Part B

Name : Kshitij Alwadhi

Entry Num : 2019EE10577

(I) Gradient calculation (Activation functions)

(1) Sigmoid $\sigma(x) = \dfrac{1}{1+e^{-x}}$.

$$\frac{d\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$$= \frac{-1}{(1+e^{-x})^2} \times \frac{d}{dx}\left(e^{-x}\right)$$

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

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

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

$$\frac{d}{dx}\sigma(x) = (1 - \sigma(x)) \cdot \sigma(x)$$

Now consider,

$$\sigma(g(\omega))$$

$$\frac{\partial}{\partial \omega}\left(\sigma(g(\omega))\right) = \frac{\partial}{\partial g}\left(\sigma(g(\omega))\right) \cdot \frac{\partial g(\omega)}{\partial \omega}$$

$$= \sigma(g) \cdot (1-\sigma(g)) \cdot \frac{\partial g(\omega)}{\partial \omega}$$

② Hyperbolic Tangent

Representing this as:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$\frac{d}{dx}(\tanh(x)) = \frac{d}{dx}\left(\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}\right)$$

$$= \frac{(e^{x} + e^{-x}) \times (e^{x} + e^{-x}) - (e^{x} - e^{-x}) \cdot (e^{x} - e^{-x})}{(e^{x} + e^{-x})^2}$$

$$= 1 - \left(\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}\right)^2$$

$\longrightarrow \tanh(x)$

$$\frac{d}{dx}(\tanh(x)) = 1 - (\tanh(x))^2$$

Now consider,

$\tanh(g(\omega))$

$$\frac{d}{d\omega}\tanh(g(\omega)) = \frac{\partial}{\partial g} \times \left(\tanh(g(\omega))\right) \times \frac{\partial g(\omega)}{\partial \omega}$$

$$= \left(1 - (\tanh(g(\omega)))^2\right) \cdot \frac{\partial g(\omega)}{\partial \omega}$$

③ Relu

$$\mathrm{Relu}(x) = \begin{cases} x, & x \geq 0 \\ 0, & \text{o.w.} \end{cases}$$

$$\frac{\partial\, \mathrm{Relu}(x)}{\partial x} = \begin{cases} 1, & x \geq 0 \\ 0, & \text{o.w.} \end{cases}$$

Now consider,

$\mathrm{relu}(g(\omega))$

$$\frac{\partial}{\partial \omega}\mathrm{relu}(g(\omega)) = \frac{\partial}{\partial g}\left(\mathrm{relu}(g(\omega))\right) \cdot \frac{\partial g(\omega)}{\partial \omega}$$

$$= \begin{cases} \frac{\partial g(\omega)}{\partial \omega}, & g(\omega) \geq 0 \\ 0, & \text{o.w} \end{cases}$$

II Gradient calculation (Loss functions)

1 Cross - Entropy Loss

(*) Assumed binary classification for all.

$$L = -\sum_{i=1}^{n} y_i \cdot \ln \hat{y}_i + (1-y_i) \cdot \ln(1-\hat{y}_i)$$

$$\hat{y}_i = g(w, x_i)$$

$$\frac{\partial L}{\partial w} = -\sum_{i=1}^{n} y_i \cdot \frac{1}{\hat{y}_i} \cdot \frac{\partial g}{\partial w} + \frac{(1-y_i)}{(1-\hat{y}_i)} \cdot \left(-\frac{\partial g}{\partial w}\right)$$

$$\frac{\partial L}{\partial w} = -\sum_{i=1}^{n} \left(\frac{y_i}{\hat{y}_i} - \frac{(1-y_i)}{(1-\hat{y}_i)}\right) \cdot \left(\frac{\partial g}{\partial w}\right)$$

where $\hat{y}_i = g(w, x_i)$

② <u>Hinge Loss</u>

$x_i \rightarrow$ i^{th} input

$y_i \rightarrow$ i^{th} gold output

$$L = \sum_{i=1}^{n} l_i$$

$$l_i = \max(0, 1 - y_i \cdot \hat{y}_i)$$

$$\hat{y}_i = g(\omega, x_i)$$

$$L = \sum_{i=1}^{n} \max(0, 1 - y_i \cdot \hat{y}_i)$$

$$\frac{\partial L}{\partial \omega} = \sum_{i=1}^{n} \frac{\partial}{\partial \omega} \max(0, 1 - y_i \cdot \hat{y}_i)$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \omega} \max(0, 1 - y_i \cdot g(\omega, x_i))$$

Applying chain rule

$$= \sum_{i=1}^{n} \frac{\partial}{\partial g} \max(0, 1 - y_i \cdot g(\omega, x_i)) \cdot \frac{\partial g}{\partial \omega}$$

$$\frac{\partial}{\partial g} (\max(0, 1 - y_i \cdot g(\omega, x_i))) = -y_i$$

$$\frac{\partial L}{\partial \omega} = \sum_{i=1}^{n} \begin{cases} -y_i \cdot \dfrac{\partial g}{\partial \omega}, & y_i \cdot g(x_i, \omega) < 1 \\ 0, & \text{o.w} \end{cases}$$

③ L_1 loss

$$L = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$= \sum_{i=1}^{n} L_i$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{n} \frac{\partial L_i}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial w}$$

$$\hat{y}_i = g(w, x_i)$$

$$\frac{\partial \hat{y}_i}{\partial w} = \frac{\partial\, g(w, x_i)}{\partial w}$$

$$\frac{\partial L_i}{\partial \hat{y}_i} = \frac{\partial}{\partial \hat{y}_i} \left(|y_i - \hat{y}_i| \right)$$

signum fn.

$$= -\,\mathrm{sgn}(y_i - \hat{y}_i) = \mathrm{sgn}(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{n} \mathrm{sgn}(\hat{y}_i - y_i) \cdot \frac{\partial\, g(w, x_i)}{\partial w}$$

(4) Huber Loss

Huber loss is defined as,

for each data point: i.

$$L_i = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 , & |y_i - \hat{y}_i| \leq \delta \\ \delta |y_i - \hat{y}_i| - \frac{\delta^2}{2} , & o.w. \end{cases}$$

$$\frac{\partial L_i}{\partial w} = \frac{\partial L_i}{\partial \hat{y}_i} \cdot \frac{\partial g(w, x_i)}{\partial w} \qquad \hat{y}_i = g(w, x_i)$$

$$\frac{\partial L_i}{\partial \hat{y}_i} = \begin{cases} y_i - \hat{y}_i , & |y_i - \hat{y}_i| \leq \delta \\ -\delta \frac{|y_i - \hat{y}_i|}{(y_i - \hat{y}_i)} , & o.w. \end{cases}$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{n} \begin{cases} (y_i - \hat{y}_i) \cdot \frac{\partial g(w, x_i)}{\partial w} , & |y_i - \hat{y}_i| \leq \delta \\ -\delta \frac{|y_i - \hat{y}_i|}{(y_i - \hat{y}_i)} \cdot \frac{\partial g(w, x_i)}{\partial w} , & o.w. \end{cases}$$

5 L_2 loss

$$L_i = (y_i - \hat{y}_i)^2$$

$$L = \frac{1}{2} \sum_{i=1}^{n} L_i$$

for cancel when derivative

$$L = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$\hat{y}_i \rightarrow g(w, x_i)$

$$\frac{\partial L}{\partial w} = \frac{1}{2} \sum_{i=1}^{n} 2(y_i - \hat{y}_i) \cdot (-1) \cdot \frac{\partial g(w, x_i)}{\partial w}$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{n} (\hat{y}_i - y_i) \cdot \frac{\partial g(w, x_i)}{\partial w}$$

(6) Cosine Similarity

$$L = \sum_{i=1}^{n} \frac{y_i \cdot \hat{y_i}}{|y_i| \cdot |\hat{y_i}|} \qquad \hat{y_i} = g(\omega, x_i)$$

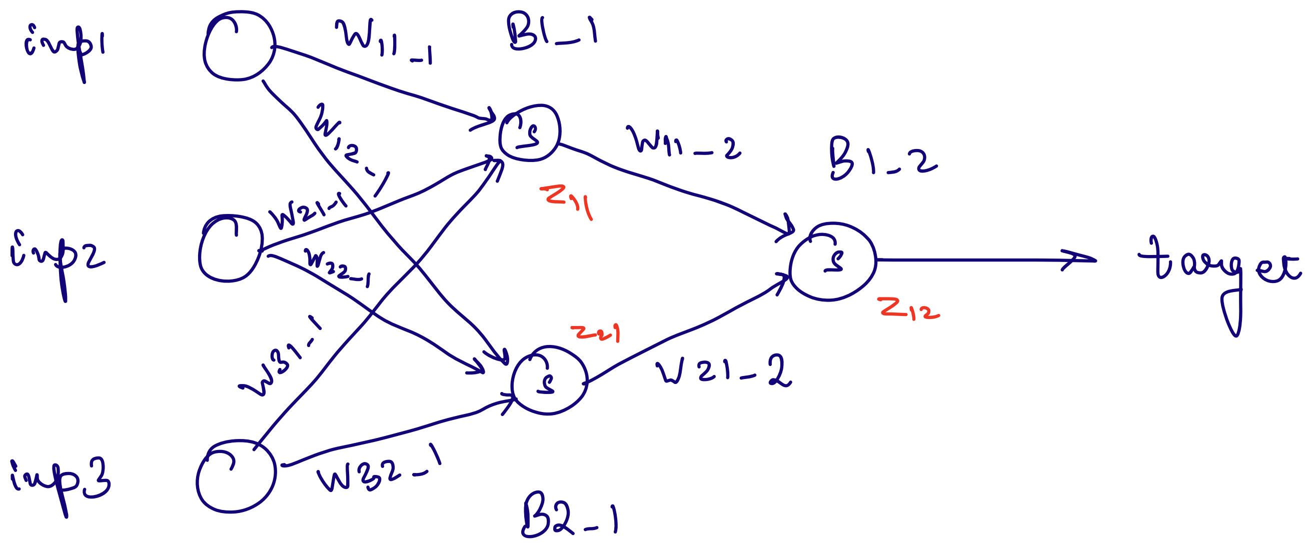
$$\frac{\partial L}{\partial \omega} = \sum_{i=1}^{n} \frac{y_i}{|y_i|} \cdot \underbrace{\frac{\partial}{\partial \hat{y_i}} \left(\frac{\hat{y_i}}{|\hat{y_i}|} \right)}_{\text{Signum function.}} \cdot \left(\frac{\partial \hat{y_i}}{\partial \omega} \right)$$

$$\frac{\partial L}{\partial \omega} = \sum_{i=1}^{n} \begin{cases} \frac{y_i}{|y_i|} \left(\frac{\partial \hat{y_i}}{\partial \omega} \right), & \hat{y_i} > 0 \\ -\frac{y_i}{|y_i|} \left(\frac{\partial \hat{y_i}}{\partial \omega} \right), & \text{ow.} \end{cases}$$

$$\begin{bmatrix} W_{11} & W_{21} & W_{31} \\ W_{12} & W_{22} & W_{32} \end{bmatrix} \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

$(2 \times 3) \times (3 \times 1) = (2 \times 1)$

III Hand calculation of gradients



$$\begin{bmatrix} z_{11} \\ z_{21} \end{bmatrix} = \sigma\left(W_1^T \underset{\text{inp}}{x} + \underset{\begin{bmatrix} B_{1\_1} \\ B_{2\_1} \end{bmatrix}}{b} \right)$$

On solving, (using MATLAB)

$$\begin{bmatrix} z_{11} \\ z_{21} \end{bmatrix} = \begin{bmatrix} -0.67 \\ 0.55 \end{bmatrix}$$

z_{12} further depends on z_{11}, z_{12}.

$$z_{12} = \sigma(w_2^T \underset{\vec{z}_1}{x} + b)$$

$$= \sigma(W_{11\_2} \cdot z_{11} + W_{21\_2} \cdot z_{21} + B_{1\_2})$$

$$z_{12} = 0.574$$

$$L = \frac{1}{2}(\underset{\hat{y}}{z_{12}} - \underset{\text{target}}{t})^2$$

$$\frac{\partial L}{\partial z_{12}} = (z_{12} - t)$$
$$= 1.394$$

$$y = \sigma(wx+c)$$
$$y = \frac{1}{1+e^{-(wx+c)}}$$
$$\frac{dy}{dx} = \frac{-w \cdot e^{-(wx+c)}}{(1+e^{-(wx+c)})^2}$$
$$= w \cdot \sigma(wx+c) \cdot (1-\sigma(wx+c))$$

variable multiplied with the variable $\frac{dy}{dx}$.

$$\frac{\partial L}{\partial w_{11-2}} = \frac{\partial L}{\partial z_{12}} \times \frac{\partial z_{12}}{\partial w_{11-2}}$$
$$= \frac{\partial L}{\partial z_{12}} \times (z_{11}) \times z_{12} \times (1-z_{12})$$
$$= 0.1148$$

Similarly,

$$\frac{\partial L}{\partial w_{21-2}} = \frac{\partial L}{\partial z_{12}} \times (z_{2}) \times z_{12} \times (1-z_{12})$$
$$= 0.2162$$

$$\frac{\partial L}{\partial B_{1-2}} = \frac{\partial L}{\partial z_{12}} \times z_{12} \times (1-z_{12})$$
$$= 0.3409$$

$$z_{11} = \sigma(w_{11-1} \cdot inp_1 + w_{21-1} \cdot inp_2 + w_{31-1} \cdot inp_3 + B_{1-1})$$

$$z_{21} = \sigma(w_{12-1} \cdot inp_1 + w_{22-1} \cdot inp_2 + w_{32-1} \cdot inp_3 + B_{2-1})$$

$$\frac{\partial L}{\partial w_{11\_1}} = \frac{\partial L}{\partial z_{12}} \times \frac{\partial z_{12}}{\partial w_{11\_1}}$$

further chain rule

$$\frac{\partial L}{\partial w_{11\_1}} = \frac{\partial L}{\partial z_{12}} \times \left(\frac{\partial z_{12}}{\partial z_{11}} \times \frac{\partial z_{11}}{\partial w_{11\_1}} \right)$$

$$z_{12} = \sigma(w_{11\_2} \cdot z_{11} + w_{21\_2} \cdot z_{21} + B_{1\_1})$$

$$\frac{\partial z_{12}}{\partial z_{11}} = w_{11\_2} \cdot \sigma(z_{12}) \cdot (1 - \sigma(z_{12}))$$

$$z_{11} = \sigma(w_{11\_1} \cdot inp_1 + w_{21\_1} \cdot inp_2 + w_{31\_1} \cdot inp_3 + B_{1\_1})$$

$$\frac{\partial z_{11}}{\partial w_{11\_1}} = z_{11} \times (1 - z_{11}) \times inp_1$$

Similarly,

$$\frac{\partial z_{11}}{\partial w_{21\_1}} = z_{11} \times (1 - z_{11}) \times inp_2$$

$$\frac{\partial z_{11}}{\partial w_{31\_1}} = z_{11} \times (1 - z_{11}) \times inp_3$$

$$\frac{\partial z_{11}}{\partial B_{1\_1}} = z_{11} \times (1 - z_{11})$$

Same thing done for $\frac{\partial z_{21}}{\partial (\ )}$ ✓.

$$\frac{\partial L}{\partial w_{11\_1}} = \left(\frac{\partial L}{\partial z_{12}}\right) \times \left(w_{11\_2} \cdot \sigma(z_{12}) \cdot (1-\sigma(z_{12}))\right)$$

$$\times \left(z_{11} \times (1-z_{11}) \times inp_1\right)$$

Substituting, we get:

$$\frac{\partial L}{\partial w_{11\_1}} = 0.0042$$

Similarly,

$$\frac{\partial L}{\partial w_{21\_1}} = 0.0415$$

$$\frac{\partial L}{\partial w_{31\_1}} = -0.0308$$

$$\frac{\partial L}{\partial B_{1\_1}} = -0.0381$$

$$\frac{\partial L}{\partial w_{12\_1}} = -0.0037$$

$$\frac{\partial L}{\partial w_{22\_1}} = -0.0362$$

$$\frac{\partial L}{\partial w_{32\_1}} = 0.0269$$

$$\frac{\partial L}{\partial B_{2\_1}} = 0.0332$$