ECE 421 Homework 4

Name: Mark Qi

Student ID: 1006764645

Question 1,

Input Data Vector: $\underline{x} \in \mathbb{R}^{d+1}$

Two Possible Class Label: $y \in \{1, 2\}$

Weight Vectors: $\underline{w(1) \; , \; w(2)}$

Probability of $\underline{x}$ belonging to class $i \in \{1, 2\}$

$$P^{SM}(y = i \,|\, \underline{x}) = \frac{e^{\underline{w(i)^T x}}}{e^{\underline{w(1)^T x}} + e^{\underline{w(2)^T x}}}$$

For Training Example $(\underline{x_n}, y_n)$

$$e_n^{SM}\underline{(w(1), w(2))} = -\log(P^{SM}(y_n \,|\, \underline{x_n}))$$

$$= -\log\left[\frac{e^{\underline{w(y_n)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}\right]$$

Part (a),

Simplify $e_n^{SM}(\underline{w(1), w(2)})$

$$e_n^{SM}(\underline{w(1), w(2)}) = -\left[\log e^{\underline{w(y_n)^T x_n}} - \log e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}\right]$$

$$= -\underline{w(y_n)^T x_n} + \log(e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}})$$

Two Possible Situations,

$$e_n^{SM}(\underline{w(1), w(2)}) = \begin{cases} -\underline{w(1)^T x_n} + \log(e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}), & y_n = 1 \\ -\underline{w(2)^T x_n} + \log(e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}), & y_n = 2 \end{cases}$$

Hence the gradient will depend on what the label y_n is

With respect to $\underline{w(1)}$

$y_n = 1$

$$\nabla_{\underline{w(1)}} e^{SM}(\underline{w(1), w(2)}) = -\underline{x_n} + \frac{\underline{x_n}\, e^{\underline{w(1)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}$$

$y_n = 2$

$$\nabla_{\underline{w(1)}} e^{SM}(\underline{w(1), w(2)}) = \frac{\underline{x_n}\, e^{\underline{w(1)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}$$

With respect to $\underline{w(2)}$

$y_n = 1$

$$\nabla_{\underline{w(2)}} e^{SM}(\underline{w(1), w(2)}) = \frac{\underline{x_n}\, e^{\underline{w(2)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}$$

$y_n = 2$

$$\nabla_{\underline{w(2)}} e^{SM}(\underline{w(1), w(2)}) = -\underline{x_n} + \frac{\underline{x_n}\, e^{\underline{w(2)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}$$

Combine together, for $i \in \{1, 2\}$

$$\nabla_{\underline{w(i)}} e^{SM}(\underline{w(1), w(2)}) = \begin{cases} -\underline{x_n} + \dfrac{\underline{x_n}\, e^{\underline{w(i)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}, & y_n = i \\ \dfrac{\underline{x_n}\, e^{\underline{w(i)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}, & y_n \neq i \end{cases}$$

Some further simplifications

$$\nabla_{\underline{w(1)}} e^{SM}(\underline{w(1), w(2)}) = \begin{cases} \dfrac{-\underline{x_n}\, e^{\underline{w(2)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}, & y_n = 1 \\ \dfrac{\underline{x_n}\, e^{\underline{w(1)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}, & y_n = 2 \end{cases}$$

$$\nabla_{\underline{w(2)}} e^{SM}(\underline{w(1), w(2)}) = \begin{cases} \dfrac{\underline{x_n}\, e^{\underline{w(2)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}, & y_n = 1 \\ \dfrac{-\underline{x_n}\, e^{\underline{w(1)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}, & y_n = 2 \end{cases}$$

Binary Logistic Regression

$$p^{LR}(y=1\,|\,\underline{x}) = \frac{e^{\underline{w^T x}}}{1+e^{\underline{w^T x}}}$$

$$p^{LR}(y=2\,|\,\underline{x}) = 1 - p^{LR}(y=1\,|\,\underline{x}) = 1 - \frac{e^{\underline{w^T x}}}{1+e^{\underline{w^T x}}}$$

$$= \frac{1}{1+e^{\underline{w^T x}}}$$

Loss Function

$$e_n^{LR}(\underline{w}) = -\log(p^{LR}(y_n\,|\,\underline{x_n}))$$

Part (b),

$$p^{SM}(y=i\,|\,\underline{x}) = \frac{e^{\underline{w(i)^T x}}}{e^{\underline{w(1)^T x}} + e^{\underline{w(2)^T x}}}$$

$$p^{LR}(y=1\,|\,\underline{x}) = \frac{e^{\underline{w^T x}}}{1+e^{\underline{w^T x}}} = \frac{1}{1+e^{\underline{-w^T x}}}$$

$$p^{SM}(y=i\,|\,\underline{x}) = \frac{1}{e^{\underline{(w(1)-w(i))^T x}} + e^{\underline{(w(2)-w(i))^T x}}}$$

Plug in $i=1$

$$p^{SM}(y=1\,|\,\underline{x}) = \frac{1}{1+e^{\underline{(w(2)-w(1))^T x}}}$$

$P^{SM}(y=1 \mid x) = P^{LR}(y=1 \mid x)$

Is only true when $\underline{-w = w(2) - w(1)}$

$\underline{w = w(1) - w(2)}$

$y_n = 1$

$$\underline{e_n^{LR}(w)} = -\log \frac{e^{w^Tx}}{1+e^{w^Tx}}$$

$$= -\left[\log e^{w^Tx} - \log(1+e^{w^Tx})\right]$$

$$= -\underline{w^Tx} + \log(1+e^{w^Tx})$$

$$\nabla_w e_n^{LR}(w) = -x + \frac{xe^{w^Tx}}{1+e^{w^Tx}} = \frac{-x_n}{1+e^{w^Tx}}$$

$y_n = 2$

$$\underline{e_n^{LR}(w)} = -\log \frac{1}{1+e^{w^Tx}}$$

$$= -\left[\log 1 - \log(1+e^{w^Tx})\right]$$

$$= \log(1+e^{w^Tx})$$

$$\nabla_w e_n^{LR}(w) = \frac{xe^{w^Tx}}{1+e^{w^Tx}}$$

Part (c),

Softmax Regression Model

$$y_n=1: \begin{cases} w(1)_{k+1} = w(1)_k - \varepsilon^{SM} \nabla_{w(1)} e_n^{SM}(\underline{w(1), w(2)}) \\ w(2)_{k+1} = w(2)_k - \varepsilon^{SM} \nabla_{w(2)} e_n^{SM}(\underline{w(1), w(2)}) \end{cases}$$

$$y_n=2: \begin{cases} w(1)_{k+1} = w(1)_k - \varepsilon^{SM} \nabla_{w(1)} e_n^{SM}(\underline{w(1), w(2)}) \\ w(2)_{k+1} = w(2)_k - \varepsilon^{SM} \nabla_{w(2)} e_n^{SM}(\underline{w(1), w(2)}) \end{cases}$$

Binary Logistic Regression

$$y_n=1: \underline{w_{k+1}} = \underline{w_k} - \varepsilon^{LR} \underline{\nabla_w} e_n^{LR}(\underline{w})$$

$$y_n=2: \underline{w_{k+1}} = \underline{w_k} - \varepsilon^{LR} \underline{\nabla_w} e_n^{LR}(\underline{w})$$

$$\underline{w_k = w(1)_k - w(2)_k}$$

Focusing on $y_n=1$, For Softmax Regression Model

$$\underbrace{w(1)_{k+1} - w(2)_{k+1}}_{w_{k+1}} = \underbrace{w(1)_k - w(2)_k}_{w_k} - \varepsilon^{SM}$$
$$\left(\nabla_{w(1)} e_n^{SM}(\underline{w(1), w(2)}) - \nabla_{w(2)} e_n^{SM}(\underline{w(1), w(2)}) \right)$$

$$\nabla_{w(1)} e_n^{SM}(\underline{w(1), w(2)}) - \nabla_{w(2)} e_n^{SM}(\underline{w(1), w(2)})$$

$$= \frac{-\underline{x_n}\, e^{\underline{w(2)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}} - \frac{\underline{x_n}\, e^{\underline{w(2)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}}$$

$$= \frac{-2\,\underline{x_n}\, e^{\underline{w(2)^T x_n}}}{e^{\underline{w(1)^T x_n}} + e^{\underline{w(2)^T x_n}}} = \frac{-2\,\underline{x_n}}{e^{\underline{(w(1)-w(2))^T x_n}} + 1}$$

$$= \frac{-2\underline{x_n}}{1 + e^{\underline{w^T x_n}}}$$

Hence, When $y_n = 1$

$$\underline{w_{k+1}} = w_k + \varepsilon^{SM} \frac{2\underline{x_n}}{1 + e^{\underline{w^T x_n}}}$$

Focusing on $y_n = 1$, For Binary Logistic Regression

$y_n = 1$: $\underline{w_{k+1}} = \underline{w_k} - \varepsilon^{LR} \underline{\nabla_w e_n^{LR}}(\underline{w})$

$$\underline{\nabla_w} e_n^{LR}(\underline{w}) = \frac{-x_n}{1+e^{\underline{w^T x}}} \quad \text{when } y_n = 1$$

$$w_{k+1} = w_k + \varepsilon^{LR} \frac{\underline{x_n}}{1+e^{\underline{w^T x}}}$$

Thus,

$$\boxed{2\varepsilon^{SM} = \varepsilon^{LR}}$$

Same conclusion can be made for $y_n = 2$

Question 2,

Scalar Valued Function $f: \mathbb{R}^n \longrightarrow \mathbb{R}$

Vector $p \in \mathbb{R}^n$

First Order Taylor Approximation Of $f(x+p)$

$$f(x+p) \approx f(x) + \nabla f(x)^T p$$

$$E(u,v) = e^u + e^{2v} + e^{uv} + u^2 - 3uv + 4v^2 - 3u - 5v$$

u and v are scalars

Part (a), $(u,v) = (0,0)$

$$\hat{E}_1(\Delta u, \Delta v) \approx E(0,0) + \nabla E(u,v)^T \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$$

$$= e^0 + e^0 + e^0 + 0^2 - 3(0)(0) + 4(0)^2 - 3(0) - 5(0) +$$

$$\left[(e^u + ve^{uv} + 2u - 3v - 3) \quad (2e^{2v} + ue^{uv} - 3u + 8v - 5) \right]_{(0,0)} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$$

$$= 3 + \begin{bmatrix} (-2) & (-3) \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$$

$$= 3 - 2\Delta u - 3\Delta v$$

$$\boxed{\begin{aligned} a &= 3 \\ a_u &= -2 \\ a_v &= -3 \end{aligned}}$$

Part (b),

$$\left\|(\Delta u, \Delta v)\right\|_2 = \sqrt{\Delta u^2 + \Delta v^2} = 0.5$$

$$\Delta u^2 + \Delta v^2 = 0.25$$

$$-\frac{\nabla E(u,v)}{\left\|\nabla E(u,v)\right\|_2} = \frac{1}{\sqrt{2^2+3^2}}\begin{vmatrix} 2 \\ 3 \end{vmatrix} = \begin{vmatrix} \frac{2}{\sqrt{13}} \\ \frac{3}{\sqrt{13}} \end{vmatrix} = \Delta z$$

$$\Delta \hat{z} = 0.5 \Delta z = \begin{vmatrix} \frac{1}{\sqrt{13}} \\ \frac{1.5}{\sqrt{13}} \end{vmatrix}$$

$$\hat{E}_1(\Delta \hat{z}_1, \Delta \hat{z}_2) = 3 - 2\left(\frac{1}{\sqrt{13}}\right) - 3\left(\frac{1.5}{\sqrt{13}}\right)$$

$$= 1.197$$