Juhi Patel - JPP2464

1.1 $y = f(x) = w \cdot x + b = 2x_1 + x_2 + b$

$$\frac{dy}{dx_1} = w_1 = 2$$

$$\frac{dy}{dx_2} = w_2 = 1$$

1.2 $MSE = \frac{1}{4}\left((0-0.1)^2 + (1-0.95)^2 + (1-1.10)^2 + (0-0.2)^2\right)$

$= 0.015625$

1.3 i) $y = w \cdot x + b = 2x_1 + x_2 + 3$

$\hat{y}_1 = 0 + 0 + 3 = 3$

$\hat{y}_2 = 0 + 1 + 3 = 4$

$\hat{y}_3 = 2 + 0 + 3 = 5$

$\hat{y}_4 = 2 + 1 + 3 = 6$

$MSE = \frac{1}{4}\left((3-0)^2 + (4-1)^2 + (5-1)^2 + (6-0)^2\right)$

$= 17.5$

ii) $\hat{y}_i = w \cdot x_i + b = w_1 x_{1i} + w_2 x_{2i} + b$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (w_1 x_{1i} + w_2 x_{2i} + b - y_i)^2$$

$$\frac{d(MSE)}{d(w_1)} = \frac{2}{n}\sum_{i=1}^{n} (w_1 x_{1i} + w_2 x_{2i} + b - y_i)\, x_{1i}$$

$= \frac{2}{4}\left[(3-0)\times 0 + (4-1)\times 0 + (5-1)\times 1 + (6-0)\times 1\right.$

$= 5$

$$\frac{d(MSE)}{d(w_2)} = \frac{2}{n}\sum_{i=1}^{n} (w_1 x_{1i} + w_2 x_{2i} + b - y_i)\, x_{2i}$$

$= \frac{2}{4}\left[(3-0)\times 0 + (4-1)\times 1 + (5-1)\times 0 + (6-0)\times 1\right]$

$= 4.5$

direction $(-5, -4.5, -8)$

paper only

iii) $\frac{d(MSE)}{dw_1} = \frac{d(MSE)}{dw_2} = \frac{d(MSE)}{db} = 0$

$f(x) = w \cdot x + b = w_1 x_1 + w_2 x_2 + b = w_1 x_1 + w_2 x_2 + b$

$\hat{y}_1 = b \rightarrow \hat{y}_1 - y_1 = b$

$\hat{y}_2 = w_2 + b \rightarrow \hat{y}_2 - y_2 = w_2 + b - 1$

$\hat{y}_3 = w_1 + b \rightarrow \hat{y}_3 - y_3 = w_1 + b - 1$

$\hat{y}_4 = w_1 + w_2 + b \rightarrow \hat{y}_4 - y_4 = w_1 + w_2 + b$

$\frac{d(MSE)}{dw_1} = 0 \rightarrow \sum_{i=1}^{n} (\hat{y}_i - y_i) x_{1i} = 0$

$w_1 + b_0 - y \quad w_1 + w_2 + b = 0 \rightarrow 2w_1 + w_2 + 2b - 1 = 0$

$\frac{d(MSE)}{dw_2} = 0 \rightarrow \sum_{i=1}^{n} (\hat{y}_i - y_i) x_{2i} = 0$

$\rightarrow w_2 + b + w_1 + w_2 + b = 0 \rightarrow w_1 + 2w_2 + 2b - 1 = 0$

$\frac{d(MSE)}{db} = 0 \rightarrow \sum_{i=1}^{n} (\hat{y}_i - y_i) = 0$

$\rightarrow 2w_1 + 2w_2 + 4b - 2 = 0 \rightarrow w_1 + w_2 + 2b - 1 = 0$

$w_1 = 0, \ w_2 = 0, \ w_3 = \frac{1}{2}$

Extra Credit on python

(2.1)

$x = [1, -1, 3, 4, 4]$

kernel $= [1, 1]$

for stride $= 1$

$conv(k, x) = [1-1, -1+3, 3+4, 4+4]$

$= [0, 2, 7, 8]$

for stride $= 2$,

$Conv(k, x = [1-1, 3+4] = [0, 7]$

(2.2) i)

$$I = \begin{bmatrix} 0.1 & -0.6 & 0.4 & 0.8 \\ -0.4 & 0.3 & 0.9 & 0.2 \\ 0.5 & 0.2 & 0.8 & -0.7 \\ 0.3 & 0.7 & -0.4 & 0.1 \end{bmatrix} \qquad K_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad K_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$Conv(K_1, I)$, stride $= [1, 1]$

$$= \begin{bmatrix} 0.1+0.3 & -0.6+0.9 & 0.4+0.2 \\ -0.4+0.2 & 0.3+0.8 & 0.9-0.7 \\ 0.5+0.7 & 0.2-0.4 & 0.8+0.1 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.3 & 0.6 \\ -0.2 & 1.1 & 0.2 \\ 1.2 & -0.2 & 0.9 \end{bmatrix}$$

$Conv(K_2, I)$, Stride $= [1, 1]$

$$\begin{bmatrix} -0.6-0.4 & 0.4+0.3 & 0.8+0.9 \\ 0.3+0.5 & 0.9+0.2 & 0.2+0.8 \\ 0.2+0.3 & 0.8+0.7 & -0.7-0.4 \end{bmatrix} = \begin{bmatrix} -1.0 & 0.7 & 1.7 \\ 0.8 & 1.1 & 1.0 \\ 0.5 & 1.5 & -1.1 \end{bmatrix}$$

ii) each channel has dimensions 3x 3 in 2.2i
if the stride is [2.2] channel would be 2x 2

iii) filter should 4x1 , stride 1 or ~~2~~ or any other

iv) maxpool (I) with filters 2x 2 stride = 1

$$\begin{bmatrix} 0.3 & 0.9 & 0.9 \\ 0.5 & 0.9 & 0.9 \\ 0.7 & 0.8 & 0.8 \end{bmatrix} = \text{dimensions } 3 \times 3$$

Extra Credit

advantages of max pooling
(i) easier to do computationally
(ii) easier extraction when it comes to extracting most prominent feautures

When do you think max-pooling operation may not be advantageous?

You would not want to extract prominent features only while still keeping subtle features to remain in network - the max-pool would discard them

(3.1) i) False
ii) True
iii) False
iv) False

(3.2) $pr(x = cat) = \text{sigmoid}(w'x)$

$$= \frac{1}{1 + \exp(-w_1 x_{11} - w_2 x_{12})}$$

given $x_1 = [2,1]$, & $w = [1,1]$

i) $pr(x = dog) =$

$$1 - \frac{1}{1 + \exp(-3)} = 1 - 0.952$$

<u>4.74%</u>

$Pr(x = cat) = 0.952 = $ <u>95.25%</u>

ii) $\|x - x_1\|_{int} \leq 0.1$

$x = [a, b]$ then $x - x_1 = [a-2, b-1]$

Case 1 : $a > 2$ & $b > 1$

$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & a-2 > b-1 \\ b-1 ; & b-1 > a-2 \end{cases} = \begin{cases} a-2, & a-b > 1 \\ b-1, & a-b < 1 \end{cases}$$

Case 2 $a < 2$ and $b > 1$

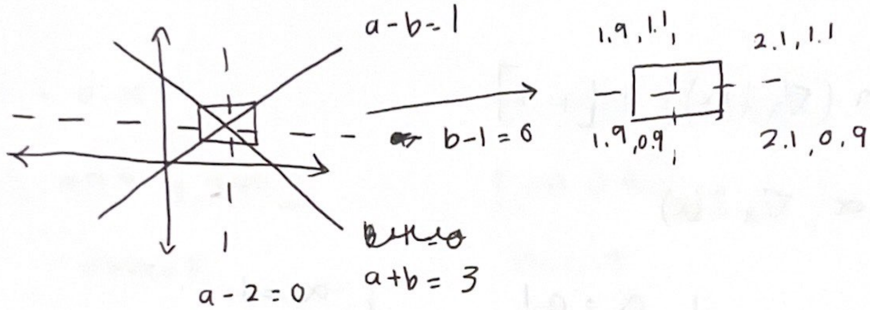
$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & 2-a > b-1 \\ b-1 ; & 2-a < b-1 \end{cases} = \begin{cases} a-2 ; & a+b < 3 \\ b-1 ; & a+b > 3 \end{cases}$$

Case 3 $a > 2$ and $b < 1$

$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & a-2 > 1-b \\ b-1 & a-2 < 1-b \end{cases} = \begin{cases} a-2 ; & a+b > 3 \\ b-1 ; & a+b < 3 \end{cases}$$

case 4 $a < 2$ and $b < 2$

$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & 2-a > 1-b \\ b-1 ; & 2-a < 1-b \end{cases} = \begin{cases} a-2, & a-b < 1 \\ b-1, & a-b > 1 \end{cases}$$



iii)

$$X_{optimiz} = \max_x \left(1 - \frac{1}{1+\exp(-w'x)} \right)$$

s.t. $\|x - x_1\|_{int} < 0.1$

given $x_1 = [2, 1]$

iv)

$$f(x) = \text{Sigmoid}(w'x) = \frac{1}{1+\exp(-w'x)}$$

$$f(x) = \frac{d}{dx}\left[\frac{1}{1+\exp(-w'x)}\right]$$

$$= \frac{-1}{(1+\exp(-w'x))^2} \times \exp(-w'x) \times -w'$$

$$= \frac{w'\exp(-w'x)}{(1+\exp(-w'x))^2}$$

for $w = [1, 1]$ and $x_1 = [2\ 1]$

$$\nabla_x f(x) = \frac{\exp(-3)}{(1+\exp(-3))^2}[1\ 1] = [0.045\ \ 0.045]$$

$$\text{sign}(\nabla_x f(x)) = [1\ 1]$$

$$X_{attack} = x_1\ \alpha\ \nabla_x f(x)$$

$\alpha = 0.001$	$\alpha = 0.1$	$\alpha = 1$
[1.99995, 0.99995]	[1.995, 0.995]	1.995, 0.995
threat model	threat model	threat model

$$X_{attack} = x_1 - \alpha\ \text{sign}(\nabla_x f(x))$$

$\alpha = 0.001$	$\alpha = 0.1$	$\alpha = 1$
[1.999, 0.999]	[1.9, 0.9]	[1, 0]
threat model	threat model	not in threat model

ii) $\|x - x_1\|_{int} \le 0.1$

$x = [a, b]$ then $x - x_1 = [a-2, b-1]$

Case 1 : $a > 2$ & $b > 1$

$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & a-2 > b-1 \\ b-1 ; & b-1 > a-2 \end{cases} = \begin{cases} a-2 , & a-b > 1 \\ b-1 , & a-b < 1 \end{cases}$$

Case 2 $a < 2$ and $b > 1$

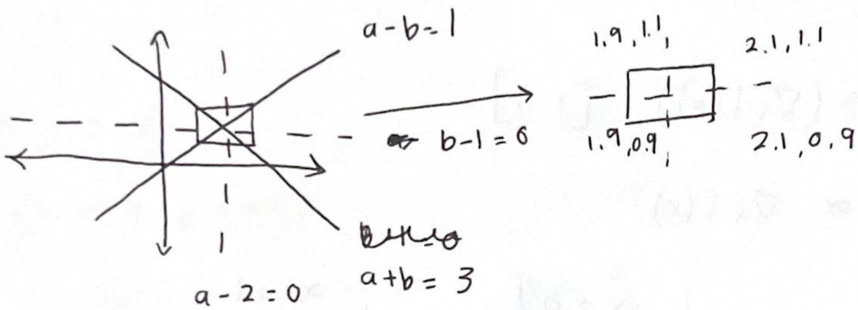
$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & 2-a > b-1 \\ b-1 ; & 2-a < b-1 \end{cases} = \begin{cases} a-2 ; & a+b < 3 \\ b-1 ; & a+b > 3 \end{cases}$$

Case 3 $a > 2$ and $b < 1$

$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & a-2 > 1-b \\ b-1 & a-2 < 1-b \end{cases} = \begin{cases} a-2 ; & a+b > 3 \\ b-1 ; & a+b < 3 \end{cases}$$

Case 4 $a < 2$ and $b < 2$

$$\|x - x_1\|_{int} = \begin{cases} a-2 ; & 2-a > 1-b \\ b-1 ; & 2-a < 1-b \end{cases} = \begin{cases} a-2 , & a-b < 1 \\ b-1 , & a-b > 1 \end{cases}$$



iii)

$$X_{optimiz} = \max_x \left(1 - \frac{1}{1+\exp(-w'x)} \right)$$

s.t. $\|x - x_1\|_{int} < 0.1$

given $x_1 = [2, 1]$

(v) $f(x) = 1 - \text{sigmoid}(w'x) = 1 - \frac{1}{(1+\exp(-w'x))^2}$

$$f'(x) = \frac{df(x)}{dx} = \frac{-w'\exp(-w'x)}{(1+\exp(-w'x))^2}$$

$\nabla_x f(x) = [-0.045, -0.045]$, $\text{sign}(\nabla_x f(x)) = [-1, -1]$

$X_{attack} = X_1 + \alpha \nabla_x f(x)$

$X_{attack} = X_1 + \alpha \nabla_x f(x)$

$\alpha = 0.001$	$\alpha = 0.1$	$\alpha = 1$
[1.99995, 0.99995]	[1.995, 0.995]	[1.955, 0.955]
threat	threat	threat

$\alpha = 0.001$	$\alpha = 0.1$	$\alpha = 1$
[1.999, 0.999]	[1.9, 0.9]	[1, 0]
threat	threat	not threat

vi) we see that new points lie in the threat model when we use gradient values. This means the new images aren't very different. But when we use the signs of the gradient it can have a stronger attack. This can allow the new image to be close to the target image. In this case, targeted and untargeted attack make no difference to the result since there are only two case.