Briana Zhao
U93987617
EC414
HW 9

1

(1.1) $\nabla_{\hat{y}} L(\hat{y}, y) = \nabla_{\hat{y}}\left(-y\ln(\hat{y}) - (1-y)\ln(1-\hat{y})\right)$

$$= \frac{-y}{\hat{y}} - \frac{(-1)(1-y)}{(1-\hat{y})} = \frac{-y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})} = \frac{-y(1-\hat{y}) + \hat{y}(1-y)}{\hat{y}(1-\hat{y})}$$

$$= \frac{-y + \hat{y}y + \hat{y} - \hat{y}y}{\hat{y}(1-\hat{y})} = \frac{-y+\hat{y}}{\hat{y}(1-\hat{y})}$$ and since $h^{(2)} = \hat{y}$,

$$\boxed{\nabla_{\hat{y}} L(\hat{y}, y) = \frac{-y+\hat{y}}{\hat{y}(1-\hat{y})} = \frac{-y+h^{(2)}}{h^{(2)}(1-h^{(2)})}}$$

(1.2) $\nabla_{a^{(2)}} J = g \odot f'(a^{(2)})$ $\quad f'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{(1+e^{-z}-1)}{(1+e^{-z})} \cdot \frac{1}{(1+e^{-z})} = (1-f(z))f(z)$

so $g \odot f'(a^{(2)}) = g \odot f(a^{(2)})(1-f(a^{(2)}))$

$= g \odot h^{(2)}(1-h^{(2)})$ since $h^{(2)} = f(a^{(2)})$

$= \underbrace{\frac{-y+h^{(2)}}{h^{(2)}(1-h^{(2)})}}_{g \text{ from } 1.1} \odot h^{(2)}(1-h^{(2)})$ so $\boxed{\nabla_{a^{(2)}} J = h^{(2)} - y}$

(1.3) $\nabla_{b^{(2)}} J = g$

from 1.2, $g = \nabla_{a^{(2)}} J$ so $\boxed{\nabla_{b^{(2)}} J = h^{(2)} - y}$

(1.4) $\nabla_{W^{(2)}} J = g h^{(2-1)T} = g h^{(1)T}$ so $\boxed{\nabla_{W^{(2)}} J = (h^{(2)} - y) \cdot h^{(1)T}}$

(1.5) $\nabla_{h^{(1)}} J = W^{(2)T} g$

$g = h^{(2)} - y$ so $\boxed{\nabla_{h^{(1)}} J = W^{(2)T} \cdot (h^{(2)} - y)}$

(1.6) $g \leftarrow \nabla_{a^{(1)}} J = g \odot f'(a^{(1)})$

$$= g \odot f(a^{(1)})(1 - f(a^{(1)})) = g \odot h^{(1)}(1 - h^{(1)})$$

from 1.4, we had $g \leftarrow \nabla_{h^{(1)}} J = W^{(2)T}(h^{(2)} - y)$

so $g \odot h^{(1)}(1 - h^{(1)}) = W^{(2)T}(h^{(2)} - y) \odot h^{(1)}(1 - h^{(1)}) = \nabla_{a^{(1)}} J = g$

$\nabla_{b^{(1)}} J = g$ so $\boxed{\nabla_{b^{(1)}} J = W^{(2)T}(h^{(2)} - y) \odot h^{(1)}(1 - h^{(1)})}$

$\nabla_{W^{(1)}} J = g h^{(1-1)T}$ so $\boxed{\nabla_{W^{(1)}} J = \left[W^{(2)T}(h^{(2)} - y) \odot h^{(1)}(1 - h^{(1)})\right] \cdot h^{(0)T}}$

where $h^{(0)} = x$, $h^{(1)} = \sigma(b^{(1)} + W^{(1)}x)$, $h^{(2)} = \sigma(b^{(2)} + W^{(2)}\sigma(b^{(1)} + W^{(1)}x))$

2.2.2 $f = \frac{1}{1+e^{-w^T x}} = (1 + e^{-w^T x})^{-1}$

$\frac{df}{dx} = (-w^T e^{-w^T x}) \cdot -1(1 + e^{-w^T x})^{-2}$ so $\boxed{\frac{df}{dx} = \frac{w^T e^{-w^T x}}{(1 + e^{-w^T x})^2}}$

$\frac{df}{dw} = (-x e^{-w^T x}) \cdot -1(1 + e^{-w^T x})^{-2}$ so $\boxed{\frac{df}{dw} = \frac{x e^{-w^T x}}{(1 + e^{-w^T x})^2}}$

$$w^T x = \begin{bmatrix} 2 & -3 & -3 \end{bmatrix} \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix} = -2 + 6 - 3 = 1$$

so $\frac{w^T e^{-w^T x}}{(1 + e^{-w^T x})^2} = \frac{w^T e^{-1}}{(1 + e^{-1})^2} = \frac{.367879}{1.87109} \begin{bmatrix} 2 \\ -3 \\ -3 \end{bmatrix} = \begin{bmatrix} 0.3932 \\ -0.5898 \\ -0.5898 \end{bmatrix}$ ✓

and $\frac{x e^{-w^T x}}{(1 + e^{-w^T x})^2} = \frac{x e^{-1}}{(1 + e^{-1})^2} = \frac{0.367879}{1.87109} \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.1966 \\ -0.3932 \\ 0.1966 \end{bmatrix}$ ✓