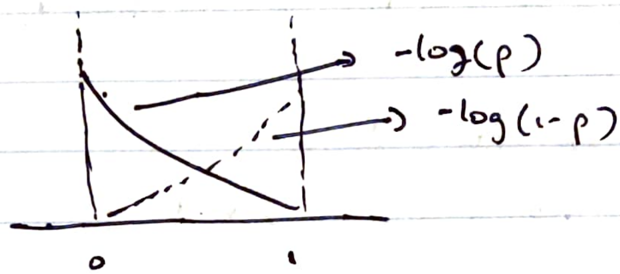
SHAMEKH SIDDIQUI

HW-04

(1) a) cross entropy loss $= -y\log(p) - (1-y)\log(1-p)$

$y=0 \Rightarrow \text{loss} = -\log(1-p)$

$y=1 \Rightarrow \text{loss} = -\log(p)$



- This is better than MSE as we do not have the issue of vanishing gradient.
- Here output is in range of [0,1], classification of input is based on the threshhold of the probability of data belonging to particular class.

b) Back propogation cost function:

$$L_\lambda(w,b) = \frac{1}{2}\sum_{i=1}^{N} \text{loss}\left(h(x_i), y^i\right) + \frac{\lambda}{2}\sum_{l=1}^{n-1} \|w^{(l)}\|_F^2$$

considering BCE for each layer

$$L_\lambda\{w^{(l)}, b^{(l)}\}_{l=1}^{n-1} = -\frac{1}{2}\sum_{j=1}^{S_n} y_j \log\left(a_j^{(n)}(x^i)\right) +$$

$$(1-y_j^i)\log(1-a_j^{(n)}(x^i))) + \lambda/2 \sum_{l=1}^{n-1} \|w^{(l)}\|_F^2$$

using gradient descent :

i) initialize $w^{(l)} \in R^{s_{l+1} \times s_l}$, $b^{(l)} \in s^{l+1}$

ii) Until convergence do,

for $l = 1, 2, \ldots, n-1$

$$a)\ w^{(l)(t+1)} \leftarrow w^{(l)t} - \rho \frac{\partial L_\lambda}{\partial w^{(l)}}\bigg|_{w^{(l)(t)}}$$

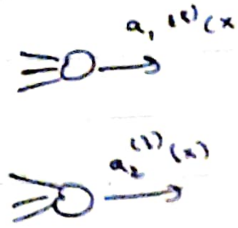
$$b)\ b^{(l)(t+1)} \leftarrow b^{(l)(t)} - \rho \frac{\partial L_\lambda}{\partial b^{(l)}}\bigg|_{b^{(l)(t)}}$$

}

$$\therefore \frac{\partial L}{\partial w^{(l)}} = \frac{-1}{2}\sum_{i=1}^{N} \frac{y_j^i \log(a_j^{(n)}(x^i)) + (1-y_j^i)\log(1-a_j^{(n)}(x^i))}{\partial w^{(l)}}$$

$$+ \lambda w^{(l)}$$

$$\frac{\partial L}{\partial b^{(l)}} = \frac{-1}{2}\sum_{i=1}^{N} \frac{y_j^i \log(a_j^{(n)}(x^i)) + (1-y_j^i)\log(1-a_j^{(n)}(x^i))}{\partial b^{(l)}}$$



$$\begin{bmatrix} a_1^{(3)}(x) \\ a_2^{(3)}(x) \end{bmatrix} = h(x^i) \text{ . and } \begin{bmatrix} y_1^i \\ y_2^i \end{bmatrix} = y^i$$

$$\delta_i^{(n)} \triangleq \frac{\partial J}{\partial z_i^{(n)}} = \frac{\partial - \frac{1}{2} \sum_j y_j \log(a_j^{(n)}(x)) + (1-y_j) \log(1-a_j^{(n)}(x))}{\partial z_i^{(n)}} \longrightarrow (1)$$

$$\delta_i^{(n)} \triangleq \frac{\partial J}{\partial z_i^{(n)}} = \frac{\partial - \frac{1}{2}[y_i \log(a_i^{(n)}) + (1-y_i)\log(1-a_i^{(n)})]}{\partial z_i^{(n)}}$$

$$= -\frac{1}{2} \left[\frac{\partial y_i \log(a_j^{(n)})}{\partial z_i^{(n)}} + \frac{\partial (1-y_i)\log(1-a_i^{(n)})}{\partial z_i^{(n)}} \right]$$

$$a_i^{(n)} = f(z_i^{(n)})$$

$$\therefore \delta_i^{(n)} = -\frac{1}{2} \left[\frac{\partial y_i \log(f(z_i^{(n)}))}{\partial z_i^{(n)}} + \frac{\partial (1-y_i)\log(1-f(z_i^{(n)}))}{\partial z_i^{(n)}} \right]$$

$$\delta_i = -\frac{1}{2} \left[y_i \frac{f'(z_i^{(n)})}{f(z_i^{(n)})} + (1-y_i) \frac{-f'(z_i(n))}{1-f(z_i^{(n)})} \right] \longrightarrow (2)$$

diff ② w.r.t $z^{(n-1)}$

$$\therefore \delta_j^{(n-1)} \triangleq \frac{\partial J}{\partial z_j^{(n-1)}} = \sum_{i=1}^{S_n} \frac{\partial J}{\partial z_i^{(n)}} \times \frac{\partial z_i^{(n)}}{\partial z_j^{(n-1)}}$$

$$= \delta_j^{(n)} \times \frac{\partial z_i^{(n)}}{\partial z_j^{(n-1)}}$$

$$z_i^{(n)} = \omega_{11}^{(n-1)} \times a_{11}^{(n-1)} + \cdots + \omega_{ij}^{(n-1)} a_{ij}^{(n-1)} + \cdots$$

$$\omega_{is_{n-1}}^{(n-1)} \cdot a_{s_{n-1}}^{(n-1)}$$

$$a_{i1}^{(n-1)} = f(z_1^{(n-1)})$$

$$\therefore \frac{\partial z_i^{(n)}}{\partial z_j^{(n-1)}} = \omega_{ij}^{(n-1)} \times \frac{\partial a_{ij}^{(n-1)}}{\partial z_j}$$

$$= \omega_{ij}^{(n-1)} \times \frac{\partial f(z_j^{(n-1)})}{\partial z_j}$$

$$\frac{\partial z_i^{(n)}}{\partial z_j^{(n-1)}} = \omega_{ij}^{(n-1)} \times f'(z_j^{(n-1)})$$

$$\boxed{\delta_j^{(n-1)} \triangleq \frac{\partial J}{\partial z_j^{(n-1)}} = \sum_{i=1}^{s_n} \delta_i^{(n)} \times w_{ij}^{(n-1)} \times f'(z_j^{(n-1)})} \quad \text{③}$$

generalize derivatives for layer l,

$$\boxed{\sum_{i=1}^{s_l} \delta_i^{(l)} \cdot w_{ij}^{(l-1)} \cdot f'(z_j^{(l-1)}) = \delta_j^{(l-1)} \triangleq \frac{\partial J}{\partial z_j^{(l-1)}}}$$

diff. J w.r.t. w_{ij} & b_i

$$\frac{\partial J}{\partial w_{ij}^{(l)}} = \frac{\partial J}{\partial z_i^{(l+1)}} \times \frac{\partial z_i^{(l+1)}}{\partial w_{ij}^{(l)}}$$

$$\left[z_i^{(l+1)} = \text{---} + w_{ij}^{(l)} a_j^{(l)} + \text{---} \; b_i^{(l)} \quad \text{\& using ① \& ②} \right]$$

$$\boxed{\frac{\partial J}{\partial w_{ij}} = \delta_i^{(l+1)} \cdot a_j^{(l)}}$$

similarly
$$\frac{\partial J}{\partial b_i^{(l)}} = \frac{\partial J}{\partial z_i^{(l+1)}} \times \frac{\partial z_i^{(l+1)}}{\partial b_i^{(l)}}$$

$$\frac{\partial J}{\partial b_i^{(l)}} = \delta_i^{(l+1)} \times 1$$

$$\frac{\partial J}{\partial w^{(l)}} = \begin{bmatrix} \vdots & \vdots & \\ \frac{\partial J}{\partial w^{(l)}_{i1}} & \frac{\partial J}{\partial w^{(l)}_{i2}} & \cdots \\ \vdots & \vdots & \end{bmatrix} = \begin{bmatrix} \vdots & \vdots \\ \delta_i^{(l+1)} a_1^{(l)} & \delta_2^{(l+1)} a_2^{(l)} \\ \vdots & \vdots \end{bmatrix}$$

$$= \begin{bmatrix} \vdots \\ \delta_i^{(l+1)} \\ \vdots \end{bmatrix} \begin{bmatrix} a_1^{(l)} & a_2^{(l)} & \cdots \end{bmatrix} = \delta^{(l+1)} \cdot a^{(l)^T}$$

$$\frac{\partial J}{\partial b^{(l)}} = \delta_i^{(l+1)} = \begin{bmatrix} \delta_1^{(l+1)} \\ \delta_2^{(l+1)} \\ \vdots \end{bmatrix} = \delta^{(l+1)}$$

c) First derivative of sigmoid function:

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

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

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

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

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

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

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

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

$$\sigma'(z) = \sigma(z) \times (1 - \sigma(z))$$

② a) function a_1 is at z_1

function a_2 is at z_2

$$z_1 = x_1 w_3 + x_2 w_5 + w_1$$

$$z_2 = x_1 w_4 + x_2 w_6 + w_2$$

Activation

$$a_1 = c z_1$$
$$= c(x_1 w_3 + x_2 w_5 + w_1)$$

$$a_2 = c z_2$$
$$= c(x_1 w_4 + x_2 w_6 + w_2)$$

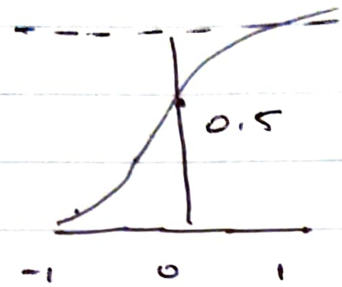
$$z_3 = w_7 + w_8 a_1 + w_9 a_2$$
$$= w_7 + w_8 c[x_1 w_3 + x_2 w_5 + w_1] + w_9 c[x_1 w_4 + x_2 w_6 + w_2]$$

$$= x_1 [c w_3 w_8 + c w_4 w_9] + x_2 [c w_5 w_8 + c w_6 w_9] + [w_7 + c w_1 w_8 + c w_2 w_9]$$

$$a_3 = P(y=1 \mid x, w) = \frac{1}{1+e^{-z_3}}$$

$$= \frac{1}{1 + e^{-\left(x_1 [c w_3 w_8 + c w_4 w_9] + x_2 [c w_5 w_8 + c w_6 w_9] + [w_7 + c w_1 w_8 + c w_2 w_9]\right)}}$$

Thus we get a sigmoid function ($\frac{1}{1+e^{x}}$)



→ sigmoid function, this gives us final classification boundary.

b) $z_3 = x_1 [cw_3 w_8 + cw_4 w_9]$

$+ x_2 [cw_5 w_8 + cw_6 w_9]$

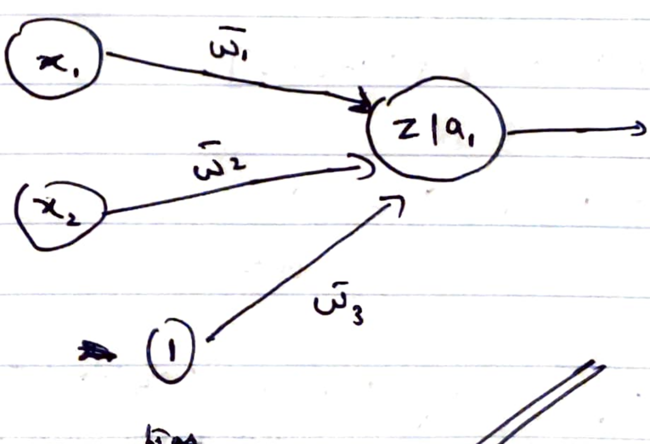
$+ [w_7 + cw_1 w_8 + cw_2 w_9]$

<u>weights</u>

$\bar{w}_1 = cw_3 w_8 + cw_4 w_9$

$\bar{w}_2 = cw_5 w_8 + cw_6 w_9$

$\bar{w}_3 = cw_1 w_8 + cw_2 w_9 + w_7$



c) In such a case, ~~the i~~ we get linear activation of the inputs. If we remove hidden layer we can still represent the multilayered neural network.

TRUE!

③ Assigning <u>linear</u> activations functions for <u>layer 1</u>:

$L(z_1) = c(x_1 w_1 + x_2 w_3)$ from $z_1 = x_1 w_1 + x_2 w_3$

$L(z_2) = c(x_1 w_2 + x_2 w_4)$ from $z_2 = x_1 w_2 + x_2 w_3$

Therefore we get $z_2 = w_5 [c(x_1 w_1 + x_2 w_3)] + w_6 [c(x_1 w_2 + x_2 w_4)]$

$= x_1 (c(w_1 w_5 + w_2 w_6)) + x_2 (c(w_3 w_5 + w_4 w_6))$

Assign <u>sigmoid function</u> for <u>layer 2</u>

$$s(z^{(2)}) = \frac{1}{1 + \exp(-x_1(c(w_1 w_5 + w_2 w_6)) - x_2(c(w_3 w_5 + w_4 w_6)))}$$

$$= \frac{1}{1 + \exp(x_1 \beta_1 + x_2 \beta_2)}$$

where

$$\beta_1 = -c(\omega_1\omega_5 + \omega_2\omega_6)$$

$$\beta_2 = -c(\omega_3\omega_5 + \omega_4\omega_6)$$