Machine Learning Course 2024 Spring: Homework 3

April 28, 2024

1 Problem 1

Let

$$\boldsymbol{W}_1 = \begin{bmatrix} w_{1,11} & w_{1,12} & \cdots & w_{1,1n} \\ w_{1,21} & w_{1,22} & \cdots & w_{1,2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1,d1} & w_{1,d2} & \cdots & w_{1,dn} \end{bmatrix} \in \mathbb{R}^{d\times n},\ \boldsymbol{W}_2 = \begin{bmatrix} w_{2,11} & w_{2,12} & \cdots & w_{2,1q} \\ w_{2,21} & w_{2,22} & \cdots & w_{2,2q} \\ \vdots & \vdots & \ddots & \vdots \\ w_{2,n1} & w_{2,n2} & \cdots & w_{2,nq} \end{bmatrix} \in \mathbb{R}^{n\times q},$$

$$\boldsymbol{b}_1 = \begin{bmatrix} b_{1,1} \\ b_{1,2} \\ \vdots \\ b_{1,n} \end{bmatrix} \in \mathbb{R}^n,\ \boldsymbol{b}_2 = \begin{bmatrix} b_{2,1} \\ b_{2,2} \\ \vdots \\ b_{2,q} \end{bmatrix} \in \mathbb{R}^q.$$

Then, we have

$$\boldsymbol{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} = \boldsymbol{W_1}^\top \boldsymbol{x} + \boldsymbol{b_1} = \begin{bmatrix} \sum_{m=1}^d w_{1,m1} \cdot x_m + b_{1,1} \\ \sum_{m=1}^d w_{1,m2} \cdot x_m + b_{1,2} \\ \vdots \\ \sum_{m=1}^d w_{1,mn} \cdot x_m + b_{1,n} \end{bmatrix} \in \mathbb{R}^q,$$

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_q \end{bmatrix} = \boldsymbol{W_2}^\top \boldsymbol{a} + \boldsymbol{b_2} = \begin{bmatrix} \sum_{l=1}^n w_{2,l1} \cdot a_l + b_{2,1} \\ \sum_{l=1}^n w_{2,l2} \cdot a_l + b_{2,2} \\ \vdots \\ \sum_{l=1}^n w_{2,lq} \cdot a_l + b_{2,q} \end{bmatrix} \in \mathbb{R}^q.$$

Solution:

□

The loss function is

$$\text{Loss} = -\sum_{i=1}^{q} y_i^s \ln(\hat{y}_i) = -\sum_{i=1}^{q} \left((1-\epsilon)y_i + \frac{\epsilon}{q}\right) \ln(\hat{y}_i)\,. \tag{1}$$

The derivative of the loss function w.r.t $\hat{y}_i$ $(1 \le i \le q)$ is

$$\begin{aligned} \frac{\partial \text{Loss}}{\partial \hat{y}_i} &= -\sum_{i=1}^{q} \frac{\partial \left((1-\epsilon)y_i + \frac{\epsilon}{q}\right) \ln(\hat{y}_i)}{\partial \hat{y}_i} \\ &= -(1-\epsilon)\sum_{i=1}^{q} \frac{y_i}{\hat{y}_i} - \frac{\epsilon}{q}\sum_{i=1}^{q} \frac{1}{\hat{y}_i}. \end{aligned} \tag{2}$$

The derivative of the soft-max function w.r.t z_j $(1 \le j \le q)$ is

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{\frac{\partial \exp(z_i)}{\partial z_j} \sum_{k=1}^{q} \exp(z_k) - \exp(z_i) \frac{\partial \sum_{k=1}^{q} \exp(z_k)}{\partial z_j}}{\left(\sum_{k=1}^{q} \exp(z_k)\right)^2}. \tag{3}$$

When $i = j$, we have

$$\begin{aligned} \frac{\partial \hat{y}_i}{\partial z_j} &= \frac{\exp(z_i) \sum_{k=1}^{q} \exp(z_k) - \exp(z_i)\exp(z_j)}{\left(\sum_{k=1}^{q} \exp(z_k)\right)^2} \\ &= \hat{y}_i(1-\hat{y}_j). \end{aligned} \tag{4}$$

When $i \neq j$, we have

$$\begin{aligned} \frac{\partial \hat{y}_i}{\partial z_j} &= \frac{0 \sum_{k=1}^{q} \exp(z_k) - \exp(z_i)\exp(z_j)}{\left(\sum_{k=1}^{q} \exp(z_k)\right)^2} \\ &= -\hat{y}_i\hat{y}_j. \end{aligned} \tag{5}$$

So, the derivative of the soft-max function w.r.t z_j is

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1-\hat{y}_j), & if\ i = j, \\ -\hat{y}_i\hat{y}_j, & if\ i \neq j. \end{cases} \tag{6}$$

According to Eq.(2) and Eq.(6), the derivative of the loss function w.r.t z_j is

$$\begin{aligned}
\frac{\partial \text{Loss}}{\partial z_j} &= \frac{\partial \text{Loss}}{\partial \hat{y_i}}\frac{\partial \hat{y_i}}{\partial z_j} \\
&= -(1-\epsilon)\sum_{i=1}^{q}\frac{y_i}{\hat{y_i}}\frac{\partial \hat{y_i}}{\partial z_j} - \frac{\epsilon}{q}\sum_{i=1}^{q}\frac{1}{\hat{y_i}}\frac{\partial \hat{y_i}}{\partial z_j} \\
&= (1-\epsilon)\left(-\frac{y_j}{\hat{y_j}}\hat{y_j}(1-\hat{y_j}) + \sum_{i=1,i\neq j}^{q}\frac{y_i}{\hat{y_i}}\hat{y_i}\hat{y_j}\right) + \frac{\epsilon}{q}\left(-\frac{1}{\hat{y_j}}\hat{y_j}(1-\hat{y_j}) + \sum_{i=1,i\neq j}^{q}\frac{1}{\hat{y_i}}\hat{y_i}\hat{y_j}\right) \\
&= (1-\epsilon)(-y_j + y_j\hat{y_j} + \sum_{i=1,i\neq j}^{q} y_i\hat{y_j}) + \frac{\epsilon}{q}(-1 + \hat{y_j} + \sum_{i=1,i\neq j}^{q}\hat{y_j}) \\
&= (1-\epsilon)(-y_j + \sum_{i=1}^{q} y_i\hat{y_j}) + \frac{\epsilon}{q}(-1 + \sum_{i=1}^{q}\hat{y_j}) \\
&= (1-\epsilon)(\hat{y_j} - y_j) + \frac{\epsilon}{q}(q\hat{y_j} - 1) \\
&= \hat{y_j} - (1-\epsilon)y_j - \frac{\epsilon}{q}.
\end{aligned} \tag{7}$$

The derivative of z_j w.r.t $w_{2,kj}$ $(1 \le k \le n, 1 \le j \le q)$ is

$$\frac{\partial z_j}{\partial w_{2,kj}} = \frac{\partial(\sum_{l=1}^{n} w_{2,lj}\cdot a_l + b_{2,j})}{\partial w_{2,kj}} = a_k. \tag{8}$$

According to Eq.(7) and Eq.(8), we have

$$\frac{\partial \text{Loss}}{\partial w_{2,kj}} = \frac{\partial \text{Loss}}{\partial z_j}\frac{\partial z_j}{\partial w_{2,kj}} = \left(\hat{y_j} - (1-\epsilon)y_j - \frac{\epsilon}{q}\right)\cdot a_k. \tag{9}$$

Thus,

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{W}_2} = \begin{bmatrix} \left(\hat{y_1} - (1-\epsilon)y_1 - \frac{\epsilon}{q}\right)\cdot a_1 & \left(\hat{y_2} - (1-\epsilon)y_2 - \frac{\epsilon}{q}\right)\cdot a_1 & \cdots & \left(\hat{y_n} - (1-\epsilon)y_n - \frac{\epsilon}{q}\right)\cdot a_1 \\ \left(\hat{y_1} - (1-\epsilon)y_1 - \frac{\epsilon}{q}\right)\cdot a_2 & \left(\hat{y_2} - (1-\epsilon)y_2 - \frac{\epsilon}{q}\right)\cdot a_2 & \cdots & \left(\hat{y_n} - (1-\epsilon)y_n - \frac{\epsilon}{q}\right)\cdot a_2 \\ \vdots & \vdots & \ddots & \vdots \\ \left(\hat{y_1} - (1-\epsilon)y_1 - \frac{\epsilon}{q}\right)\cdot a_q & \left(\hat{y_2} - (1-\epsilon)y_2 - \frac{\epsilon}{q}\right)\cdot a_q & \cdots & \left(\hat{y_n} - (1-\epsilon)y_n - \frac{\epsilon}{q}\right)\cdot a_q \end{bmatrix}. \tag{10}$$

The derivative of z_j w.r.t $b_{2,j}$ $(1 \le j \le q)$ is

$$\frac{\partial z_j}{\partial b_{2,j}} = \frac{\partial(\sum_{l=1}^{n} w_{2,lj}\cdot a_l + b_{2,j})}{\partial b_{2,j}} = 1. \tag{11}$$

According to Eq.(7) and Eq.(11), we have

$$\frac{\partial \text{Loss}}{\partial b_{2,j}} = \frac{\partial \text{Loss}}{\partial z_j}\frac{\partial z_j}{b_{2,j}} = \hat{y_j} - (1-\epsilon)y_j - \frac{\epsilon}{q}. \tag{12}$$

Thus,

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{b}_2} = \begin{bmatrix} \hat{y_1} - (1-\epsilon)y_1 - \frac{\epsilon}{q} \\ \hat{y_2} - (1-\epsilon)y_2 - \frac{\epsilon}{q} \\ \vdots \\ \hat{y_q} - (1-\epsilon)y_q - \frac{\epsilon}{q} \end{bmatrix}. \tag{13}$$

The derivative of z_j w.r.t a_p $(1 \leq p \leq n)$ is

$$\frac{\partial z_j}{\partial a_p} = \frac{\partial(\sum_{l=1}^{n} w_{2,lj} \cdot a_l + b_{2,j})}{\partial a_p} = w_{2,pj}. \tag{14}$$

The derivative of the ReLU activation function w.r.t h_p $(1 \leq p \leq n)$is

$$f(h_i) = \frac{\partial a_p}{\partial h_p} = \frac{\partial \text{ReLU}(h_p)}{\partial h_p} = \begin{cases} 0, & if\ h_i < 0, \\ 1, & otherwise. \end{cases} \tag{15}$$

The derivative of h_p w.r.t $w_{1,rp}$ $(1 \leq r \leq d, 1 \leq p \leq n)$ is

$$\frac{\partial h_p}{\partial w_{1,rp}} = \frac{\partial(\sum_{m=1}^{d} w_{1,mp} \cdot x_m + b_{2,p})}{\partial w_{1,rp}} = x_r. \tag{16}$$

According to Eq.(7), Eq.(14), Eq.(15) and Eq.(16) we have

$$\frac{\partial \text{Loss}}{\partial w_{1,rp}} = \sum_{s=1}^{q} \left(\frac{\partial \text{Loss}}{\partial z_s} \frac{\partial z_s}{\partial a_p} \right) \frac{\partial a_p}{\partial h_p} \frac{\partial h_p}{\partial w_{1,rp}} = \left(\sum_{s=1}^{q} g_s \cdot w_{2,ps} \right) \cdot f(h_p) \cdot x_r. \tag{17}$$

Thus,

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{W}_1} = \begin{bmatrix} (\sum_{s=1}^{q} g_s w_{2,1s}) \cdot f(h_1) \cdot x_1 & (\sum_{s=1}^{q} g_s w_{2,2s}) \cdot f(h_2) \cdot x_1 & \cdots & (\sum_{s=1}^{q} g_s w_{2,ns}) \cdot f(h_n) \cdot x_1 \\ (\sum_{s=1}^{q} g_s w_{2,1s}) \cdot f(h_1) \cdot x_2 & (\sum_{s=1}^{q} g_s w_{2,2s}) \cdot f(h_2) \cdot x_2 & \cdots & (\sum_{s=1}^{q} g_s w_{2,ns}) \cdot f(h_n) \cdot x_2 \\ \vdots & \vdots & \ddots & \vdots \\ (\sum_{s=1}^{q} g_s w_{2,1s}) \cdot f(h_1) \cdot x_d & (\sum_{s=1}^{q} g_s w_{2,2s}) \cdot f(h_2) \cdot x_d & \cdots & (\sum_{s=1}^{q} g_s w_{2,ns}) \cdot f(h_n) \cdot x_d \end{bmatrix}, \tag{18}$$

where $g_s = \left(\hat{y_s} - (1 - \epsilon) y_s - \frac{\epsilon}{q} \right)$.

The derivative of h_p w.r.t $b_{1,p}$ $(1 \leq p \leq n)$ is

$$\frac{\partial h_p}{\partial b_{1,p}} = \frac{\partial(\sum_{m=1}^{d} w_{1,mp} \cdot x_m + b_{1,p})}{\partial b_{1,p}} = 1. \tag{19}$$

According to Eq.(7), Eq.(14), Eq.(15) and Eq.(19) we have

$$\frac{\partial \text{Loss}}{\partial b_{1,p}} = \sum_{s=1}^{q} \left(\frac{\partial \text{Loss}}{\partial z_s} \frac{\partial z_s}{\partial a_p} \right) \frac{\partial a_p}{\partial h_p} \frac{\partial h_p}{\partial b_{1,p}} = \left(\sum_{s=1}^{q} g_s \cdot w_{2,ps} \right) \cdot f(h_p) \cdot x_r. \tag{20}$$

Thus,

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{b}_1} = \begin{bmatrix} (\sum_{s=1}^{q} g_s \cdot w_{2,1s}) \cdot f(h_1) \\ (\sum_{s=1}^{q} g_s \cdot w_{2,2s}) \cdot f(h_2) \\ \vdots \\ (\sum_{s=1}^{q} g_s \cdot w_{2,ns}) \cdot f(h_n) \end{bmatrix}. \tag{21}$$

2 Problem 2

First, we consider the feed-forward process:

$$\boldsymbol{h} = \boldsymbol{W}_1^\top \boldsymbol{x} + \boldsymbol{b}_1 = \begin{bmatrix} -13 \\ 2 \\ -5 \end{bmatrix}, \boldsymbol{a} = \text{ReLU}(\boldsymbol{h}) = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix},$$

$$\boldsymbol{z} = \boldsymbol{W}_2^\top \boldsymbol{a} + \boldsymbol{b}_2 = \begin{bmatrix} 0 \\ -3 \\ 1 \end{bmatrix}, \hat{\boldsymbol{y}} = \text{Softmax}(\boldsymbol{z}) = \begin{bmatrix} 0.2654 \\ 0.0132 \\ 0.7214 \end{bmatrix},$$

$$\text{Loss} = -\sum_{i=1}^{q} y_i^s \ln(\hat{y}_i) = 1.5266.$$

Then, we compute the back-propagation using Eq.(10), Eq.(13), Eq.(18) and Eq.(21):

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{W}_2} = \begin{bmatrix} a_1 \cdot (\hat{y_1} - 0.8) & a_1 \cdot (\hat{y_2} - 0.1) & a_1 \cdot (\hat{y_3} - 0.1) \\ a_2 \cdot (\hat{y_1} - 0.8) & a_2 \cdot (\hat{y_2} - 0.1) & a_2 \cdot (\hat{y_3} - 0.1) \\ a_3 \cdot (\hat{y_1} - 0.8) & a_3 \cdot (\hat{y_2} - 0.1) & a_3 \cdot (\hat{y_3} - 0.1) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ -1.0692 & -0.1736 & 1.2428 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{b}_2} = \begin{bmatrix} \hat{y_1} - 0.8 \\ \hat{y_2} - 0.1 \\ \hat{y_3} - 0.1 \end{bmatrix} = \begin{bmatrix} -0.5346 \\ -0.0868 \\ 0.6214 \end{bmatrix},$$

$$\begin{aligned} \frac{\partial \text{Loss}}{\partial \boldsymbol{W}_1} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & \left(\sum_{s=1}^{q} g_s w_{2,2s}\right) \cdot x_2 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & [-0.5346 \times (-2) + (-0.0868) \times (-4) + 0.6214 \times (-2)] \cdot 2 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.3472 & 0 \end{bmatrix}, \end{aligned}$$

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{b}_1} = \begin{bmatrix} 0 \\ \sum_{s=1}^{q} g_s \cdot w_{2,2s} \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ -0.5346 \times (-2) + (-0.0868) \times (-4) + 0.6214 \times (-2) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.1736 \\ 0 \end{bmatrix}.$$

Given the gradients above, we update the parameters as follows:

$$\boldsymbol{W}_2 \leftarrow \boldsymbol{W}_2 - \eta \cdot \frac{\partial \text{Loss}}{\partial \boldsymbol{W}_2} = \begin{bmatrix} 3 & -4 & 1 \\ -2 & -4 & -2 \\ -4 & -2 & 3 \end{bmatrix} - 0.1 \times \begin{bmatrix} 0 & 0 & 0 \\ -1.0692 & -0.1736 & 1.2428 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & -4 & 1 \\ -1.8931 & -3.9826 & -2.1243 \\ -4 & -2 & 3 \end{bmatrix},$$

$$\boldsymbol{b}_2 \leftarrow \boldsymbol{b}_2 - \eta \cdot \frac{\partial \text{Loss}}{\partial \boldsymbol{b}_2} = \begin{bmatrix} 4 \\ 5 \\ 5 \end{bmatrix} - 0.1 \times \begin{bmatrix} -0.5346 \\ -0.0868 \\ 0.6214 \end{bmatrix} = \begin{bmatrix} 4.0535 \\ 5.0087 \\ 4.9379 \end{bmatrix},$$

$$\boldsymbol{W}_1 \leftarrow \boldsymbol{W}_1 - \eta \cdot \frac{\partial \text{Loss}}{\partial \boldsymbol{W}_1} = \begin{bmatrix} -5 & 2 & 2 \\ -5 & 2 & -1 \end{bmatrix} - 0.1 \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.3472 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} -5 & 2 & 2 \\ -5 & 1.9653 & -1 \end{bmatrix},$$

$$\boldsymbol{b}_1 \leftarrow \boldsymbol{b}_1 - \eta \cdot \frac{\partial \text{Loss}}{\partial \boldsymbol{b}_1} = \begin{bmatrix} -3 \\ -2 \\ -3 \end{bmatrix} - 0.1 \times \begin{bmatrix} 0 \\ 0.1736 \\ 0 \end{bmatrix} = \begin{bmatrix} -3 \\ -2.0174 \\ -3 \end{bmatrix}.$$