COMP 540 HW 02

Lyu Pan (lp28), Yuhui Tong (yt30)

January 28, 2018

1 Gradient and Hessian of $NLL(\theta)$ for logistic regression.

1.1

Proof. We explicitly express $\frac{\partial g(z)}{\partial z}$ and $g(z)\,(1-g(z))$ in terms of function of z respectively, i.e.,

$$\frac{\partial g(z)}{\partial z} = (1+e^{-z})^{-2}(-1)(-1)e^{-z} = \frac{e^{-z}}{(1+e^{-z})^2}, \tag{1}$$

and

$$g(z)\,(1-g(z)) = \frac{1}{1+e^{-z}}\left(1-\frac{1}{1+e^{-z}}\right) = \frac{e^{-z}}{(1+e^{-z})^2}. \tag{2}$$

Comparing above two equations gives

$$\frac{\partial g(z)}{\partial z} = g(z)\,(1-g(z)) \tag{3}$$

□

1.2

Proof. The negative log likelihood function of logistic regression model is described by

$$NLL(\theta) = -\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\theta(x^{(i)})\right) + (1-y^{(i)})\log\left(1-h_\theta(x^{(i)})\right)\right]. \tag{4}$$

Taking the derivative of Eq. (4) on θ_j gives:

$$\begin{aligned}
\frac{\partial NLL(\theta)}{\partial\theta_j} &= -\sum_{i=1}^{m}\left[y^{(i)}\frac{1}{h_\theta(x^{(i)})} - (1-y^{(i)})\frac{1}{1-h_\theta(x^{(i)})}\log\right]\frac{\partial h_\theta(x^{(i)})}{\partial\theta_j} && (5)\\
&= -\sum_{i=1}^{m}\left[y^{(i)}\frac{1}{h_\theta(x^{(i)})} - (1-y^{(i)})\frac{1}{1-h_\theta(x^{(i)})}\log\right]h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right)x_j^{(i)} && (6)\\
&= -\sum_{i=1}^{m}\left[y^{(i)} - h_\theta(x^{(i)})\right]x_j^{(i)} && (7)\\
&= \sum_{i=1}^{m}\left[h_\theta(x^{(i)}) - y^{(i)}\right]x_j^{(i)}, && (8)
\end{aligned}$$

where in the second step we make use of the the chain rule of derivative and the result of Sec. 1.1, i.e.

$$\frac{\partial h_\theta(x^{(i)})}{\partial \theta_j} = \frac{\partial g(z)}{\partial z} \cdot \frac{\partial z}{\partial \theta_j} = g(z)\,(1-g(z))\,x_j^{(i)} = h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right)x_j^{(i)},\ z=\theta^T x. \tag{9}$$

Finally, vectorizing Eq. (8) leads to the conclusion that

$$\frac{\partial NLL(\theta)}{\partial \theta} = \sum_{i=1}^{m}\left[h_\theta(x^{(i)}) - y^{(i)}\right]x^{(i)}. \tag{10}$$

□

1.3

Proof. Take ∀ nonzero column vector u, we have scalar $\alpha(u)$:

$$\alpha(u) = u^T H u = u^T X^T S X u = v^T S v \tag{11}$$

where $v = Xu$. Doing some linear algebra (by making use of the diagonality of S matrix) gives us the expression of the scalar

$$\begin{aligned}
\alpha(u) &= v^T S v = \sum_i \sum_j v_i S_{i,j} v_j &(12)\\
&= \sum_i \sum_j v_i \delta_{i,j} S_{i,i} v_j &(13)\\
&= \sum_i v_i S_{i,i} v_i &(14)\\
&= \sum_i (v_i)^2\, h_\theta(x^{(i)})\left[1-h_\theta(x^{(i)})\right]. &(15)
\end{aligned}$$

Considering the fact that X is full rank, it is implied that $v = Xu \neq 0, \forall$nonzero u, which futher indicating that every term in the summation of Eq. (15) is positive (assuming $h_\theta(x) \in (0,1)$), i.e.,

$$(v_i)^2\, h_\theta(x^{(i)})\left[1-h_\theta(x^{(i)})\right] > 0, \forall\ i. \tag{16}$$

Therefore we have

$$u^T H u > 0,\ \forall \text{ nonzero } u \tag{17}$$

i.e., matrix H is positive definite. □

2 Properties of L2 regularized logistic regression.