Lecture 16
Solving GLMs via IRWLS

09 November 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Notes

- problem set 5 posted; due next class

Notes

- problem set 5 posted; due next class
- problem set 6, November 18th

Goals for today

- fixed PCA example from last time
- how to solve logistic regression via weighted least squares
- classification problem on image corpus

fixed PCA example from last time

solving logistic regression

GLMs

Recall that we define generalized linear models such that the mean of y is some function of $X\beta$, rather than directly equal to it:

$$\mathbb{E}(y|X) = g^{-1}(X\beta)$$

With g, called the *link function*, equal to some fixed and known function.

Logistic regression

The logistic regression function uses g equal to the logit function to describe a distribution with $y \in \{0, 1\}$. Specifically, we have the following description of the statistical model:

$$\begin{aligned} \mathbb{E}(y|X) &= \text{logit}^{-1}\left(X\beta\right) \\ &= \frac{1}{1 + e^{-X\beta}} \end{aligned}$$

Now that we have discussed how to solve the ordinary least squares equation, you may wonder how we would go about solving the logistic regression problem.

Recall that the likelihood of the data point y_i given its mean p_i is equal to:

$$L_i(y_i|p_i) = \exp\left\{ y_i \cdot \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right\}$$

There are two terms here that depend on p_i in different ways:

$$L_i(y_i|p_i) = \exp\left\{ y_i \cdot \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right\}$$

There are two terms here that depend on p_i in different ways:

$$L_i(y_i|p_i) = \exp\left\{ y_i \cdot \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right\}$$

The first term is simply equal to the projection of the regression vector $x_i^t\beta$.

There are two terms here that depend on p_i in different ways:

$$L_i(y_i|p_i) = \exp\left\{ y_i \cdot \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right\}$$

The first term is simply equal to the projection of the regression vector $x_i^t\beta$.

We can write the second term using:

$$\begin{aligned} p_i &= \frac{1}{1+e^{-x_i^t\beta}} \\ 1-p_i &= \frac{e^{-x_i^t\beta}}{1+e^{-x_i^t\beta}} \\ &= \frac{1}{1+e^{x_i^t\beta}} \\ \log(1-p_i) &= -1 \cdot \log(1+e^{x_i^t\beta}) \end{aligned}$$

We can then write the log-likelihood of the full model, in vector form, as:

$$\begin{aligned} l(y_i|x_i,\beta) &= \sum_i \left(y_i \cdot x_i^t\beta - \log(1+^{x_i^t\beta}) \right) \\ &= y^t X\beta - \log(1+e^{X\beta}) \end{aligned}$$

To find the MLE estimator, we want to find zeros of the derivative of the log-likelihood. The derivative is given by:

$$\begin{aligned}\nabla_\beta l(y_i|x_i,\beta) &= \nabla y^t X\beta - \nabla \log(1+e^{X\beta}) \\ &= X^t y - \nabla \log(1+e^{X\beta})\end{aligned}$$

To find the MLE estimator, we want to find zeros of the derivative of the log-likelihood. The derivative is given by:

$$\begin{aligned}\nabla_\beta l(y_i|x_i,\beta) &= \nabla y^t X\beta - \nabla \log(1+e^{X\beta}) \\ &= X^t y - \nabla \log(1+e^{X\beta})\end{aligned}$$

The second term can be calculated by writing the gradient out component wise.

In other words, we see from a simple application of the chain rule that:

$$\frac{\partial}{\partial \beta_k} \log(1 + e^{X\beta}) = \frac{1}{1 + e^{X\beta}} \cdot x_k^t \cdot e^{X\beta}$$

In other words, we see from a simple application of the chain rule that:

$$\frac{\partial}{\partial \beta_k} \log(1 + e^{X\beta}) = \frac{1}{1 + e^{X\beta}} \cdot x_k^t \cdot e^{X\beta}$$

Which can be simplified in terms of p:

$$\frac{1}{1 + e^{X\beta}} \cdot x_k^t \cdot e^{X\beta} = x_k^t p$$

In other words, we see from a simple application of the chain rule that:

$$\frac{\partial}{\partial \beta_k} \log(1 + e^{X\beta}) = \frac{1}{1 + e^{X\beta}} \cdot x_k^t \cdot e^{X\beta}$$

Which can be simplified in terms of p:

$$\frac{1}{1 + e^{X\beta}} \cdot x_k^t \cdot e^{X\beta} = x_k^t p$$

In vector form gives the gradient as:

$$\nabla \log(1 + e^{X\beta}) = X^t p$$

So now we have an explicit form of the gradient of the log-likelihood:

$$\begin{aligned}\nabla_\beta l(y_i|x_i,\beta) &= X^t y - \nabla \log(1+e^{X\beta}) \\ &= X^t y - X^t p \\ &= X^t(y-p)\end{aligned}$$

So now we have an explicit form of the gradient of the log-likelihood:

$$\begin{aligned}\nabla_\beta l(y_i|x_i,\beta) &= X^t y - \nabla \log(1+e^{X\beta}) \\ &= X^t y - X^t p \\ &= X^t(y-p)\end{aligned}$$

We cannot simply set this to zero because p is a non-linear function of X. However it does tell us that the residual in logistic regression should be uncorrelated with the X matrix (just as was the case with linear regression).

To find a numerical solution to the zeros of the function $\nabla_\beta l(y_i|x_i, \beta)$, the Newton–Raphson method can be used.

To find a numerical solution to the zeros of the function $\nabla_\beta l(y_i|x_i, \beta)$, the Newton–Raphson method can be used.

The technique is best described in a series of illustractions.

> *http://euler.stat.yale.edu/~tba3/stat612/lectures/lec16/img/NewtonIteration_Ani.gif*

We saw that the scalar version of using Newton–Raphson to determine the optimal values of f is given by (remember the derivatives are one more than the illustration because we want critical points rather than zeros of f):

$$\beta^{(k+1)} = \beta^{(k)} - \frac{f'(\beta^{(k)})}{f''(\beta^{(k)})}$$

We saw that the scalar version of using Newton–Raphson to determine the optimal values of f is given by (remember the derivatives are one more than the illustration because we want critical points rather than zeros of f):

$$\beta^{(k+1)} = \beta^{(k)} - \frac{f'(\beta^{(k)})}{f''(\beta^{(k)})}$$

The multivariate version of Newton–Raphson applies the following set of updates:

$$\beta^{(k+1)} = \beta^{(k)} - H^{-1}(\beta^{(k)})\nabla f(\beta^{(k)})$$

Where H is the Hessian matrix of all second order partial derivatives.

What is the Hessian of the log-likelihood? The gradient was $X^t(y - p)$, so we can quickly see that the Hessian only depends on the term $X^t p$.

In componentwise form, we see that:

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} &= \frac{\partial}{\partial \beta_j} \sum_i x_{i,k} \cdot p_i \\ &= \sum_i x_{i,k} \cdot \frac{\partial p_i}{\partial \beta_j}\end{aligned}$$

The partial derivative of p_i with respect to β_j is given by the following (it is just calculus with a clever grouping of terms at the end):

$$\begin{aligned}
\frac{\partial p_i}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j} \frac{1}{1 + e^{-x_i^t \beta}} \\
&= \frac{1}{(1 + e^{-x_i^t \beta})^2} \cdot -1 \cdot x_{i,j} \cdot e^{-x_i^t \beta} \\
&= -x_{i,j} \cdot \left(\frac{1}{1 + e^{-x_i^t \beta}} \right) \cdot \left(\frac{e^{-x_i^t \beta}}{1 + e^{-x_i^t \beta}} \right) \\
&= -x_{i,j} \cdot p_i \cdot (1 - p_i) \\
&= -x_{i,j} \cdot \mathrm{Var}(y_i)
\end{aligned}$$

The partial derivative of p_i with respect to β_j is given by the following (it is just calculus with a clever grouping of terms at the end):

$$\begin{aligned}
\frac{\partial p_i}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j} \frac{1}{1 + e^{-x_i^t \beta}} \\
&= \frac{1}{(1 + e^{-x_i^t \beta})^2} \cdot -1 \cdot x_{i,j} \cdot e^{-x_i^t \beta} \\
&= -x_{i,j} \cdot \left(\frac{1}{1 + e^{-x_i^t \beta}} \right) \cdot \left(\frac{e^{-x_i^t \beta}}{1 + e^{-x_i^t \beta}} \right) \\
&= -x_{i,j} \cdot p_i \cdot (1 - p_i) \\
&= -x_{i,j} \cdot \text{Var}(y_i)
\end{aligned}$$

We don't actually need the final line, but include it as it helps to give some intutition to the next set of steps.

This now yields the Hessian as:

$$Hl(\beta|x_i, \beta) = -X^t DX$$

Where D is an n-by-n diagonal matrix with components equal to $p_i \cdot (1 - p_i)$.

And therefore the Newton-Raphson step is given by:

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} - H^{-1}(\beta^{(k)})\nabla f(\beta^{(k)}) \\ &= \beta^{(k)} + (X^t D^{(k)} X)^{-1} X^t (y - p^{(k)})\end{aligned}$$

And therefore the Newton-Raphson step is given by:

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} - H^{-1}(\beta^{(k)})\nabla f(\beta^{(k)}) \\ &= \beta^{(k)} + (X^t D^{(k)} X)^{-1} X^t (y - p^{(k)})\end{aligned}$$

Look familiar?

If we let $S = D^{1/2}$, the standard deviations of y, and set

$$X' = SX$$
$$z = S^{-1}(y - p)$$

If we let $S = D^{1/2}$, the standard deviations of y, and set

$$\begin{aligned} X' &= SX \\ z &= S^{-1}(y - p) \end{aligned}$$

The least squares regression of z on X' gives:

$$\begin{aligned} (X'^t X')^{-1} X'^t z &= (X^t S^2 X)^{-1} X^t S S^{-1}(y - p) \\ &= (X^t D X)^{-1} X^t (y - p) \end{aligned}$$

So solving the logistic regression amounts to solving a sequence of weighted least squares models.

Notice that the scheme weights observations more if they have predicted probabilities close to 0 or 1.

Notice that the scheme weights observations more if they have predicted probabilities close to 0 or 1. Does this make sense based on our other lecture?

Notice that the scheme weights observations more if they have predicted probabilities close to 0 or 1. Does this make sense based on our other lecture?

Another interpretation is that it put the highest weight on points with the lowest variance.

This method generalizes to generalized linear models with exponential families, and has some fairly deep connections to Fischer information.

Why is this important?

Why is this important?

1. Explains many of the results on problem set 4; in particular when linear and logistic regression can be expected to give similar predictions and when they don't.

Why is this important?

1. Explains many of the results on problem set 4; in particular when linear and logistic regression can be expected to give similar predictions and when they don't.
2. Helps explain and address convergence properties in generalized linear models.

Why is this important?

1. Explains many of the results on problem set 4; in particular when linear and logistic regression can be expected to give similar predictions and when they don't.
2. Helps explain and address convergence properties in generalized linear models.
3. The logistic regression version of the normal equations can be used to establish large sample theory convergence results.

Why is this important?

1. Explains many of the results on problem set 4; in particular when linear and logistic regression can be expected to give similar predictions and when they don't.
2. Helps explain and address convergence properties in generalized linear models.
3. The logistic regression version of the normal equations can be used to establish large sample theory convergence results.
4. Gives us an idea of how the geometric and analytic concepts underlying ridge regression, principal component analysis, and (later) lasso regression connect to generalized linear models.