①

$X = \{[0\ 0]^T, [1\ 0]^T, [0\ 1]^T, [1\ 1]^T\}$

$y = \underset{x \in X}{f^*(x)} = \{0, 1, 1, 0\}$

$J(\theta) = \frac{1}{4} \sum_{x \in X} [f^*(x) - f(x;\theta)]^2$

for linear function, $f(x;\theta) = f(x;w,b) = x^T w + b$. $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

$\therefore J(\theta) = J(w,b) = \frac{1}{4} \sum_{x \in X} \{f^*(x) - f(x;w,b)\}^2 = \frac{1}{4} \sum_{x \in X} \{f^*(x) - x^T w - b\}^2$

To minimize $J(\theta)$, we find θ s.t. $\nabla_\theta (J(\theta)) = 0 = \nabla_{w,b} (J(w,b))$

$\nabla_w (J) = \frac{1}{4} \nabla_w \sum_{x \in X} \{f^*(x) - x^T w - b\}^2$

$= \frac{1}{4} \nabla_w \Big[\Big(0 - [0\ 0]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\Big)^2 + \Big(1 - [0\ 1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\Big)^2 + \Big(1 - [1\ 0]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\Big)^2 + \Big(0 - [1\ 1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\Big)^2 \Big]$

$= \frac{1}{4} \nabla_w \Big[(0-b)^2 + (1 - w_2 - b)^2 + (1 - w_1 - b)^2 + (0 - w_1 - w_2 - b)^2 \Big]$

$= \frac{1}{4} \cdot \Big[2.(0-b).0 + 2.(1-w_2-b).\nabla_w (1-w_2-b) + 2.(1-w_1-b).\nabla_w (1-w_1-b) + 2.(-w_1-w_2-b).\nabla_w (-w_1-w_2-b) \Big]$

$= \frac{2}{4} \Big[(1-w_2-b)\begin{bmatrix} 0 \\ -1 \end{bmatrix} + (1-w_1-b)\begin{bmatrix} 1 \\ 0 \end{bmatrix} + -(w_1+w_2+b)\begin{bmatrix} -1 \\ -1 \end{bmatrix} \Big]$

$$= \frac{1}{2} \cdot \begin{bmatrix} (0) + (1 - w_1 - b) + (w_1 + w_2 + b) \\ (0) + (1 - w_2 - b) + (w_1 + w_2 + b) \end{bmatrix}$$

$$= \frac{1}{2} \cdot \begin{bmatrix} w_2 \\ w_1 \end{bmatrix}$$

Setting to zero, we have, $\boxed{w_1 = w_2 = 0.}$

For bias,

$$\nabla_b (J) = \frac{1}{4} \nabla_b \sum_{x \in X} \{ f^*(x) - x^T w - b \}^2$$

$$= \frac{1}{4} \cdot \nabla_b \left[b^2 + (1 - w_2 - b)^2 + (1 - w_1 - b)^2 + (w_1 + w_2 + b)^2 \right]$$

$$= \frac{1}{4} \left[2b + 2(1 - w_2 - b)(-1) + 2(1 - w_1 - b)(-1) + 2(w_1 + w_2 + b) \right]$$

$$= \frac{1}{4} \left[2b - 2 + 2w_2 + 2b - 2 + 2w_1 + 2b + 2w_1 + 2w_2 + 2b \right]$$

$$= \frac{1}{4} \left[8b - 4 \right] \quad \ldots \text{ setting } w_1 = w_2 = 0.$$

$$= 2b - 1$$

Setting $\nabla_b (J) = 0$, $\quad 2b - 1 = 0$

$$\Rightarrow b = \frac{1}{2} \quad \Rightarrow \boxed{b = 0.5}$$

Thus, for XOR problem, $w = [0 \ 0]^T$ & $b = 0.5$ minimizes the MSE loss function.

$$\hat{y} = x^T w + b = [x_1 \ x_2] \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + 0.5 = [x_1 \ x_2] \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.5 = \underline{0.5} \ \forall x.$$

■

(4)

$$\hat{y} = \sigma(x^T w + b)$$

where, $\sigma = \frac{1}{1 - e^{-x}}$.

$$\therefore \hat{y} = \frac{1}{1 - e^{-(x^T w + b)}}$$

Given,

training examples → {+ve}
testing examples → {+ve, -ve}

We know that, $\sigma'(x) = \frac{\partial \sigma}{\partial x}(x) = \sigma(x) \cdot (1 - \sigma(x)) \quad \forall x$

Now, $\hat{y} = f(x; w, b) = \sigma(x^T w + b)$

Consider MSE → $\frac{1}{n} \sum_{x \in X} \{f^*(x) - f(x; w, b)\}^2$

where, $f^*(x)$ → desired predictions.

Say, we have one example, i.e., size(x) = 1.

$\therefore \nabla_w (J(w)) = 0$ (to minimize MSE loss)

$$= 2 \cdot \{f^*(x) - f(x; w, b)\} \cdot [\nabla (f^*(x)) - \nabla (f(x; w, b))]$$

$$= -2 \{f^*(x) - f(x; w, b)\} \cdot \sigma'(x^T w + b) .$$

Now,

$$\sigma'(x^T w + b) = \sigma(x^T w + b) \cdot (1 - \sigma(x^T w + b))$$

$$= \frac{1}{1 + e^{-z}} \left(1 - \frac{1}{1+e^{-z}}\right)$$

$$= \frac{1+e^{-z} - 1}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$= \frac{1}{(1+e^{z})(1+e^{-z})} .$$

$$* = \frac{e^{z}}{(1+e^{z})^2}$$

$(z = x^T w + b)$

(i) Bias term, b → converges
because it is scalar & independant of training & testing examples.

(ii) Weights, w → converge, to minimize log loss. Regardless of values.

(iii) Training loss → converges
the model tries to minimize log loss on training examples.

(iv) Testing loss → depends on testing values.

Q-3

(i) for $l = K$:- $\nabla_{w^{(l)}} \hat{y}_K^{(i)} = \nabla_{w^{(l)}} \frac{e^{z_l}}{\sum_j e^{z_j}}$

$$= \frac{\left(\sum_j e^{z_j}\right) \nabla_{w^{(l)}} e^{z_l} - e^{z_l} \nabla_w \left(\sum_j e^{z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$= \frac{\left(\sum_j e^{z_j}\right) e^{z_l} \nabla_{w^{(l)}} \left(X^T w^{(l)} + b_K\right) - \left(\sum_j e^{z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$= \frac{\left(\sum_j e^{z_j}\right) e^{z_l} \nabla_{w^{(l)}} \left(X^T w^{(l)} + b_K\right) - e^{z_l} \nabla_{w^{(l)}} \left(e^{z_l}\right)}{\left(\sum e^{z_j}\right)^2}$$

$$= \frac{\left(\sum_j e^{z_j}\right) e^{z_l} X^{(i)} - e^{z_l} X^{(i)} e^{z_l}}{\left(\sum_j e^{z_j}\right)^2}$$

only vector part here is $X^{(i)}$:

$$\nabla_{w^{(l)}} \hat{y}_K^{(i)} = X^{(i)} \left(\frac{e^{z_l}}{\sum_j e^{z_j}}\right) \left(\frac{1 - e^{z_l}}{\sum_j e^{z_j}}\right)$$

$$\boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \hat{y}_l^{(i)} \left(1 - \hat{y}_l^{(i)}\right)}$$

(ii) Now, $l \neq k$,

$$\nabla_{w^l} \hat{y}_k^{(i)} = \nabla_w \left(\frac{e^{z_k}}{\sum_j e^{z_j}}\right) = \frac{\left(\sum_j e^{z_j}\right) \nabla_{w^l} e^{z_k} - e^{z_k} \nabla_{w^l} \left(\sum_j e^{z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$= \frac{\left(\sum_j e^{z_j}\right) e^{z_k} (0) - e^{z_k} \left(\nabla_{w^l} e^{z_l}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$= \frac{-e^{z_k}}{\left(\sum_j e^{z_j}\right)^2} \; e^{z_l} \nabla_{w^l} \left(X^T w^l + b_k\right)$$

$$= -\frac{e^{z_k}}{\left(\sum_j e^{z_j}\right)} \cdot \frac{e^{z_l}}{\left(\sum_j e^{z_j}\right)} \cdot x^{(i)}$$

$$= -\hat{y}_k^{(i)} \cdot \hat{y}_k^{(i)} \cdot x^{(i)}$$

$$= \boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = -x^{(i)} \hat{y}_k^{(i)} \hat{y}_l^{(i)}}$$

(iii) $\nabla_{w^{(l)}} f_{CE}(W,b) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{K=1}^{c} y_K^{(i)} \nabla_{w^{(l)}} \log \hat{y}_K^{(i)}$

$= -\frac{1}{n} \sum_{i=1}^{n} \sum_{K=1}^{c} y_K^{(i)} \left(\frac{\nabla_{w^{(l)}} \hat{y}_K^{(i)}}{\hat{y}_K^{(i)}} \right)$

$= -\frac{1}{n} \left(\sum_{i=1}^{n} \left(\sum_{K \neq l} y_K^{(i)} \left(\frac{-x^{(i)} \hat{y}_K^{(i)} \hat{y}_l^{(i)}}{\hat{y}_K^{(i)}} \right) \right. \right.$

$\left. \left. y_l^{(i)} \left(\frac{x^{(i)} \hat{y}_l^{(i)} (1 - \hat{y}_l^{(i)})}{\hat{y}_l^{(i)}} \right) \right) \right)$

$= -\frac{1}{n} \left(\sum_{i=1}^{n} \left[x^{(i)} y_l^{(i)} (1 - \hat{y}_l^{(i)}) - x^{(i)} \hat{y}_l^{(i)} \sum_{K \neq l} y_K^{(i)} \right] \right)$

$= -\frac{1}{n} \sum_{i=1}^{n} \left[x^{(i)} y_l^{(i)} - x^{(i)} y_l^{(i)} \hat{y}_l^{(i)} - x^{(i)} \hat{y}_l^{(i)} \left(\sum_K y_K^{(i)} - y_l^{(i)} \right) \right]$

$= -\frac{1}{n} \sum_{i=1}^{n} \left[x^{(i)} y_l^{(i)} - x^{(i)} \hat{y}_l^{(i)} \right]$

$$\boxed{\nabla_{w^{(l)}} f_{CE}(W,b) = -\frac{1}{n} \sum_{i=1}^{n} x^{(i)} (y^{(i)} - \hat{y}^{(i)})}$$

$\nabla_b f_{CE}(W,b) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{K=1}^{c} y_K^{(i)} \nabla_b (\log \hat{y}_K^{(i)})$

$= -\frac{1}{n} \sum_{i=1}^{n} \sum_{K=1}^{c} y_K^{(i)} \left(\frac{\nabla_b \hat{y}_K^{(i)}}{\hat{y}_K^{(i)}} \right) \quad \cdots\cdots (1)$

Now, $\nabla_b \hat{y}_K^{(i)} = \nabla_b \frac{e^{z_K}}{\sum_j e^{z_j}} = \frac{\left(\sum_j e^{z_j}\right) \nabla_b e^{z_K} - e^{z_K} \nabla_b \left(\sum_j e^{z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$

Now, the term $\nabla_b e^{z_k} = e^{z_k} \begin{bmatrix} \partial/\partial b_1 \\ \partial/\partial b_2 \\ \vdots \\ \partial/\partial b_k \\ \vdots \\ \partial/\partial b_c \end{bmatrix} (x^T w^{(k)} + b_k)$

$$\nabla_b e^{z_k} = e^{z_k} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \rightarrow k^{th} \text{ position}$$

$$\therefore \text{As } \nabla_b \hat{y}_k^{(i)} = \frac{\left(\sum_j e^{z_j}\right) \nabla_b e^{z_k} - e^{z_k} \left(\sum_j \nabla_b e^{z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$\therefore \nabla_b \hat{y}_k^{(i)} = \left(\frac{\sum_j e^{z_j}}{\left(\sum_j e^{z_j}\right)^2}\right) \cdot e^{z_k} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix} \rightarrow k+n^{th} \text{ position} - \frac{e^{z_k}}{\left(\sum_j e^{z_j}\right)^2} \sum_j e^{z_j} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \leftarrow j+n^{th} \text{ position}$$

$$\therefore \nabla_b \hat{y}_k^{(i)} = \frac{e^{z_k}}{\left(\sum_j e^{z_j}\right)} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \rightarrow k^{th} \text{ position} - \frac{e^{z_k}}{\left(\sum_j e^{z_j}\right)^2} \begin{bmatrix} e^{z_1} \\ e^{z_2} \\ \vdots \\ e^{z_c} \end{bmatrix}$$

$$= \hat{y}_k^{(i)} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \rightarrow k^{th} \text{ position} - \hat{y}_k^{(i)} \begin{bmatrix} e^{z_1} / (e^{z_1})^2 \\ e^{z_2} / (e^{z_2})^2 \\ \vdots \\ e^{z_c} / e^{(z_c)^2} \end{bmatrix}$$

$$\therefore \frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \rightarrow k^{th} \text{ position} - \begin{bmatrix} \hat{y}_1^{(i)} \\ \hat{y}_2^{(i)} \\ \vdots \\ \hat{y}_c^{(i)} \end{bmatrix}$$

$$\frac{y_k^{(i)} \nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ y_k^{(i)} \\ \vdots \\ 0 \end{bmatrix} - y_k^{(i)} \begin{bmatrix} \hat{y}_1^{(i)} \\ \hat{y}_2^{(i)} \\ \vdots \\ \hat{y}_c^{(i)} \end{bmatrix}$$

k^{th} position

$$\therefore \sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} \right) = \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ \vdots \\ y_c^{(i)} \end{bmatrix} - \left(\sum_{k=1}^{c} y_k^{(i)} \right) \begin{bmatrix} \hat{y}_1^{(i)} \\ \hat{y}_2^{(i)} \\ \vdots \\ \hat{y}_c^{(i)} \end{bmatrix}$$

$$\therefore \sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} \right) = \vec{y}^{(i)} - (1)\, \vec{\hat{y}}^{(i)}$$

$$\therefore -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} \right) \cdots \frac{1}{n} \sum_{i=1}^{n} \left(\vec{y}^{(i)} - \vec{\hat{y}}^{(i)} \right)$$

from ①

$$\therefore \nabla_b f_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \left(\vec{y}^{(i)} - \vec{\hat{y}}^{(i)} \right)$$