Ques 01 $y = f(x) = w.x + b$; $w = [2, 1]$; $b = 3$

RUCHI SHARMA

1.1 $x = [4, 2]$

RS58898

$$\frac{\partial y}{\partial x_1} = w_1 = 2$$

$$\frac{\partial y}{\partial x_2} = w_2 = 1$$

1.2 y-true = [0, 1, 1, 0]

y-pred = [0.1, 0.95, 1.1, 0.2]

$$MSE = \frac{1}{n}\Sigma(y\text{-pred} - y\text{-true})^2$$

$$= \frac{1}{4}\left[(0.1-0)^2 + (0.95-1)^2 + (1.1-1)^2 + (0.2-0)^2\right]$$

$$= 0.015625$$

1.3 i)

x1	x2	y	y = w.x + b where w = [2, 1], b = 3
0	0	0	3
0	1	1	4
1	0	1	5
1	1	0	6

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

$$= 17.5$$

ii) to determine the direction → take gradient of loss fn w.r.t w & b

↳ $L = MSE = \frac{1}{n}(y - w_1x_1 - w_2x_2 - b)^2$

$$\frac{\partial L}{\partial w_1} = \sum \frac{2}{n}(y - w_1x_1 - w_2x_2 - b)(-x_1) \Rightarrow \left|\frac{\partial L}{\partial w_1}\right| = 5$$

$$\frac{\partial L}{\partial w_2} = \sum \frac{2}{n}(y - w_1x_1 - w_2x_2 - b)(-x_2) \Rightarrow \left|\frac{\partial L}{\partial w_2}\right| = 4.5$$

$$\frac{\partial L}{\partial b} = \sum \frac{2}{n}(y - w_1x_1 - w_2x_2 - b)(-1) \Rightarrow \left|\frac{\partial L}{\partial b}\right| = 8$$

$\Rightarrow$ we pick the highest magnitude gradient as our direction for descent.

since loss ~~or~~ is defined as $\frac{1}{n}(y - w_1x_1 - w_2x_2 - b)^2$

we want our weights to move in the direction towards 0 to ↓ loss by max amount

iii) loss min

$$\Rightarrow \frac{\partial L}{\partial w} = 0 \Rightarrow \begin{aligned} &1 - w_1 - b - w_1 - w_2 - b = 0 \\ &1 - w_2 - b - w_1 - w_2 - b = 0 \\ &1 - 2w_1 - 2w_2 - 2b = 0 \end{aligned}$$

solving these eqn $\Rightarrow$ $w_1 = w_2 = 0$ & $b = 0.5$

Solving using Gradient Descent. (some iterations)

let learning rate $= \lambda = 0.1$

loss at $[w_1, w_2] = [2\ 1]$ and $b = 3$ is 17.5

we have grads as $\frac{\partial L}{\partial w_1} = 5$; $\frac{\partial L}{\partial w_2} = 4.5$; $\frac{\partial L}{\partial w_3} = 8$

updating :

$$w_1 = w_1 - 0.1\frac{\partial L}{\partial w_1} = 2 - 0.1 \times 5 = 1.5$$

$$w_2 = 1 - 0.1 \times 4.5 = 0.55$$

$$\cancel{w_3}\ b = 3 - 0.1 \times 8 = 2.2$$

calculating y's with updated values

0,0 → 2.2
0,1 → 2.75
1,0 → 3.7
1,1 → 4.25

updated loss = $\frac{1}{4}\left((0-2.2)^2 + (1-2.75)^2 + (1-3.7)^2 + (0-4.25)^2\right)$

= 8.31375

(loss ↓'d)

calculating updated gradient

$\frac{\partial J}{\partial w_1} = 3$; $\frac{\partial J}{\partial w_2} = 3.475$; $\frac{\partial J}{\partial b} = 5.45$

(grads ↓d)

again, updating :

$w_1 = w_1 - 3\lambda = 1.5 - 0.1 \times 3 = 1.2$

$w_2 = 0.55 - 0.1 \times 0.3475 = 0.2$

$b = 2.2 - 0.1 \times 5.45 = 1.655$

continuing further

min loss is attained at $w_1 = 0$

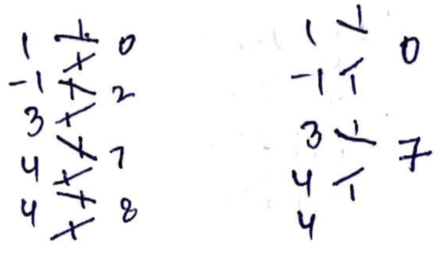
$w_2 = 0$

$b = 0.5$

(as also seen in code)

Ques 2

2.1 $x = [1, -1, 3, 4, 4]$



stride = 1 → we get $[0, 2, 7, 8]$

stride = 2 → we get $[0, 7]$ but it is incomplete.

2.2 i) filter

1	0
0	1

image

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

for stride 1

channel 1

$1\times0.1 + 0\times-0.6 + -0.4\times0 + 1\times0.3 = 0.4$	0.3	0.6
-0.2	1.1	0.2
1.2	-0.2	0.9

for filter

0	1
1	0

and stride = 1

Channel 2

$0\times0.1 + 1\times-0.6 + 1\times-0.4 + 0\times0.3 = -1.0$	0.7	1.7
0.8	1.1	1.0
0.5	1.5	-1.1

ii) dimension for channel 1 = 3×3
chanel 2 = 3×3

with stride = 2
dimension for channel 1 = 2×2
channel 2 = 2×2

iii) for output to be 1-D, we need to use a 4×1 or 1×4 filter (or 4×4!).
i.e kernel size should have atleast one dimension common to the image

iv) maxpooling with stride = 1 gives → 3×3

0.3	0.9	0.9
0.5	0.9	0.9
0.7	0.8	0.8

Ques 3

3.1
i) false
ii) true
iii) false
iv) false

3.2 $f(x) = Pr(x = cat) = sigmoid(w'x)$

i) $x_1 = [2, 1]$ cat

$$Pr(x = cat) = \frac{1}{1 + e^{-w'x}} = \frac{1}{1 + e^{-3}} = 0.952$$

$Pr(x = dog) = 1 - 0.952 = 0.048$

ii) given threat model → if $||x - x_1||_{inf} <= 0.1$

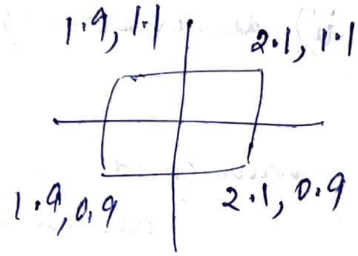
let $x = (a_1, a_2)$ then $x - x_1 = (a_1 - 2, a_2 - 1)$

#1 → if $a_1 > 2$ & $a_2 > 1$

$$||x - x_1||_{inf} = \begin{cases} a_1 - 2 & \text{when } a_1 - 2 > a_2 - 1 \\ a_2 - 1 & \text{when } a_2 - 1 > a_1 - 1 \end{cases}$$

#2 → if $a_1 < 1$ and $a_2 > 1$ then

$$||x - x_1||_{inf} = \begin{cases} a_1 - 1 ; & a_1 + a_2 < 3 \\ a_2 - 1 ; & a_1 + a_2 > 3 \end{cases}$$



#3 → if $a_1 > 2$ and $a_2 < 1$

$$||x - x_1||_{inf} = \begin{cases} a_1 - 2 ; & a_1 + a_2 > 3 \\ a_2 - 1 ; & a_1 + a_2 < 3 \end{cases}$$

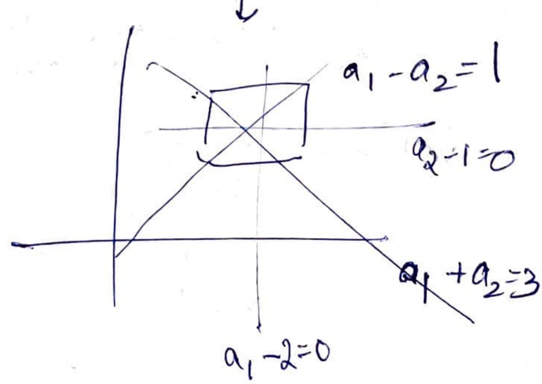
#4 → $a_1 < 2$ and $a_2 < 2$

$$||x - x_1||_{inf} = \begin{cases} a_1 - 2 ; & 2 - a_1 > 1 - a_2 \\ a_2 - 1 ; & 2 - a_1 < 1 - a_2 \end{cases}$$

$$= \begin{cases} a_1 - 2 ; & a_1 - a_2 < 1 \\ a_2 - 1 ; & a_1 - a_2 > 1 . \end{cases}$$

Max $|\, |a_1 - 2|, |a_2 - 1| \,|$

↓



iv) undirected attack

step size $= 0.001, 0.1, 1$ given $f(x) = \text{sigmoid}(w'x)$

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

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

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

$$\text{grad } f(x) = \frac{\exp(-3)}{[1+\exp(-3)]^2} [1 \quad 1]$$

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

(#) $x_{attack} = x_1 - \alpha \nabla_x f(x)$

$\alpha = 0.001 \rightarrow x_{attack} = [1.9995 \quad 0.9999]$ ✓ in threat model

$\alpha = 0.1 \rightarrow x_{attack} = [1.995, \ 0.995]$ ✓ in threat model

$\alpha = 1 \rightarrow x_{attack} = [1.995, \ 0.955]$ ✓ in threat model

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

$\alpha = 0.001 \rightarrow x_{attack} = [1.999, 0.999]$ ✓ in threat model

$\alpha = 0.01 \rightarrow x_{attack} = [1.9, 0.9]$ ✓ in threat model

$\alpha = 1 \rightarrow x_{attack} = [1 \quad 0]$ ✓ not in threat model.

v) Directed attack

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

$$\frac{\partial f(x)}{\partial x} = (-0.045, \ -0.045)$$

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

now $x_{attack} = x_1 + \alpha \nabla_x f(x)$. $\rightarrow$

$\cdot \exp(-w^T x)$

case 1 $x_{attack} = x_1 + \alpha \nabla_x f(x)$

$\alpha = 0.001 \rightarrow (1.9995, 0.9995)$ ✓ in threat model

$\alpha = 0.1 \rightarrow (1.995, 0.995)$ ✓ in threat model

$\alpha = 1 \rightarrow (1.955, 0.955)$ ✓ in threat model

Case 2 x_{attack} $x_1 + \alpha\, \mathrm{sgn}(\nabla_x f(x))$

$(1.999, 0.999)$ ✓ in threat model

$(1.9, 0.9)$ ✓ in threat model

$(1, 0)$ ✗ not in threat model

vi) ~~poor~~ OBSERVATIONS: for all points, in case of using actual gradient, the attack is within threat model but for the case when we use sgn(grad f) we get one point to be outside threat model thus indicating a stronger attack with high α.

however, results are same for target & untargeted so doesn't seem to make a difference for just 2 x's i.e x_1, x_2