HW1.1

Saturday, April 23, 2022 8:47 PM

P1. 1. (a) 1.

For hidden layers: $a = ReLU(z) = \begin{cases} z, & z>0 \\ 0, & \text{otherwise} \end{cases}$

output layer: $a_1^{[3]} = sigmoid(z_1^{[3]})$

$$= \frac{1}{1+e^{-z_1^{[3]}}}$$

1st hidden layer: $z_1^{[1]} = \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = 3$.

$a_1^{[1]} = 3$.

$z_2^{[1]} = \begin{bmatrix} -3 & -5 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = -1$

$a_2^{[1]} = 0$.

$z_3^{[1]} = \begin{bmatrix} 3 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = 1$.

$a_3^{[1]} = 1$.

2nd layer: $z_1^{[2]} = \begin{bmatrix} 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} - 3 = 2 - 3 = -1$

$a_1^{[2]} = 0$.

$z_2^{[2]} = \begin{bmatrix} 1 & 1 & -4 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} + 2 = -1 + 2 = 1$.

$a_2^{[2]} = 1$.

Output layer: $z_1^{[3]} = \begin{bmatrix} 2 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} - 2 = 2 - 2 = 0$

$a_1^{[3]} = \frac{1}{1+e^0} = \frac{1}{2}$.

2. $L(y, \hat{y}) = -[y \log(\hat{y}) + (1-y)\log(1-\hat{y})]$

$\hat{y} = a_1^{[3]} = 1/2$, $y = 1$, $\alpha = 0.8$

① Output layer:

$$da_1^{[3]} = d\hat{y} = \frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = -\frac{1}{\hat{y}}.$$

$$dz_1^{[3]} = da_1^{[3]} \cdot \frac{\partial a_1^{[3]}}{\partial z_1^{[3]}} = \left(-\frac{1}{\hat{y}}\right) \sigma(z_1^{[3]}) \left[1 - \sigma(z_1^{[3]})\right]$$

$da_1 = -a/\hat{y} \quad \hat{y}+1-\hat{y}. \quad \hat{y}.$

$$dz_1^{[3]} = da_1^{[3]} \cdot \frac{\partial a_1^{[3]}}{\partial z_1^{[3]}} = \left(-\frac{1}{\hat{y}}\right) \sigma(z_1^{[3]}) \left[1-\sigma(z_1^{[3]})\right]$$
$$= -2\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right) = -\frac{1}{2}$$

$$dw_1^{[3]} = \frac{\partial L}{\partial w_1^{[3]}} \cdot \frac{\partial z_1^{[3]}}{\partial w_1^{[3]}} = a_1^{[2]} dz_1^{[3]} = 0 \cdot \left(-\frac{1}{2}\right) = 0$$

$$dw_2^{[3]} = a_2^{[2]} dz_1^{[3]} = -\frac{1}{2}.$$

$$db^{[3]} = dz_1^{[3]} = -\frac{1}{2}.$$

update: $w_{1,new}^{[3]} = w_1^{[3]} - \alpha \cdot dw_1^{[3]} = 2 - 0 = 2.$

$$w_{2,new}^{[3]} = w_{2,new}^{[3]} - \alpha \cdot dw_2^{[3]} = 2 - 0.8 \cdot \left(-\frac{1}{2}\right) = 2.4.$$

$$b_{new}^{[3]} = b^{[3]} - \alpha \cdot db^{[3]} = -2 - 0.8\left(-\frac{1}{2}\right) = -1.6$$

② 2nd hidden layer:

$$da_1^{[2]} = dz_1^{[3]} \cdot \frac{\partial z_1^{[3]}}{\partial a_1^{[2]}} = w_1^{[3]} \cdot dz_1^{[3]} = 2\left(-\frac{1}{2}\right) = -1.$$

$$da_2^{[2]} = w_2^{[3]} dz_1^{[3]} = -1$$

$\because$ for hidden layers. $a = \sigma(z) = ReLU(z)$.

$\therefore \sigma'(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise} \end{cases}$

$\Rightarrow dz_1^{[2]} = da_1^{[2]} \cdot \sigma'(z_1^{[2]}) = da_1^{[2]} \cdot \sigma'(-1) = 0.$

$$dz_2^{[2]} = da_2^{[2]} \cdot \sigma'(1) = -1.$$

$$dw_1^{[2]} = dz_1^{[2]} \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \\ a_3^{[2]} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$dw_2^{[2]} = dz_2^{[2]} \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \\ a_3^{[2]} \end{bmatrix} = -\begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 0 \\ -1 \end{bmatrix}$$

$$db_1^{[2]} = dz_1^{[2]} = 0.$$

$$db_2^{[2]} = dz_2^{[2]} = -1$$

update: $w_{1,new}^{[2]} = w_1^{[2]} - \alpha \cdot dw_1^{[2]} = w_1^{[2]} = \begin{bmatrix} 0 \\ -1 \\ 2 \end{bmatrix}$

$$w_{2,new}^{[2]} = w_2^{[2]} - \alpha\, dw_2^{[2]} = \begin{bmatrix} 1 \\ 1 \\ -4 \end{bmatrix} + 0.8 \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.4 \\ 1 \\ -3.2 \end{bmatrix}.$$

$$b_{1,new}^{[2]} = b_1^{[2]} - 0 = -3.$$

$$b_{2,new}^{[2]} = b_2^{[2]} - \alpha\, db_2^{[2]} = 2 - 0.8(-1) = 2.8$$

③ 1st hidden layer:

$$da_1^{[1]} = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} dz_1^{[2]} \\ dz_2^{[2]} \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = -1.$$

$$da_2^{[1]} = \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = -1.$$

$$da_3^{[1]} = \begin{bmatrix} 2 & -4 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = 4.$$

$$dz_1^{[1]} = da_1^{[1]} \sigma'(z_1^{[1]}) = -1 \cdot 1 = -1$$

$$dz_2^{[1]} = da_2^{[1]} \sigma'(z_2^{[1]}) = -1 \cdot 0 = 0$$

$$dz_3^{[1]} = da_3^{[1]} \sigma'(z_3^{[1]}) = 4 \cdot 1 = 4$$

$$dW_1^{[1]} = dz_1^{[1]} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

$$dW_2^{[1]} = dz_2^{[1]} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$dW_3^{[1]} = dz_3^{[1]} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 8 \\ -4 \end{bmatrix}$$

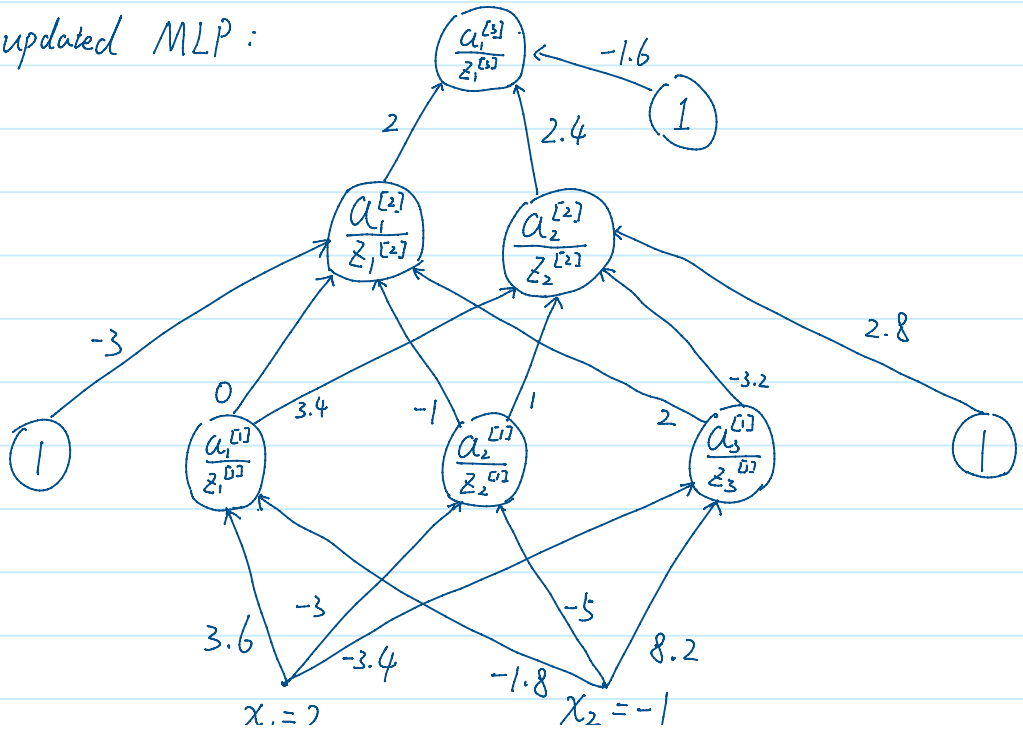
no bias when computing 1st hidden layer.

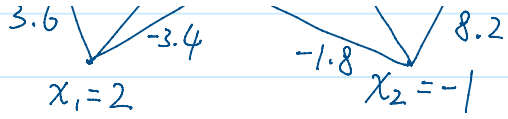
$$\therefore W_{1,\text{new}}^{[1]} = W_1^{[1]} - \alpha\, dW_1^{[1]} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} - 0.8 \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.6 \\ -1.8 \end{bmatrix}$$

$$W_{2,\text{new}}^{[1]} = W_2^{[1]} = \begin{bmatrix} -3 \\ -5 \end{bmatrix}$$

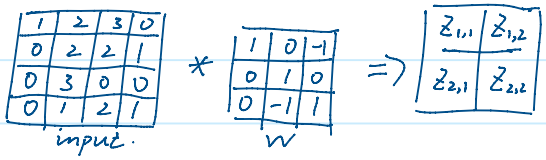
$$W_{3,\text{new}}^{[1]} = \begin{bmatrix} 3 \\ 5 \end{bmatrix} - 0.8 \begin{bmatrix} 8 \\ -4 \end{bmatrix} = \begin{bmatrix} -3.4 \\ 8.2 \end{bmatrix}$$

updated MLP:





b) 1.



$z_{1,1} = 1(1) + 3(-1) + 2(1) + 3(-1) = -3$.

$z_{1,2} = 2(1) + 2(1) = 4$.

$z_{2,1} = 2(-1) + 3(1) + 1(-1) + 2(1) = 2$.

$z_{2,2} = 2(1) + 1(-1) + 2(-1) + 1 = 0$.

∴ output after filter w ⇒

-3	4
2	0

2. Max Pool 2×2 ⇒ output: 4

3. ∵ $L(y, \hat{y}) = (y - \hat{y})^2$.

$d\hat{y} = \frac{\partial L}{\partial \hat{y}} = -2(y - \hat{y}) = -2(2 - \hat{y}) = 2\hat{y} - 4$.

$\frac{\partial L}{\partial w_{11}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_{11}} = 2\hat{y} - 4 = 8 - 4 = 4$.

$\frac{\partial L}{\partial w_{12}} = 0$

$\frac{\partial L}{\partial w_{13}} = -(2\hat{y} - 4) = -4$

$\frac{\partial L}{\partial w_{21}} = 0 = \frac{\partial L}{\partial w_{23}}$

$\frac{\partial L}{\partial w_{22}} = 2\hat{y} - 4) = 4$.

$\frac{\partial L}{\partial w_{31}} = 0$.

$\frac{\partial L}{\partial w_{32}} = -4$.

$\frac{\partial L}{\partial w_{33}} = 4$.

$$\therefore \begin{bmatrix} \frac{\partial L}{\partial W_{11}} & \frac{\partial L}{\partial W_{12}} & \frac{\partial L}{\partial W_{13}} \\ \frac{\partial L}{\partial W_{21}} & \frac{\partial L}{\partial W_{22}} & \frac{\partial L}{\partial W_{23}} \\ \frac{\partial L}{\partial W_{31}} & \frac{\partial L}{\partial W_{32}} & \frac{\partial L}{\partial W_{33}} \end{bmatrix} = \begin{bmatrix} 4 & 0 & -4 \\ 0 & 4 & 0 \\ 0 & -4 & 4 \end{bmatrix}$$

c) 1. $\frac{\partial S_t}{\partial S_{t-1}} = \frac{\partial[\sigma(UX_t + WS_{t-1})]}{\partial S_{t-1}}$

$\frac{\partial S_{t-1}}{\partial S_{t-2}} = \frac{\partial[\sigma(UX_{t-2} + WS_{t-2})]}{\partial S_{t-2}}$

$\vdots$

$\frac{\partial S_{k+1}}{\partial S_k} = \frac{\partial[\sigma(UX_k + WS_k)]}{\partial S_k}$

$\therefore \frac{\partial S_t}{\partial S_k} = \frac{\partial S_t}{\partial S_{t-1}} \cdot \frac{\partial S_{t-1}}{\partial S_{t-2}} \cdots \frac{\partial S_{k+1}}{\partial S_k} = \prod_{j=k+1}^{t} \frac{\partial h_j}{h_{j-1}}$

$= \frac{\partial[\sigma(UX_{t-1} + WS_{t-1})]}{\partial S_{t-1}} \cdot \frac{\partial[\sigma(UX_{t-2} + WS_{t-2})]}{\partial S_{t-2}} \cdots \frac{\partial[\sigma(UX_k + WS_k)]}{\partial S_k}$

$= \prod_{j=k+1}^{t} W^T X \left(\sigma(UX_{j-1} + WS_{j-1})[1 - \sigma(UX_{j-1} + WS_{j-1})]\right)$

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

2. if $\text{eigen}(W) < 1$, the model suffers from gradient vanishing.

HW1.2

Sunday, April 24, 2022 12:56 AM

2. (a) $\min_x f(x)$

$$x(1) = x(0) - \alpha \cdot \nabla f(x)$$

$$x(2) = x(1) - \alpha \cdot \nabla f(x)$$

$\nabla f(x)$ – gradient of $f(x)$

α – learning rate.

if α is too high, the model will overfit,
if α is too low, the model will underfit.

(b) $V_1 = \beta(V_0) + (1-\beta)\nabla f(x) = (1-\beta)\nabla f(x)$

$$x(1) = x(0) - \alpha V_1$$

$$= x(0) - \alpha(1-\beta)\nabla f(x)$$

$$V_2 = \beta V_1 + (1-\beta)\nabla f(x) = \beta(1-\beta)\nabla f(x) + (1-\beta)\nabla f(x)$$

$$= (1-\beta)(1+\beta)\nabla f(x)$$

$$x(2) = x(1) - \alpha V_2$$

$$= x(0) - \alpha(1-\beta)\nabla f(x) - \alpha(1-\beta)(1+\beta)\nabla f(x)$$

$$= x(0) - \alpha(1-\beta)(2+\beta)\nabla f(x)$$

$\nabla f(x)$ – gradient of $f(x)$.

β – exponential decay rate.

V_t – momentum at time t.

α – learning rate.

(c) $V_1 = \beta_1 V_0 + (1-\beta_1)\nabla f(x)$.

$= (1-\beta_1)\nabla f(x)$.

$S_1 = \beta_2 S_0 + (1-\beta_2)[\nabla f(x)]^2$.

$X(1) = X(0) - \frac{\alpha}{\sqrt{S_1}+\varepsilon} V_1$.

$V_2 = \beta_1 V_1 + (1-\beta_1)\nabla f(x)$.

$S_2 = \beta_2 S_1 + (1-\beta_2)[\nabla f(x)]^2$.

$X(2) = X(1) - \frac{\alpha}{\sqrt{S_2}+\varepsilon} V_2$.

α — learning rate

β_1 - exponential decay rate for first moment.

β_2 - ..... for second moment.

V_t - first moment at time t.

S_t - second moment at time t.

$\nabla f(x)$ - gradient of $f(x)$

(d) $\hat{y} = W^T X + b$; $L(y, \hat{y}) = (y-\hat{y})^2$.

1. $X = [1, -2, 3, 2, 4]$, $W = [0, 3, 1, 4, -1]$, $b = -1$

$$\hat{y} = W^T X + b = [0\ 3\ 1\ 4\ -1]\begin{bmatrix} 1 \\ -2 \\ 3 \\ 2 \\ 4 \end{bmatrix} - 1 = 0$$

$\therefore L = (y-\hat{y})^2 = 4$.

2. $\min_{\hat{y}} L(y, \hat{y}) = \min_{\hat{y}} (y-\hat{y})^2$, $\hat{y} = W^T X + b$, $y = 2$.

$\Rightarrow \min_{W,b} L(y, W^T X + b) = \min_{W,b} [y - (W^T X + b)]^2$.

$$\Rightarrow \nabla_w L = -2X(2 - W^T X + b) = -4X = [-4 \;\; 8 \;\; -12 \;\; -8 \;\; -16]^T$$

$$\nabla_b L = -2(2-0) = -4.$$

update $\Rightarrow$

$$\begin{cases} W_{new} = W - \alpha \nabla_w L = \begin{bmatrix} 0 \\ 3 \\ 1 \\ 4 \\ -1 \end{bmatrix} - \begin{bmatrix} -0.4 \\ 0.8 \\ -1.2 \\ -0.8 \\ -1.6 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 2.2 \\ 2.2 \\ 4.8 \\ 0.6 \end{bmatrix} \\ b_{new} = b - \alpha \nabla_b L = -1 + 0.4 = -0.6 \end{cases}$$

3. $$V_{w,new} = \beta V_w + (1-\beta)\nabla_w L = 0.9[0.5 \;\; 0 \;\; -0.5 \;\; 0 \;\; 0]^T + 0.1[-4 \;\; 8 \;\; -12 \;\; -8 \;\; -16]^T$$

$$= [0.45 \;\; 0 \;\; -0.45 \;\; 0 \;\; 0]^T + [-0.4 \;\; 0.8 \;\; -1.2 \;\; -0.8 \;\; -1.6]^T$$

$$= [0.05 \;\; 0.8 \;\; -1.65 \;\; -0.8 \;\; -1.6]^T$$

$$V_{b,new} = \beta V_b + (1-\beta)\nabla_b L = 0.9(0.2) + 0.1(-4)$$

$$= 0.18 - 0.4$$

$$= -0.22$$

$$W_{new} = W - \alpha V_{w,new} = \begin{bmatrix} 0 \\ 3 \\ 1 \\ 4 \\ -1 \end{bmatrix} - 0.1 \begin{bmatrix} 0.05 \\ 0.8 \\ -1.65 \\ -0.8 \\ -1.6 \end{bmatrix} = \begin{bmatrix} -0.005 \\ 2.92 \\ 1.165 \\ 4.08 \\ -0.84 \end{bmatrix}.$$

$$b_{new} = b - \alpha V_{b,new} = -1 + 0.022 = -0.978$$

(e) 1. $$z = W^T X + b = [0 \;\; 3 \;\; 1 \;\; 4 \;\; -1] \begin{bmatrix} 1 \\ -2 \\ 3 \\ 2 \\ 4 \end{bmatrix} - 1 = 0.$$

$$\hat{y} = \frac{1}{1+e^0} = \frac{1}{2}.$$

$$L(y, \hat{y}) = -\left(\log\left(\tfrac{1}{2}\right)\right) = \log(2).$$

2. $$\nabla_w L = \frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w} = X(\hat{y} - y) = X\left(\tfrac{1}{2} - 1\right) = -\tfrac{1}{2}X = \begin{bmatrix} -0.5 \\ 1 \\ -1.5 \\ -1 \\ -2 \end{bmatrix}$$

$$\nabla_b L = \frac{\partial L}{\partial b} = -\frac{1}{2}.$$

$\nabla_b L = \frac{\partial L}{\partial b} = -\frac{1}{2}$.

$$W_{new} = W - \alpha \nabla_W L = [0 \;\; 3 \;\; 1 \;\; 4 \;\; -1]^T - 0.1[-0.5 \;\; 1 \;\; -1.5 \;\; -1 \;\; -2]^T$$

$$= [0.05 \;\; 2.9 \;\; 1.15 \;\; 4.1 \;\; -0.8]^T$$

$$b_{new} = b - \alpha \nabla_b L = -1 + 0.1(0.5) = -0.95$$

HW1.3

Sunday, April 24, 2022 3:01 AM

P3.

1. degree ① matrix · · adjacency ② matrix ·

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \cdot \qquad \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Laplacian matrix = ① - ②

$$= \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$$

2. $L = U \Lambda U^T$

$\Lambda = diag(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$

$\lambda_1 = 0, \quad \lambda_2 = 2, \quad \lambda_3 = \lambda_4 = 4.$

$$U = \begin{bmatrix} -0.5 & 0.7071 & 0.4914 & -0.0924 \\ -0.5 & 0 & -0.3607 & 0.7874 \\ -0.5 & 0 & -0.6221 & -0.6025 \\ -0.5 & -0.7071 & 0.4914 & -0.0924 \end{bmatrix}$$

the eigenvalues of graph Laplacian matrix are non-negative (i.e. ≥ 0).

In other word, the matrix is

matrix are non-negative (i.e. ≥ 0).

In other word, the matrix is positive semi-definite. It is true for all graph Laplacian matrix, since $V^T L V = \lambda V^T V \geq 0$. where V is an eigenvector of matrix L.

3. $\hat{x} = U^T x = U^T \cdot \begin{bmatrix} -2 \\ -1 \\ -2 \\ 2 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0 \\ 3.5704 \\ 0.0478 \end{bmatrix}$.