CSE 616 : Neural Networks

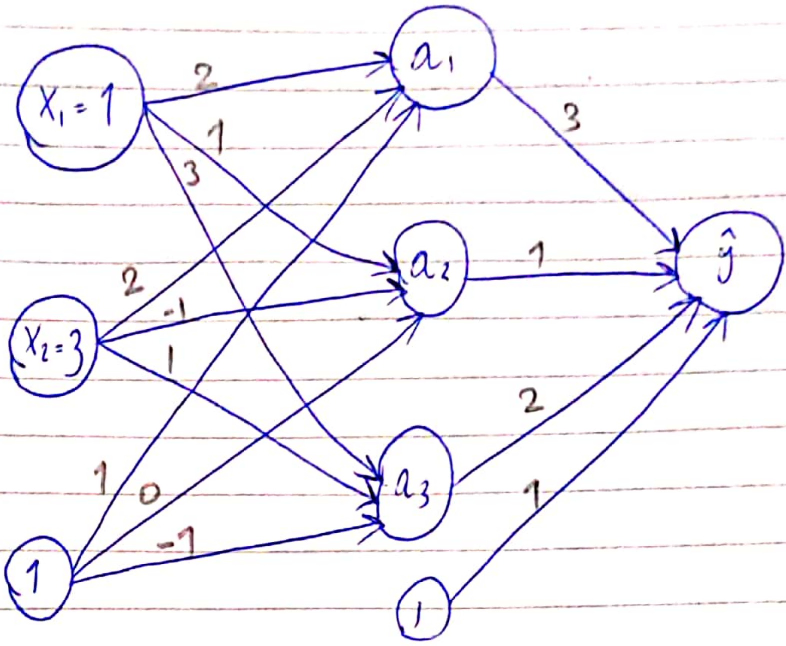
Assignment 1

حمدي أسامة محمد السيد

ID : 2100966

Mechatronics PG

1)



$y = 32$

(a) $\hat{y} = ?$ when activations are identity function

hidden layer :

$$a_1 = \sigma\left(\sum_{i=1}^{2}\left(w_{i1}^{[1]} x_i\right) + 1\right) = 1\left((2\times1) + (2\times3) + 1\right) = 9$$

$$a_2 = \sigma\left(\sum_{i=1}^{2}\left(w_{i2}^{[1]} x_i\right) + 0\right) = 1\left((1\times1) + (-1\times3) + 0\right) = -2$$

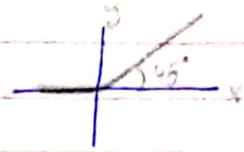
$$a_3 = \sigma\left(\sum_{i=1}^{2}\left(w_{i3}^{[1]} x_i\right) - 1\right) = 1\left((3\times1) + (1\times3) - 1\right) = 5$$

Output Layer :

$$\hat{y} = \sigma\left(\sum_{i=1}^{3}\left(w_{i1}^{[2]} a_i\right) + 1\right) = 1\left((3\times9) + (1\times-2) + (2\times5) + 1\right)$$

$$\hat{y} = 36$$, error $= \hat{y} - y = 36 - 32 = 4$

1b) $\hat{y} = ?$ when activations are ReLU function $\max(0, x)$



hidden layers:

$a_1 = \sigma\left(\sum_{i=1}^{2}(w_{i1}^{[1]} x_i) + 1\right) = ReLU(9) = 9$

$a_2 = \sigma\left(\sum_{i=1}^{2}(w_{i2}^{[1]} x_i) + 0\right) = ReLU(-2) = 0$

$a_3 = \sigma\left(\sum_{i=1}^{2}(w_{i3}^{[1]} x_i) - 1\right) = ReLU(5) = 5$

Output layer:

$$\hat{y} = \sigma\left(\sum_{i=1}^{3}(w_{i1}^{[2]} a_i) + 1\right) \qquad \boxed{\hat{y} = 38}$$

$$= ReLU\left((3\times 9) + (1\times 0) + (2\times 5) + 1\right) = 38$$

1c) $J = (\hat{y} - y)^2$ using $\hat{y}$ in (1a) $\longrightarrow J = (36 - 32)^2 \Rightarrow J = 16$

$$\frac{\partial J}{\partial b_1^{[2]}} = ? = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b_1^{[2]}}$$

$$\boxed{\frac{\partial J}{\partial b_1^{[2]}} = 2(\hat{y} - 32) \cdot 1 \rightarrow \text{at } \hat{y} = 36 \rightarrow \frac{\partial J}{\partial b_1^{[2]}} = 8}$$

$$\frac{\partial J}{\partial w_{21}^{[2]}} = ? = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_{21}^{[2]}}$$

$$\boxed{\frac{\partial J}{\partial w_{21}^{[2]}} = 2(\hat{y} - 32) \times -2 \xrightarrow{\text{at } \hat{y} = 36} \frac{\partial J}{\partial w_{21}^{[2]}} = -16}$$

$$J = (\hat{y} - 32)^2$$

$$\frac{\partial J}{\partial \hat{y}} = 2(\hat{y} - 32)$$

$$\hat{y} = w_{11}^{[2]} a_1 + w_{21}^{[2]} a_2 + w_{31}^{[2]} a_3 + b_1^{[2]}$$

$$\frac{\partial \hat{y}}{\partial b_1^{[2]}} = 1$$

$$\frac{\partial \hat{y}}{\partial w_{21}^{[2]}} = a_2 = -2$$ from Q (1a)

$$\frac{\partial \hat{y}}{\partial a_2} = w_{21}^{[2]}$$

1c) completion

$$\frac{\partial J}{\partial b_2^{[1]}} = ? = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b_2^{[1]}} = \underbrace{\frac{\partial J}{\partial \hat{y}}}_{2(\hat{y}-y)} \cdot \underbrace{\frac{\partial \hat{y}}{\partial a_2}}_{W_{21}^{[2]}} \times \underbrace{\frac{\partial a_2}{\partial b_2^{[1]}}}_{1}$$

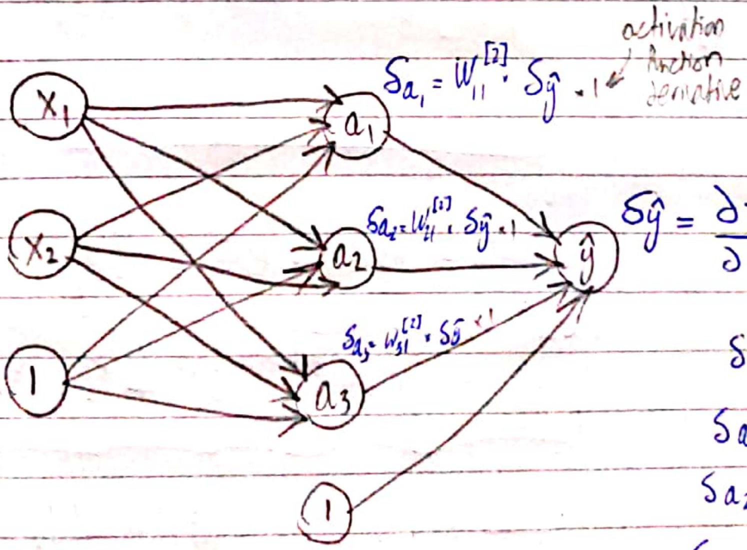
substitute $y = 32$, $\hat{y} = 36$, $W_{21}^{[2]} = 1$ $\longrightarrow$ $\boxed{\frac{\partial J}{\partial b_2^{[1]}} = 8}$

$$\frac{\partial J}{\partial W_{13}^{[1]}} = ? = \frac{\partial J}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial W_{13}^{[1]}} = \underbrace{\frac{\partial J}{\partial \hat{y}}}_{2(\hat{y}-y)} \times \underbrace{\frac{\partial \hat{y}}{\partial a_3}}_{W_{31}^{[2]}} \times \underbrace{\frac{\partial a_3}{\partial W_{13}^{[1]}}}_{x_1}$$

$\hat{y} = 36$, $y = 32$, $W_{31}^{[2]} = 2$, $x_1 = 1$

$$\boxed{\frac{\partial J}{\partial W_{13}^{[1]}} = 16}$$

1d)



activation ~~deri~~ derivative of identity function

$$\delta\hat{y} = \frac{\partial J}{\partial \hat{y}} \times 1$$

$y = 32$, $\hat{y} = 36$

$\delta\hat{y} = 2(\hat{y} - y) = 8$

$\delta a_1 = 6(\hat{y} - y) = 24$

$\delta a_2 = 2(\hat{y} - y) = 8$

$\delta a_3 = 4(\hat{y} - y) = 16$

Update formula: $W_{ij} = W_{ij} - \eta\, \delta_j z_i$ $\quad i \xrightarrow{W_{ij}} j$

$b_2^{[1]} = b_2^{[1]} - 2 \times \delta a_2 \times 1 \longrightarrow \boxed{b_2^{[1]} = 0 - 2 \times 8 \times 1 = -16}$

$W_{13}^{[1]} = W_{13}^{[1]} - 2 \times \delta a_3 \times x_1 \rightarrow \boxed{W_{13}^{[1]} = 3 - 2 \times 16 \times 1 = -29}$

Note that large learning rate lead to divergence.

1e) Splitting the data into testing and training sets enable us to check the robustness of the machine learning model to unseen data.

By using the training data to determine the model capacity with acceptable error, we then check the model performance on the testing data to verify its ability of generalization. if the model performed poorly on the testing dataset, overfitting occurred and can be solved by regularization or dropout or similar techniques that reduce the training dataset accuracy in return of higher testing set accuracy. Choosing the best model performing on test set will be a good indication of model generalization

---

2)

$$f = \sin(g_1) + g_2^2$$

$$g_1 = x_1 e^{x_2}$$

$$g_2 = x_1 + x_2^2$$

$$\frac{\partial f}{\partial x_1} = ? = \frac{\partial f}{\partial g_1} \cdot \frac{\partial g_1}{\partial x_1} + \frac{\partial f}{\partial g_2} \cdot \frac{\partial g_2}{\partial x_1}$$

$$\frac{\partial f}{\partial x_2} = ? = \frac{\partial f}{\partial g_1} \cdot \frac{\partial g_1}{\partial x_2} + \frac{\partial f}{\partial g_2} \cdot \frac{\partial g_2}{\partial x_2}$$

$$\frac{\partial f}{\partial g_1} = \cos(g_1)$$

$$\frac{\partial f}{\partial g_2} = 2g_2$$

$$\frac{\partial g_1}{\partial x_1} = e^{x_2}$$

$$\frac{\partial g_1}{\partial x_2} = x_1 e^{x_2}$$

$$\frac{\partial g_2}{\partial x_1} = 1$$

$$\frac{\partial g_2}{\partial x_2} = 2x_2$$

$$\frac{\partial f}{\partial x_1} = \cos(g_1) \cdot e^{x_2} + 2g_2 \cdot 1$$

$$\boxed{\frac{\partial f}{\partial x_1} = \cos(x_1 e^{x_2})\, e^{x_2} + 2(x_1 + x_2^2)}$$

$$\frac{\partial f}{\partial x_2} = \cos(g_1) \cdot x_1 e^{x_2} + 2g_2 \cdot 2x_2$$

$$\boxed{\frac{\partial f}{\partial x_2} = x_1 e^{x_2} \cos(x_1 e^{x_2}) + 4x_2(x_1 + x_2^2)}$$

3)

3.1) $f(z) = \frac{1}{1+e^{-z}}$, $\frac{df}{dz}$?

by using the quotient derivative rule:

$$\frac{df}{dz} = \frac{(0)\times(1+e^{-z}) - (1)\times(-e^{-z})}{(1+e^{-z})^2}$$

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

$$\boxed{\frac{df}{dz} = \left[f(z)\,(1-f(z))\right]}$$

3.2) $f(w) = \frac{1}{1+e^{-w^T x}}$ where w is a matrix of weights, what is $f'(w)$?

by using the quotient derivative rule:

$$f'(w) = \frac{0\times(1+e^{-w^T x}) - 1(-x e^{-w^T x})}{(1+e^{-w^T x})^2}$$

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

$$f'(w) = x\, f(w)\,(1-f(w))$$

3.3) $J(w) = \frac{1}{2} \sum_{i=1}^{m} \left| w^T x^{(i)} - y^{(i)} \right|$, $\frac{dJ}{dw} = ?$

$$= \frac{1}{2}\left(|w^T x^{(1)} - y^{(1)}| + |w^T x^{(2)} - y^{(2)}| \; ---- \right)$$

by using the derivative of absolute value $\frac{d}{dz}|z| = \frac{z}{|z|}$

$$\frac{dJ}{dw} = \frac{1}{2}\left(x^{(1)} \cdot \frac{w^T x^{(1)} - y^{(1)}}{|w^T x^{(1)} - y^{(1)}|} + x^{(2)} \cdot \frac{w^T x^{(2)} - y^{(2)}}{|w^T x^{(2)} - y^{(2)}|} + \ldots \right)$$

$$\boxed{\frac{dJ}{dw} = \frac{1}{2} \sum_{i=1}^{m} \left[x^{(i)} \cdot \frac{w^T x^{(i)} - y^{(i)}}{|w^T x^{(i)} - y^{(i)}|} \right]}$$

3.4) $J(w) = \frac{1}{2}\left[\sum_{i=1}^{m} (w^T x^{(i)} - y^{(i)})^2 \right] + \lambda \|w\|_2^2$, $\frac{dJ}{dw} = ?$

$$\frac{dJ}{dw} = \frac{1}{2}\left[2 \sum_{i=1}^{m} \left(x^{(i)} \cdot (w^T x^{(i)} - y^{(i)}) \right) \right] + \lambda \frac{d}{dw}\left(\sum w_j^2 \right)$$

$$\frac{dJ}{dw} = \sum_{i=1}^{m} \left[x^{(i)} \cdot (w^T x^{(i)} - y^{(i)}) \right] + \lambda \sum_{j=1}^{n} 2 w_j$$

sum all vector of weights

$$\boxed{\frac{dJ}{dw} = \sum_{i=1}^{m} \left[x^{(i)} \cdot (w^T x^{(i)} - y^{(i)}) \right] + 2\lambda \sum_{j=1}^{n} w_j}$$

regularization hyperparameter

3.5) $J(w) = \sum_{i=1}^{m} \left[\underbrace{y^{(i)} \log\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right)} + \underbrace{(1-y^{(i)}) \log\left(1 - \frac{1}{1+e^{-w^T x^{(i)}}}\right)} \right]$

$dJ/dw = ?$

let $z = \frac{1}{1+e^{-w^T x^{(i)}}}$

$$\frac{dJ}{dw} = \sum_{i=1}^{m} \frac{y^{(i)} \cdot x^{(i)} \cdot \cancel{z} \cdot (1-z)}{\cancel{z}} + \frac{(1-y^{(i)}) \cdot (-x^{(i)} \cdot z \cdot \cancel{(1-z)})}{\cancel{(1-z)}}$$

$$\boxed{\begin{aligned}\frac{dJ}{dw} &= \sum_{i=1}^{m} \left[y^{(i)} x^{(i)} (1-z) - (1-y^{(i)}) x^{(i)} z \right] \\ &= \sum_{i=1}^{m} \left[y^{(i)} x^{(i)} \left(1 - \frac{1}{1+e^{-w^T x^{(i)}}}\right) - (1-y^{(i)}) x^{(i)} \left(\frac{1}{1+e^{-w^T x^{(i)}}}\right) \right]\end{aligned}}$$

3.6) $\nabla_w f = ?$ where $f(w) = \tanh[w^T x]$ derivative rule: $\tanh(x) \overset{d/dx}{\Downarrow} \operatorname{sech}^2(x)$

$$\boxed{\nabla_w f = \frac{\partial f}{\partial w} = x \operatorname{sech}^2(w^T x)}$$