CS446: Machine Learning, Fall 2017, Homework 1

Name: Lanxiao Bai(lbai5)

Worked individually

Problem 2

1. **Solution:** Since we see by definition

$$p(y \mid \mathbf{x}, \mathbf{w}) = \mathrm{Ber}(y \mid \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x}))$$

we can derive that

$$\begin{aligned} p(y = 1 \mid \mathbf{x}, \mathbf{w}) &= \mathrm{Ber}(y = 1 \mid \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x})) \\ &= \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x})^{1_{[y=1]}} \\ &= \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x}) \end{aligned} \tag{1}$$

$$\begin{aligned} p(y = 0 \mid \mathbf{x}, \mathbf{w}) &= \mathrm{Ber}(y = 0 \mid \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x})) \\ &= (1 - \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x}))^{1_{[y=0]}} \\ &= 1 - \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x}) \end{aligned} \tag{2}$$

2. **Solution:** The derivative of the Sigmoid function is

$$\begin{aligned} \frac{d}{dz}\mathbf{sigm}(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{e^x}{(1+e^x)^2} \qquad \text{(\textit{Weisstein} (2017))} \end{aligned}$$

We also see that

$$\frac{d}{dz}\mathbf{sigm}(z) = \mathbf{sigm}(z)(1 - \mathbf{sigm}(z)) \tag{3}$$

3. **Solution:** The likelihood of logistic regression

$$P(y \mid \mathbf{x}, \mathbf{w}) = \prod_{x_i : y_i = 1} \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i)^{1_{[y_i=1]}} \prod_{x_i : y_i = 0} (1 - \mathbf{sigm}(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i))^{1_{[y_i=0]}} \tag{4}$$

4. **Solution:** Base on equation (4), we have the log likelihood function

$$\begin{aligned}\mathcal{L}(y \mid \mathbf{X}, \mathbf{w}) &= \log P(y \mid \mathbf{X}, \mathbf{w}) \\ &= \sum_{i=1}^{N} [y_i \log \mathbf{sigm}(\mathbf{w^T x}) + (1 - y_i) \log(1 - \mathbf{sigm}(\mathbf{w^T x}))]\end{aligned}$$

Thus, we have the gradient

$$\begin{aligned}\nabla \mathcal{L}(y \mid \mathbf{X}, \mathbf{w}) &= \frac{d}{d\mathbf{w}} \mathcal{L}(y \mid \mathbf{X}, \mathbf{w}) \\ &= \frac{d\mathcal{L}(y \mid \mathbf{X}, \mathbf{w})}{d\mathbf{sigm}(\mathbf{w^T X})} \frac{d\mathbf{sigm}(\mathbf{w^T X})}{d\mathbf{w^T X}} \frac{d\mathbf{w^T X}}{d\mathbf{w}} \\ &= \sum_{i=1}^{N} \left(\frac{y_i}{\mathbf{sigm}(\mathbf{w^T x}_i)} - \frac{1 - y_i}{1 - \mathbf{sigm}(\mathbf{w^T x}_i)} \right) \mathbf{sigm}(\mathbf{w^T x}_i)(1 - \mathbf{sigm}(\mathbf{w^T x}_i))\mathbf{x}_i \\ &= \sum_{i=1}^{N} (y_i - \mathbf{sigm}(\mathbf{w^T x}_i))\mathbf{x}_i\end{aligned}$$

Thus, we finally get the update rule for gradient descent

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \sum_{i=1}^{N} (y_i - \mathbf{sigm}(\mathbf{w^T x}_i))\mathbf{x}_i$$

References

WEISSTEIN, E. W. (2017). Sigmoid function.
URL http://mathworld.wolfram.com/SigmoidFunction.html