamma) &= \frac{1}{2}\gamma^2 + \gamma|\beta - \gamma|, \quad \text{and} \\ f(\beta + \gamma) &= \frac{1}{2}\gamma^2 + \gamma|\beta + \gamma|. \end{aligned}$$

Now, because $\gamma \geq 0$, it follows that $f(\beta - \gamma) < f(\beta + \gamma)$ if and only if $|\beta - \gamma| < |\beta + \gamma|$ if and only if $\beta > 0$. Moreover, if $\beta > \gamma$, then $|\beta - \gamma| = \beta - \gamma$, and hence

$$\begin{aligned} f(\beta - \gamma) - f(0) &= \frac{1}{2}\gamma^2 + \gamma|\beta - \gamma| - \frac{1}{2}\beta^2 \\ &= \gamma(\beta - \gamma) - \frac{1}{2}(\beta + \gamma)(\beta - \gamma) \\ &= -\frac{1}{2}(\beta - \gamma)^2 \\ &< 0. \end{aligned}$$

That is, $f(\beta - \gamma) < f(0)$ if and only if $\beta > \gamma$. By analogous argument, we find that $f(\beta + \gamma) < f(0)$ if and only if $\beta < -\gamma$. Combining all of these results, we obtain

$$x^* = \begin{cases} \beta - \gamma & \text{if} \quad \beta > \gamma, \\ \beta + \gamma & \text{if} \quad \beta < -\gamma, \\ 0 & \text{otherwise.} \end{cases}$$

Of course, the above expression is precisely $x^* = S(\beta, \gamma)$. $\square$

Lemma 2. Consider the optimization problem

$$\min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

for some $k \in \{1, \ldots, p\}$. It follows that the minimizer is given by

$$\hat{\beta}_k = \left(\sum_{i=1}^{n} w_i x_{ik}^2 \right)^{-1} \sum_{i=1}^{n} w_i x_{ik} \left(z_i - \sum_{j \neq k} \beta_j x_{ij} \right).$$

Proof. We first remark that minimizing our objective function is equivalent to maximizing our quadratic approximation g, as the two functions differ by a constant (with respect to β_k) and in signature. For convenience, we refine earlier notation via

$$\eta_i \equiv \sum_{j=1}^{p} \beta_j x_{ij} \quad \text{and} \quad \eta_i^{(k)} \equiv \eta_i - \beta_k x_{ik}$$

for each $i \in \{1, \ldots, n\}$. Note that $\eta_i^{(k)}$ does not depend on β_k. Because $-g$ is convex with respect β_k, it has a unique minimizer $\hat{\beta}_k$ satisfying

$$\begin{aligned}
0 &= -n \frac{\partial g}{\partial \beta_k} \\
&= \frac{\partial}{\partial \beta_k} \left(\frac{1}{2} \sum_{i=1}^{n} w_i \Big(z_i - \eta_i \Big)^2 \right) \\
&= \sum_{i=1}^{n} w_i \Big(\eta_i - z_i \Big) \frac{\partial \eta_i}{\partial \beta_k} \\
&= \sum_{i=1}^{n} w_i x_{ik} \Big(\hat{\beta}_k x_{ik} + \eta_i^{(k)} - z_i \Big) \\
&= \left(\sum_{i=1}^{n} w_i x_{ik}^2 \right) \hat{\beta}_k - \sum_{i=1}^{n} w_i x_{ik} \left(z_i - \eta_i^{(k)} \right).
\end{aligned}$$

Rearranging the above expression gives the desired result. $\square$

Lemma 3. With $\hat{\beta}_k$ defined as above,

$$\min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} = \min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2} (\beta_k - \hat{\beta}_k)^2 + \lambda_k |\beta_k| \right\}, \quad \text{where}$$

$$\lambda_k \equiv \left(\frac{1}{n} \sum_{i=1}^{n} w_i x_{ik}^2 \right)^{-1} \lambda.$$

Proof. For convenience, we consider the following definition: for any functions $h, k : \mathbb{R}^p \to \mathbb{R}$, we say that $h(\boldsymbol{\beta}) \simeq k(\boldsymbol{\beta})$ if and only if the difference $h(\boldsymbol{\beta}) - k(\boldsymbol{\beta})$ does not depend on β_k. That is, h and k define the

same optimization problem with respect to β_k. Observe that

$$\begin{aligned}
\sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 &= \sum_{i=1}^{n} w_i \Big(z_i - \eta_i \Big)^2 \\
&\simeq \sum_{i=1}^{n} w_i \left(\Big(\beta_k x_{ik} + \eta_i^{(k)} \Big)^2 - 2 z_i \Big(\beta_k x_{ik} + \eta_i^{(k)} \Big) \right) \\
&\simeq \sum_{i=1}^{n} w_i \left(\beta_k^2 x_{ik}^2 + 2 \beta_k x_{ik} \eta_i^{(k)} - 2 z_i \beta_k x_{ik} \right) \\
&= \left(\sum_{i=1}^{n} w_i x_{ik}^2 \right) \beta_k^2 - 2 \left(\sum_{i=1}^{n} w_i x_{ik} \left(z_i - \eta_i^{(k)} \right) \right) \beta_k \\
&= \left(\sum_{i=1}^{n} w_i x_{ik}^2 \right) \left(\beta_k^2 - 2 \hat{\beta}_k \beta_k \right) \\
&\simeq \left(\sum_{i=1}^{n} w_i x_{ik}^2 \right) \Big(\beta_k - \hat{\beta}_k \Big)^2 .
\end{aligned}$$

Similarly, we further obtain

$$\lambda \sum_{j=1}^{p} |\beta_j| \simeq \lambda |\beta_k| .$$

Thus, our optimization problem can be rewritten as

$$\begin{aligned}
&\min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\
&= \min_{\beta_k \in \mathbb{R}} \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} w_i x_{ik}^2 \right) \frac{1}{2} \Big(\beta_k - \hat{\beta}_k \Big)^2 + \lambda |\beta_k| \right\} .
\end{aligned}$$

Because x_{ik} is a continuous random variable in our data, if not all fitted probabilities are zero or one (a scenario we again mitigate), then $\frac{1}{n} \sum_{i=1}^{n} w_i x_{ik}^2 > 0$ almost surely. As such, dividing through by this term establishes the desired result. □

Proposition 2. With $\hat{\beta}_k$ and λ_k defined as above,

$$\arg\min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} = S \left(\hat{\beta}_k, \lambda_k \right) .$$

Proof. By Lemma 3, our minimization problem is equivalent to that of

$$\min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2} (\beta_k - \hat{\beta}_k)^2 + \lambda_k |\beta_k| \right\} .$$

By Lemma 1, the solution to this minimization problem is uniquely $S \left(\hat{\beta}_k, \lambda_k \right)$. □