Pranshu Malik
1004138916

Solution 1: Gradient Computation

We use the knowledge of the following rules to calculate the gradients: $\nabla_{\mathbf{w}}(\mathbf{w}^\top \mathbf{v}) = \nabla_{\mathbf{w}}(\mathbf{v}^\top \mathbf{w}) = \mathbf{v}$ and $\nabla_{\mathbf{w}}(\mathbf{w}^\top \mathbf{A}\mathbf{w}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{w}$ for some matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$.

i) The gradient reduces to $\nabla_{\mathbf{w}} f = 2(\mathbf{A} + \mathbf{A}^\top)\mathbf{v}$

ii) $\nabla_{\mathbf{w}} f = (\mathbf{A} + \mathbf{A}^\top)\mathbf{w}$

iii) $\nabla_{\mathbf{w}} f = \begin{bmatrix}\theta(w_1) & \dots & \theta(w_d)\end{bmatrix}^\top$, where $\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$ is the logistic function.

iv) $\nabla_{\mathbf{w}} f = \frac{\mathbf{w}}{\sqrt{1+||\mathbf{w}||^2}}$

Solution 2: Logistic Regression

For the dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^{d+1}$ is an augmented datavector $\begin{bmatrix}1 & x_{i,1} & \dots & x_{i,d}\end{bmatrix}^\top$ with label $y_i \in \{-1, 1\}$, and given any datavector $\mathbf{x}_n \in \mathbb{R}^{d+1}$, the logistic regression model outputs the probability estimate for the true class y_n as

$$\hat{p}_{\mathbf{w}}(y_n \mid \mathbf{x}_n) = \theta\left(y_n \mathbf{w}^\top \mathbf{x}_n\right). \tag{1}$$

The loss function (in-sample error) for the model is defined to be

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{-\log\left(\hat{p}_{\mathbf{w}}(y_n \mid \mathbf{x}_n)\right)}_{e_n(\mathbf{w})} \tag{2}$$

a) From (2) we can see that since there are only two classes, we can rewite the loss function as

$$\begin{aligned} E_{\text{in}}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}\Big[-I(y_n = 1)\log(\hat{p}_{\mathbf{w}}(1 \mid \mathbf{x}_n)) - I(y_n = -1)\log(\hat{p}_{\mathbf{w}}(-1 \mid \mathbf{x}_n))\Big] \\ &= \frac{1}{N}\sum_{n=1}^{N}\Big[-\frac{1+y_n}{2}\log(\hat{p}_{\mathbf{w}}(1 \mid \mathbf{x}_n)) - \frac{1-y_n}{2}\log(1-\hat{p}_{\mathbf{w}}(1 \mid \mathbf{x}_n))\Big], \end{aligned} \tag{3}$$

where $I(y_n = 1)$ and $I(y_n = -1)$ are the indicator functions. □

b) From (1) and (2) we can see that

$$e_n(\mathbf{w}) = -\log\left(\hat{p}_{\mathbf{w}}(y_n \mid \mathbf{x}_n)\right) = \log(1 + e^{-y_n \mathbf{w}^\top \mathbf{x}_n})$$

Therefore, we can also rewrite the loss function as

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\log(1 + e^{-y_n \mathbf{w}^\top \mathbf{x}_n}) \tag{4}$$

□

c) Now, from (4) we can see that the gradient of the loss function is given by

$$\begin{aligned} \nabla_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}\nabla_{\mathbf{w}}\log(1 + e^{-y_n \mathbf{w}^\top \mathbf{x}_n}) = \frac{1}{N}\sum_{n=1}^{N}\frac{-y_n \mathbf{x}_n e^{-y_n \mathbf{w}^\top \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^\top \mathbf{x}_n}} \\ &= \frac{1}{N}\sum_{n=1}^{N} -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^\top \mathbf{x}_n). \end{aligned} \tag{5}$$

Continuing with what we established in Homework 1, here notice that for a confidently missclassified point $y_n \mathbf{w}^\top \mathbf{x}_n \ll 0$ which gives $\theta(-y_n \mathbf{w}^\top \mathbf{x}_n) \approx 1$, whereas for a confidently correctly classified point, we have $y_n \mathbf{w}^\top \mathbf{x}_n \gg 0$ which gives $\theta(-y_n \mathbf{w}^\top \mathbf{x}_n) \approx 0$, meaning that there is little contribution compared to the misclassified case. □

d) From (3) notice that

$$
\begin{aligned}
\nabla_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}\Big[-\frac{1+y_n}{2}\nabla_{\mathbf{w}}\log(\hat{p}_{\mathbf{w}}(1\mid\mathbf{x}_n)) - \frac{1-y_n}{2}\nabla_{\mathbf{w}}\log(1-\hat{p}_{\mathbf{w}}(1\mid\mathbf{x}_n))\Big] \\
&= \frac{1}{2N}\sum_{n=1}^{N}\big[-(1+y_n)(1+e^{-\mathbf{w}^\top\mathbf{x}_n})\nabla_{\mathbf{w}}(1+e^{-\mathbf{w}^\top\mathbf{x}_n})^{-1} - (1-y_n)(1+e^{\mathbf{w}^\top\mathbf{x}_n})\nabla_{\mathbf{w}}(1+e^{\mathbf{w}^\top\mathbf{x}_n})^{-1}\big] \\
&= \frac{1}{2N}\sum_{n=1}^{N}\big[-(1+y_n)\mathbf{x}_n \underbrace{\frac{e^{-\mathbf{w}^\top\mathbf{x}_n}}{(1+e^{-\mathbf{w}^\top\mathbf{x}_n})}}_{\theta(-\mathbf{w}^\top\mathbf{x}_n)=1-\theta(\mathbf{w}^\top\mathbf{x}_n)} + (1-y_n)\mathbf{x}_n \underbrace{\frac{e^{\mathbf{w}^\top\mathbf{x}_n}}{(1+e^{\mathbf{w}^\top\mathbf{x}_n})}}_{\theta(\mathbf{w}^\top\mathbf{x}_n)}\big] \\
&= \frac{1}{2N}\sum_{n=1}^{N}\mathbf{x}_n\left(2\theta(\mathbf{w}^\top\mathbf{x}_n) - y_n - 1\right) \\
&= \frac{1}{N}\mathbf{X}^\top\underbrace{\left(\boldsymbol{\theta}(\mathbf{X}\mathbf{w}) - \frac{\mathbf{y}}{2} - \frac{\mathbf{I}}{2}\right)}_{p},
\end{aligned} \tag{6}
$$

where $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\top & \dots & \mathbf{x}_N^\top\end{bmatrix}^\top \in \mathbb{R}^{N\times(d+1)}$. Notice that the expression p is similar to $\mathbf{X}\mathbf{w}-\mathbf{y}$ in the gradient of the mean squared error loss function (for linear regression) derived in Homework 1. $\square$

Solution 3: Midterm 2017, Problem 4

We are given $\mathcal{D} = \left\{\left(\begin{bmatrix}1 & 1\end{bmatrix}^\top, 1\right), \left(\begin{bmatrix}1 & 0\end{bmatrix}^\top, -1\right)\right\} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\}$ and the corresponding regularized loss function for logistic regression

$$
E_{\text{in}}(\mathbf{w}) = \lambda||\mathbf{w}||^2 + \frac{1}{N}\sum_{n=1}^{N} e_n(\mathbf{w}), \tag{7}
$$

which is a shorthand for (4) except for the weight regularization term.

a) Since $\lambda = 0$, we can evaluate the loss function for $\mathbf{w} = \begin{bmatrix}w_1 & w_2\end{bmatrix}^\top$ as follows

$$
E_{\text{in}}(\mathbf{w}) = \log(1+e^{-(w_1+w_2)}) + \log(1+e^{w_1}).
$$

This expression can be minimized if $w_1 \to -\infty$ and $w_1 + w_2 \to \infty$. This can be realized with $w_2 = -2w_1$ and $w_1 \to \infty$, which will yield $E_{\text{in}} = 0$.

b) For $\lambda \gg 0$, we can consider $||w|| \ll 1$ while minimizing the loss function. For this case, we are given the Taylor approximation of the loss function for $e_n(\mathbf{w}) \approx \log(2) - \frac{1}{2}y_n\mathbf{w}^\top\mathbf{x}_n$, which we can substituted into the expression loss in part a) to then minimize it, i.e.,

$$
\begin{aligned}
E_{\text{in}}(\mathbf{w}) &= \lambda(w_1^2 + w_2^2) + \log(1+e^{-(w_1+w_2)}) + \log(1+e^{w_1}) \\
&= \lambda(w_1^2 + w_2^2) + \log(2) - \frac{1}{2}(w_1+w_2) + \log(2) + \frac{1}{2}w_1.
\end{aligned}
$$

Now to minimize the loss, we set

$$
\begin{aligned}
\frac{\partial E_{\text{in}}}{\partial w_1} = 0 &\implies 2\lambda w_1 - \frac{1}{2} + \frac{1}{2} = 0 \iff w_1 = 0 \\
\frac{\partial E_{\text{in}}}{\partial w_2} = 0 &\implies 2\lambda w_2 - \frac{1}{2} = 0 \iff w_2 = \frac{1}{4\lambda}.
\end{aligned}
$$

Thus, $\mathbf{w}^\star = \begin{bmatrix}0 & \frac{1}{4\lambda}\end{bmatrix}^\top$ is the minimizer.