The Classification Problem, Machine learning demonstrated

Written by Brian Lesko, this this is a hand written demonstration of Statistical Machine Learning theories largely originating from the book, *An Introduction to Statistical Learning*, by Gareth James.

In part for the course offering at Ohio State University, Statistics and Machine Learning 6500, Problem set 2

2/14/23

For The Logistic regression model (with one feature, X ($P=1$) Predictor)

$$\log \frac{P(X)}{1-P(X)} = \beta_0 + \beta_1 X$$

the maximum likelyhood yields the coefficients β_0 and β_1 by

where the likelyhood function measures the likelyhood of observing a particular set of data

where $\ell(\beta_0, \beta_1)$

$$= \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 X_i) - \log(1 + e^{\beta_0 + \beta_1 X_i}) \right]$$

note: with one class assigned a probability of success, this can be called a bernoulli trial, which results in a binomial distribution (also called bernoulli)

maximizing the likelihood function requires the gradient and hessian of the likelyhood function, $\ell(\beta_0, \beta_1)$

the hessian is a matrix of form

$$\begin{bmatrix} \dfrac{\partial^2 \ell}{\partial \beta_0^2} & \dfrac{\partial^2 \ell}{(\partial \beta_0)(\partial \beta_1)} \\[2ex] \dfrac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \dfrac{\partial^2 \ell}{\partial \beta_1^2} \end{bmatrix}$$

the gradient has form

$$\begin{bmatrix} \dfrac{\partial \ell}{\partial \beta_0} \\[2ex] \dfrac{\partial \ell}{\partial \beta_1} \end{bmatrix} = \begin{matrix} \displaystyle\sum_{i=1}^{n} y_i - \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x}} \\[2ex] \displaystyle\sum_{i=1}^{n} x_i \left(y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \end{matrix}$$

$\hookrightarrow$ the gradient can be used to find the parameters numerically with a gradient descent algorithm

however, finding the hessian reveals the curvature of the surface revealing if the estimate is a maximum or minimum as well as the stability

$\frown$ $+$ not stable $\quad$ $\smile$ $-$ stable

the hessian can also be used for estimating the standard errors of the coefficient estimates and the efficiency of the solution fisher info matrix and Cramer-Rao lower bound

a) The hessian

$$\begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{(\partial \beta_0)(\partial \beta_1)} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1^2} \end{bmatrix} = \begin{matrix} -\sum_{i=1}^{n} \frac{1}{\beta_0^2} & -\sum_{i=1}^{n} \frac{1}{\beta_0^2} \\ -\sum_{i=1}^{n} \frac{x_i}{\beta_1^2} & -\sum_{i=1}^{n} \frac{x_i}{\beta_1^2} \end{matrix}$$

from below

$$\frac{\partial^2 \ell}{\partial \beta_0^2} = \frac{-\partial}{\partial \beta_0}\left(\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x}}\right)$$

quotient rule

$$= -\left[\left(1+e^{\beta_0+\beta_1 x}\right) - \left(e^{\beta_0+\beta_1 x}\right)\right] \Big/ \beta_0^2$$

$\leftarrow$ sum left out

$$= -\sum_{i=1}^{n} 1/\beta_0^2$$

$$\frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} = \frac{\partial}{\partial \beta_1}\left(\frac{\partial \ell}{\partial \beta_0}\right) = \frac{-\partial}{\partial \beta_1}\left(\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x}}\right)$$

$$= -\left[x\left(1+e^{\beta_0+\beta_1 x}\right) - x\, e^{\beta_0+\beta_1 x}\right] / \beta_1^2$$

$$= \sum_{i=1}^{n} -x_i / \beta_1^2 = -\sum_{i=1}^{n} \frac{x_i}{\beta_1^2}$$

$\beta_0 + \beta_1 x_i$

$$\frac{\partial^2 \ell}{(\partial \beta_0)(\partial \beta_1)} = \frac{\partial}{\partial \beta_0}\left(\frac{\partial \ell}{\partial \beta_1}\right) = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} x_i\left(y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

$$= -\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

$$= \sum_{i=1}^{n} -\left[\left(1+e^{\beta_0+\beta_1 x_i}\right) - \left(e^{\beta_0+\beta_1 x_i}\right)\right] / \beta_0^2$$

$$= -\sum_{i=1}^{n} 1/\beta_0^2 = -\sum_{i=1}^{n} \frac{1}{\beta_0^2}$$

$$\frac{\partial^2 \ell}{\partial \beta_1^2} = \frac{\partial}{\partial \beta_1}\left(\frac{\partial \ell}{\partial \beta_1}\right) = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} x_i\left(y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

$$= -\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{\underbrace{1+e^{\beta_0+\beta_1 x_i}}_{\to P(x)}}$$

$$= \sum_{i=1}^{n} -\left[x_i\left(1+e^{\beta_0+\beta_1 x_i}\right) - x_i\left(e^{\beta_0+\beta_1 x_i}\right)\right] / \beta_1^2$$

$$= \sum_{i=1}^{n} -x_i/\beta_1^2 = -\sum_{i=1}^{n} \frac{x_i}{\beta_1^2}$$

b) the hessian can be simplified
(now negative)

$$\begin{bmatrix} \sum_{i=1}^{n} \frac{1}{\beta_0^2} & \sum_{i=1}^{n} \frac{1}{\beta_0^2} \\ \sum_{i=1}^{n} \frac{x_i}{\beta_1^2} & \sum_{i=1}^{n} \frac{x_i}{\beta_1^2} \end{bmatrix} = X^T W X$$

$$\underset{n \times 2}{X} = \begin{matrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{matrix}$$

$$\underset{n \times n}{W} = \begin{bmatrix} P(x_1)(1-P(x_1)) & 0 & 0 & \cdots \\ 0 & P(x_2)(1-P(x_2)) & & \\ 0 & 0 & P(x_i)(1-P(x_3)) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

$$P(x_i)(1-P(x_i)) =$$

where $P(x) = \dfrac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}$

...

c) the hession of LSR

Gradient

$$\frac{\partial l}{\partial \beta_0} = \sum -2(y_i - (\beta_0 + \beta_1 x_i))$$

$$\frac{\partial l}{} = \sum -2(y_i - (\beta_0 + \beta_1 x_i))$$

$\partial\beta_1$

hessian

$$\frac{\partial^2 \ell}{\partial \beta_0^2} = \sum 2$$

$$\frac{\partial^2 \ell}{\partial \beta_1^2} = \sum 2x_i^2$$

$$\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} = \sum 2x_i$$

how is this similar in terms of x?

...