Logistic Regression HW Solutions

CS 4780/5780

Spring 2018

1

$$
\begin{aligned}
\sigma(-s) &= \frac{1}{1+e^{s}} \\
&= \frac{e^{-s}}{e^{-s}(1+e^{s})} \\
&= \frac{e^{-s}}{e^{-s}+1} \\
&= \frac{e^{-s}+1-1}{e^{-s}+1} \\
&= \frac{e^{-s}+1}{e^{-s}+1} - \frac{1}{e^{-s}+1} \\
&= 1 - \frac{1}{e^{-s}+1} \\
&= 1 - \sigma(s)
\end{aligned}
$$

2

(a)

$$\begin{aligned}
\sigma'(s) &= \frac{d}{ds}(\frac{1}{1+e^{-s}}) \\
&= \frac{d}{ds}(1+e^{-s}) \cdot (-(1+e^{-s})^{-2}) \\
&= (-e^{-s}) \cdot (-(1+e^{-s})^{-2}) \\
&= \frac{e^{-s}}{(1+e^{-s})^2} \\
&= \frac{1}{1+e^{-s}} \cdot \frac{e^{-s}}{1+e^{-s}} \\
&= \frac{1}{1+e^{-s}} \cdot \frac{e^{-s}+1-1}{1+e^{-s}} \\
&= \frac{1}{1+e^{-s}} \cdot (\frac{e^{-s}+1}{1+e^{-s}} - \frac{1}{1+e^{-s}}) \\
&= \frac{1}{1+e^{-s}} \cdot (1 - \frac{1}{1+e^{-s}}) \\
&= \sigma(s)(1-\sigma(s))
\end{aligned}$$

(b) Before we find the gradient, let's first write down the log likelihood function

$$\log P(\vec{y}|X,\vec{w}) = \log \prod_{i=1}^{n} \sigma(y_i(w^T\vec{x}_i))) = \sum_{i=1}^{n} \log \sigma(y_i(w^T\vec{x}_i)))$$

where in the last equality, we use the property of the logarithm function. To find the gradient, we will first find the k-th entry of the gradient. By definition, the k-th entry of the gradient is

$$\begin{aligned}
\frac{\partial}{\partial w_k} \log P(\vec{y}|X,\vec{w}) &= \sum_{i=1}^{n} \frac{\partial}{\partial w_k} \log(\sigma(y_i(w^T\vec{x}_i))) \\
&= \sum_{i=1}^{n} \frac{\sigma(y_i(w^T\vec{x}_i))(1-\sigma(y_i(w^T\vec{x}_i)))}{\sigma(y_i(w^T\vec{x}_i))} y_i x_{ik} \\
&= \sum_{i=1}^{n} (1-\sigma(y_i(w^T\vec{x}_i)))y_i x_{ik}
\end{aligned}$$

where in the 2nd step, we apply the Chain rule. Now, using the partial derivative, we know that

$$\begin{aligned}
\nabla_w P(y|X,w) &= \sum_{i=1}^{n} \begin{bmatrix} \frac{\partial log(\sigma(y_i(w^T\vec{x}_i)))}{\partial w_1} \\ \vdots \\ \frac{\partial log(\sigma(y_i(w^T\vec{x}_i)))}{\partial w_d} \end{bmatrix} \\
&= \sum_{i=1}^{n} \begin{bmatrix} (1-\sigma(y_i(w^T\vec{x}_i)))y_i x_{i1} \\ \vdots \\ (1-\sigma(y_i(w^T\vec{x}_i)))y_i x_{id} \end{bmatrix} \\
&= \sum_{i=1}^{n} (1-\sigma(y_i(w^T x_i + b)))y_i \vec{x}_i
\end{aligned}$$

(c) Now, in order to find the Hessian, we again find the (a,b)-th entry of the Hessian. By definition,

$$\begin{aligned}
H_{ab} &= \frac{\partial^2}{\partial w_a \partial w_b} \log P(\vec{y}|X,\vec{w}) \\
&= \frac{\partial}{\partial w_a}\left(\frac{\partial}{\partial w_b} \log P(\vec{y}|X,\vec{w})\right) \\
&= \frac{\partial}{\partial w_a} \sum_{i=1}^{n} (1-\sigma(y_i(w^T\vec{x}_i)))y_i x_{ib} \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial w_a} \sigma(y_i(w^T\vec{x}_i)))y_i x_{ib} \\
&= -\sum_{i=n}^{n} \sigma(y_i(\vec{w}^T\vec{x}_i))(1-\sigma(y_i(\vec{w}^T\vec{x}_i)))y_i^2 x_{ia} x_{ib}
\end{aligned}$$

Now, we are left to show that the (a,b)-th entry of $\vec{x_i}\vec{x_i}^T = x_{ia}x_{ib}$. We can verify this by expanding $\vec{x_i}\vec{x_i}^T$ as follows:

$$\begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix} \begin{bmatrix} x_{i1} & \dots x_{id} \end{bmatrix} = \begin{bmatrix} x_{i1} \cdot x_{i1} & \dots & x_{i1} \cdot x_{id} \\ \vdots & \ddots & \vdots \\ x_{id} \cdot x_{i1} & \dots & x_{id} \cdot x_{id} \end{bmatrix}$$

By inspection, it is easy to conclude that (a,b)-th entry of $\vec{x_i}\vec{x_i}^T$ is indeed $x_{ia}x_{ib}$ and with this result, we can conclude that

$$H = -\sum_{i=n}^{n} \sigma(y_i(\vec{w}^T\vec{x}_i))(1-\sigma(y_i(\vec{w}^T\vec{x}_i)))y_i^2 \vec{x}_i \vec{x}_i^T$$

(d) To show the Hessian is negative semidefinite, observe that for any $\vec{z} \in \mathbb{R}^d$

$$\vec{z}^T H \vec{z} = -\sum_{i=n}^{n} \sigma(y_i(\vec{w}^T \vec{x}_i))(1 - \sigma(y_i(\vec{w}^T \vec{x}_i)))y_i^2 \vec{z}^T \vec{x}_i \vec{x}_i^T \vec{z}$$

Since $\vec{z}^T \vec{x}_i = \vec{x}_i^T \vec{z}$, we can rewrite the quadratic form as

$$\vec{z}^T H \vec{z} = -\sum_{i=n}^{n} \sigma(y_i(\vec{w}^T \vec{x}_i))(1 - \sigma(y_i(\vec{w}^T \vec{x}_i)))y_i^2 (\vec{z}^T \vec{x}_i)^2$$

Since the expression after the summation is non-negative, we can conclude that $\vec{z}^T H \vec{z} \leq 0$. Thus, the log likelihood function is concave and any local minimum of the log likelihood function should be global.