1 Source

To the best of my recollection, the details of the logistic regression update rule math were not reviewed in class. As such, I watched the lectures from Andrew Ng's deep learning class. Here is what I *believe* the update rule should be. If there is an error in the logic, please let me know.

2 Glossary of Notation

- $\mathbf{w}_t$ – Weight vector for epoch t
- $J(\mathbf{w}, \mathbf{x})$ – Cost function
- β – Learning rate
- $\mathcal{L}$ – Loss function
- $\hat{y}$ – Predicted output value
- y – Expected (target) classification value
- $\sigma(z)$ – Sigmoid function $\left(\frac{1}{1+e^{-z}}\right)$ with respect to z (i.e., $\mathbf{w}^{\mathsf{T}}\mathbf{x}$).

3 w Update Rules with Squared Loss

My understanding of the *batch* update rule is shown in Eq. (1).

$$\mathbf{w}_{t+1} := \mathbf{w}_t - \beta \cdot \frac{\partial J(\mathbf{w}, \mathbf{x})}{\partial \mathbf{w}} \tag{1}$$

The learning rate is defined as:

$$\beta := \eta \cdot t^{-\alpha}$$

where $\alpha = 0.9$. The cost function is the average loss as shown in Eq. (2).

$$J(\mathbf{w}, \mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\mathbf{w}, \mathbf{x}) \tag{2}$$

The loss function is the squared loss and uses the same regularizer as in homework #1 as shown in Eq. (3).

$$\mathcal{L}(\hat{y}, y, \mathbf{w}) = \frac{1}{2}(\hat{y} - y)^2 + \lambda \|\mathbf{w}\| \tag{3}$$

The predicted value $\hat{y}$ is

$$\hat{y} = \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}).$$

The derivative of the loss function $\mathcal{L}$ is:

$$\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{x})}{\partial \mathbf{w}} = (\hat{y} - y) \frac{\partial \hat{y}}{\partial \mathbf{w}} + \lambda \mathbf{w}. \tag{4}$$

Via the chain rule, we solve the derivative of the sigmoid function:

$$\frac{\partial \hat{y}(z)}{\partial \mathbf{w}} = \sigma(z)(1-\sigma(z)) = \frac{e^{-z}}{(1+e^{-z})^2}\frac{\partial z}{\partial \mathbf{w}} \tag{5}$$

Applying the chain rule again yields:

$$\frac{\partial z}{\partial \mathbf{w}} = \mathbf{x} \tag{6}$$

Combining Eq. (4), (5), and (6) shows the complete derivative of the loss function.

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w},\mathbf{x})}{\partial \mathbf{w}} &= \Big(\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}) - y\Big)\left(\frac{e^{-z}}{(1+e^{-z})^2}\right)\mathbf{x} + \lambda\mathbf{w} && (7)\\
&= \Big(\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}) - y\Big)\left(\frac{e^{-\mathbf{w}^{\mathsf{T}}\mathbf{x}}}{(1+e^{-\mathbf{w}^{\mathsf{T}}\mathbf{x}})^2}\right)\mathbf{x} + \lambda\mathbf{w} && (8)
\end{aligned}$$

For mathematical simplicity, the identity for $\sigma'(z)$ allows for a simpler form in Eq. (9).[1]

$$\frac{\partial \mathcal{L}(\mathbf{w},\mathbf{x})}{\partial \mathbf{w}} = \Big(\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}) - y\Big)\,\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x})\,\Big(1-\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x})\Big)\,\mathbf{x} + \lambda\mathbf{w} \tag{9}$$

4 w Update Rules with Logistic Loss

The more common loss function I see for logistic regression is in Eq. (10).

$$\mathcal{L}(\hat{y}, y, \mathbf{w}) = -\left(y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right) \tag{10}$$

The derivative of the logistic loss function is shown below in Eq. (11).

$$\frac{\partial \mathcal{L}(\hat{y}, y, \mathbf{w})}{\partial \mathbf{w}} = -\left(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right)\frac{\partial \hat{y}}{\partial \mathbf{w}} \tag{11}$$

We know the derivative of $\hat{y}$ from Eq. (5). Substituting that we get:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\hat{y}, y, \mathbf{w})}{\partial \mathbf{w}} &= -\left(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right)\hat{y}(1-\hat{y})\frac{\partial \mathbf{z}}{\partial \mathbf{w}} && (12)\\
&= \left(-y(1-\hat{y}) + (1-y)\hat{y})\right)\frac{\partial \mathbf{z}}{\partial \mathbf{w}} && (13)\\
&= (\hat{y} - y)\,\frac{\partial \mathbf{z}}{\partial \mathbf{w}}. && (14)
\end{aligned}$$

The complete derivative then is in Eq. (15).

$$\frac{\partial \mathcal{L}(\hat{y}, y, \mathbf{w})}{\partial \mathbf{w}} = (\hat{y} - y)\,\mathbf{x}. \tag{15}$$

[1] I am not considering the transposes. That math I would need to think more about.