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 classification value
- $\sigma(z)$ – Sigmoid function $\left(\frac{1}{1+e^{-z}}\right)$ with respect to z (i.e., $\mathbf{w}^{\mathrm{T}}\mathbf{x}$).

3 w Update Rules

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).

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

The predicted value $\hat{y}$ is

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

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

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

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

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

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.

$$\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{x})}{\partial \mathbf{w}} = (\hat{y} - y) \cdot \frac{e^{-z}}{(1+e^{-z})^2} \cdot \mathbf{x} + \lambda \mathbf{w} \tag{7}$$

$$= (\sigma(\mathbf{w}^{\mathrm{T}} \mathbf{x}) - y) \cdot \frac{e^{-\mathbf{w}^{\mathrm{T}} \mathbf{x}}}{(1+e^{-\mathbf{w}^{\mathrm{T}} \mathbf{x}})^2} \cdot \mathbf{x} + \lambda \mathbf{w} \tag{8}$$