Backpropogation

30 March 2022 13:18

Backprop → Algo → train nn (How)

training

Pre requisite

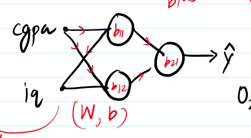
→ Gradient Descent (✓) →

→ Forward Propogation (✓) →

cgpa	iq	lpa
8	80	8
7	70	7

reg

predict



6+3 → 9 trainable

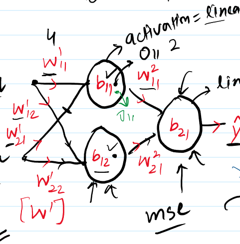
$\sigma(z)$

$O_{21} = W_{11}^2 O_{11} + W_{21}^2 O_{12} + b_{21}$

Backprop

4 students

iq	cgpa	lpa	error
80	8	3	225
60	9	5	
70	5	8	
120	7	11	



$\hat{y} = O_{21}$ (W, b) Random val

$\boxed{W \to 1, b \to 0}$

L min → $\hat{y} \to O_{21}$

Steps →

0) Init W, b ✓

1) You select a point (row) ✓

↳ student

2) Predict (lpa) → forward prop [Dot product]

3) Choose a loss function (mse)

4) Weights and bias update ✓

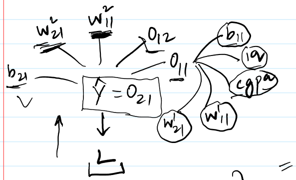
↳ Gradient Descent

→ $\boxed{18 \text{ lpa}}$

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

$(3-18)^2 = 225$ error

Hierarchy ↓ Backprop



$$\boxed{W_{new} = W_{old} - \eta \frac{\partial L}{\partial W_{old}}} \checkmark$$

=1 (0.1)

$$\boxed{W_{11\,new}^2 = W_{11\,old}^2 - \eta \frac{\partial L}{\partial W_{11}^2}}$$ ← 9 derivatives

$$\boxed{b_{new} = b_{old} - \eta \frac{\partial L}{\partial b_{old}}} \checkmark$$

$$\boxed{W_{21\,new}^2 = W_{21\,old}^2 - \eta \frac{\partial L}{\partial W_{21}^2}} \checkmark$$

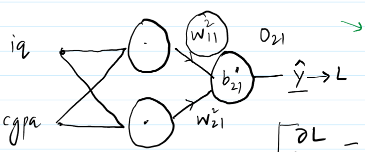
$$\boxed{b_{21\,new} = b_{21\,old} - \eta \frac{\partial L}{\partial b_{21}}} \checkmark$$

derivative of loss wrt weight

$\boxed{\frac{\partial L}{\partial W_{11}^2}}$ ← partial derivative

derivative $\frac{\partial L}{\partial W_{11}^2}$

$\frac{\partial L}{\partial W_{11}^2}, \frac{\partial L}{\partial W_{21}^2}, \frac{\partial L}{\partial b_{21}}$ | $\frac{\partial L}{\partial W_{11}^1}, \frac{\partial L}{\partial W_{21}^1}, \frac{\partial L}{\partial b_{11}}$ | $\frac{\partial L}{\partial W_{12}^1}, \frac{\partial L}{\partial W_{22}}, \frac{\partial L}{\partial b_{12}}$

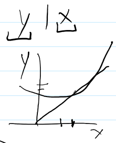


$$\frac{\partial L}{\partial W_{11}^2} = \boxed{\frac{\partial L}{\partial \hat{y}}} \times \frac{\partial \hat{y}}{\partial W_{11}^2}$$ → chain rule of differen

W_{11} → $\hat{y}$ → L (min)

$\left\{\frac{dy}{dx}\right\}$

y | x



$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}} (y - \hat{y})^2 = \boxed{-2(y - \hat{y})}$$

$$\frac{\partial \hat{y}}{\partial W_{11}^2} = \frac{\partial}{\partial W_{11}^2} [O_{11} W_{11}^2 + O_{12} W_{21}^2 + b_{21}]$$

$$= O_{11}$$

W_{21}^2 → $\hat{y}$ → L, $\frac{\partial L}{\partial W_{21}^2}$

$$\boxed{\frac{\partial L}{\partial W_{11}^2} = -2(y - \hat{y}) O_{11}} \quad 1$$

$$\frac{\partial L}{\partial W_{21}^2} = \boxed{\frac{\partial L}{\partial \hat{y}}} \times \frac{\partial \hat{y}}{\partial W_{21}^2}$$

$$\frac{\partial \hat{y}}{\partial W_{21}^2} = \frac{\partial}{\partial W_{21}^2} [O_{11} W_{11}^2 + O_{12} W_{21}^2 + b_{21}]$$

$$= O_{12}$$

$$\boxed{\frac{\partial L}{\partial W_{21}^2} = -2(y - \hat{y}) O_{12}} \quad 2$$

$\hat{y}$ → b, L

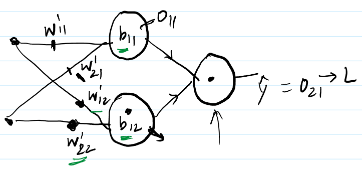
$$\frac{\partial \hat{y}}{\partial b_{21}} = 1$$

$$\boxed{\frac{\partial L}{\partial w^2_{21}} = -2(y-\hat{y})O_{12}} \quad 2$$

$\hat{y} \rightarrow b$, $L \rightarrow \hat{y}$

$$\frac{\partial \hat{y}}{\partial b_{21}} = 1$$

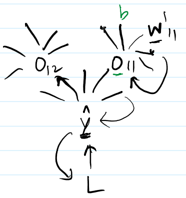
$$\boxed{\frac{\partial L}{\partial b_{21}} = -2(y-\hat{y})} \quad 3 \qquad \frac{\partial L}{\partial b_{21}} = \boxed{\frac{\partial L}{\partial \hat{y}}} \times \frac{\partial \hat{y}}{\partial b_{21}}$$



$$\frac{\partial L}{\partial w^1_{11}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{11}} \frac{\partial O_{11}}{\partial w^1_{11}} \rightarrow \boxed{-2(y-\hat{y}) w^2_{11} x_{i1}} \quad 4$$

$$\frac{\partial L}{\partial w^1_{21}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{11}} \frac{\partial O_{11}}{\partial w^1_{21}} \rightarrow \boxed{-2(y-\hat{y}) w^2_{11} x_{i2}} \quad 5$$

$$\frac{\partial L}{\partial b_{11}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{11}} \frac{\partial O_{11}}{\partial b_{11}} \rightarrow \boxed{-2(y-\hat{y}) w^2_{11}} \quad 6$$



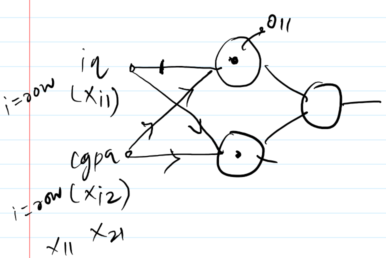
$$\frac{\partial L}{\partial w^1_{12}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial w^1_{12}} \rightarrow \boxed{-2(y-\hat{y}) w^2_{21} x_{i1}} \quad 7$$

w^2_{11}

$$\frac{\partial L}{\partial w^1_{22}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial w^1_{22}} \rightarrow \boxed{-2(y-\hat{y}) w^2_{21} x_{i2}} \quad 8$$

$$\frac{\partial \hat{y}}{\partial O_{12}} = w^2_{21}$$

$$\frac{\partial L}{\partial b_{12}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial b_{12}} \rightarrow \boxed{-2(y-\hat{y}) w^2_{21}} \quad 9$$



$$\frac{\partial O_{11}}{\partial w^1_{11}} = \frac{\partial}{\partial w^1_{11}}[iq\, w^1_{11} + cgpa\, w^1_{21} + b_{11}] = (iq) \quad \uparrow x_{i1}$$

$$\frac{\partial O_{11}}{\partial w^1_{21}} = x_{i2}$$

$$\frac{\partial O_{11}}{\partial b_{11}} = 1$$

$$\frac{\partial O_{12}}{\partial w^1_{12}} = \frac{\partial}{\partial w^1_{12}}[iq\, w^1_{12} + cgpa\, w^1_{22} + b_{12}] = iq\,(x_{i1})$$

$$\frac{\partial O_{12}}{\partial w^1_{22}} = x_{i2}$$

$$\frac{\partial O_{12}}{\partial b_{12}} = 1$$

multiple chance

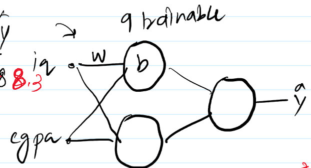
epochs

cgpa	iq	lpa	$\hat{y}$
8	80	8	8.3
6	60	6	
7	70	7	
9	90		

4 stu

9 trainable

$L\downarrow$



Steps (once again)

0) weights/bias → init → random, w=1, b=0

loop - 100/100

converge

1) for i in range(4):

$(y-\hat{y})^2$ → Backprop

1a) 1 student → forward prop → lpa) →

12 min

1b) Loss calculate (mse) →

1c) Adjust all weights and bias

The How

GD

$$\boxed{W_{new} = W_{old} - \eta \frac{\partial L}{\partial W_{old}}} \rightarrow 9 \text{ times}$$

Backpropagation Algorithm

```
epochs=5
for i in range(epochs):
    for j in range(x.shape[0]):
```

$$\frac{\partial L}{\partial w^2_{11}} = -2(y-\hat{y})O_{11}$$

$$\frac{\partial L}{\partial w^2_{21}} = -2(y-\hat{y})O_{12}$$

Ready

for i in range(epochs):

for j in range(x.shape[0]):

→ Select 1 row (random)

→ Predict (using Forward prop)

→ Calculate loss (using Loss function → mse)

→ Update weights and bias using GD

$W_n = W_0 - \eta \frac{\partial L}{\partial W}$

→ Calculate avg loss for the epoch

L_{avg}

$\frac{\partial L}{\partial w^2_{21}} = -2(y-\hat{y})O_{12}$

$\frac{\partial L}{\partial b_{21}} = -2(y-\hat{y})$

$\frac{\partial L}{\partial w^1_{11}} = -2(y-\hat{y})w^2_{11}x_{i1}$

$\frac{\partial L}{\partial w^1_{21}} = -2(y-\hat{y})w^2_{11}x_{i2}$

$\frac{\partial L}{\partial b_{11}} = -2(y-\hat{y})w^2_{11}$

$\frac{\partial L}{\partial w^1_{12}} = -2(y-\hat{y})w^2_{21}x_{i1}$

$\frac{\partial L}{\partial w^1_{22}} = -2(y-\hat{y})w^2_{21}x_{i2}$

$\frac{\partial L}{\partial b_{12}} = -2(y-\hat{y})w^2_{21}$

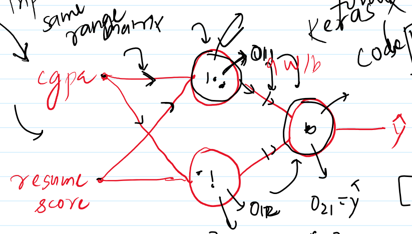
Regression Problem

80-120 NN inputs same range

0-10 cgpa (x_{i1})	resume score (x_{i2})	package (lpa)
8	8	4
7	9	5
6	10	6
5	12	7

4 students

code → A[0] [8 8]



keras code A[2] A[1] ↑ final

[2 2 1]

2 2 A[2] 9 W/b

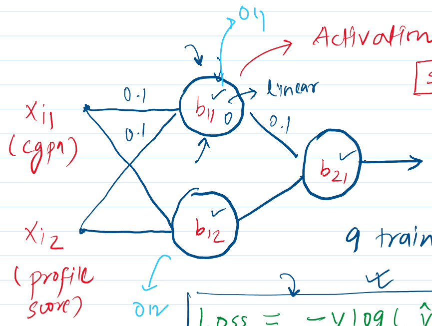
A[1] w = 0.1 b = 0

[O_{11} O_{12}] 6 3

$\hat{y} = A[2] = O_{21}$

Classification Example

cgpa	profile score	placement
8	8	1
7	9	1
6	10	0
5	5	0



Activation function $\boxed{\text{sigmoid}}$

$Z = 0.1 \times 8 + 0.1 \times 8 + 0$

$\sigma(z) = O_{11}$

9 trainable parameters

$\boxed{Loss = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})}$ $(y-\hat{y})^2$

binary cross entropy

Backpropagation Algorithm

epochs = 5

for i in range(epochs): (rows)

for j in range(x.shape[0]):

→ Select 1 row (random) → 1 by 1

→ Predict (using Forward prop) → predict

→ Calculate loss (using Loss function → ~~mse~~ bce) → { code convert }

→ Update weights and bias using GD

Backpropagation Algorithm

Cross entropy

epochs = 5 ✓

for i in range(epochs): (rows)

  for j in range(x.shape[0]):

  → Select 1 row (random) → 1 by 1

  → Predict (using Forward prop) → predict

  → Calculate loss (using Loss function → ~~mse~~ bce) → {code convert}

  → Update weights and bias using GD

$$W_n = W_0 - \eta \boxed{\frac{\partial L}{\partial W}}$$

9 w, b update

$z = W_{11}^2 O_{11} + W_{21}^2 O_{12} + b_{21}$

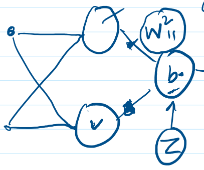
→ Calculate avg loss for the epoch

L_{avg}

$\hat{y} = \sigma(z)$

$$L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

der wrt L



$\hat{y} \to L$

$$\boxed{\frac{\partial L}{\partial W_{11}^2}} = \boxed{\frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z}} \times \frac{\partial z}{\partial W_{11}^2} = \boxed{-(y-\hat{y})O_{11}} \quad (3)$$

$-(y-\hat{y})$, O_{11}

$$\frac{\partial L}{\partial W_{21}^2} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial W_{21}^2} = \boxed{-(y-\hat{y})O_{12}}$$

O_{12}

$z_{final} = W_{11}^2 O_{11} + W_{21}^2 O_{12} + b_{21}$

$$\frac{\partial L}{\partial b_{21}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial b_{21}} = \boxed{-(y-\hat{y})}$$

1

z_{final}

$\hat{y} \to W_{11}^2 \times W_{21}^2$

$L \to$

$$\boxed{\frac{\partial L}{\partial \hat{y}}} = \frac{\partial}{\partial \hat{y}}\left[-y\log(\hat{y}) - (1-y)\log(1-\hat{y})\right]$$

$$= \frac{-y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})} = \frac{-y(1-\hat{y}) + \hat{y}(1-y)}{\hat{y}(1-\hat{y})} = \frac{-y + \cancel{y\hat{y}} + \hat{y} - \cancel{y\hat{y}}}{\hat{y}(1-\hat{y})}$$

$$= \boxed{\frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})}} \qquad \frac{\partial}{\partial z}\sigma(z) = \sigma(z)[1-\sigma(z)]$$

$\hat{y} = \sigma(z)$

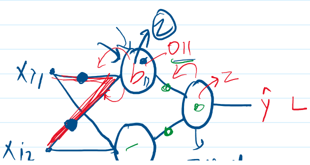
$\sigma(z) = \hat{y}$

$$\boxed{\frac{\partial \hat{y}}{\partial z}} = \frac{\partial}{\partial z}(\sigma(z)) = \sigma(z)[1-\sigma(z)] = \hat{y}(1-\hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})} \qquad \frac{\partial \hat{y}}{\partial z} = \hat{y}(1-\hat{y}) \Rightarrow \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} = \frac{-(y-\hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \times \cancel{\hat{y}(1-\hat{y})}$$

$$\to -(y-\hat{y})$$

$$\frac{\partial z_p}{\partial W_{11}^1} = W_{11}^1 X_{i1} + W_{21}^1 X_{i2} + b_{11}$$

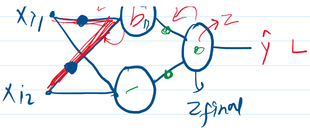


$$\frac{\partial L}{\partial W_{11}^1}, \quad \frac{\partial L}{\partial W_{21}^1}, \quad \frac{\partial L}{\partial b_{11}}$$

$\to -(y-\hat{y})$

$z_f = W_{11}^2 O_{11} + W_{21}^2 O_{12} + b_{21}$ $\quad O_{11}(1-O_{11})$

$O_{11} = \sigma(z_{prev})$



$\frac{\partial L}{\partial w'_{11}}$, $\frac{\partial L}{\partial w'_{21}}$, $\frac{\partial L}{\partial b_{11}}$

$z_f = w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}$ $\quad O_{11}(1-O_{11})$

$O_{11} = \sigma(z_{prev})$

$= \sigma(z_{prev})[1-\sigma(z_{prev})]$

$-(y-\hat{y})$

$$\frac{\partial L}{\partial w'_{11}} = \boxed{\frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_f}} \times \frac{\partial z_f}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial z_{prev}} \times \frac{\partial z_{prev}}{\partial w'_{11}}$$

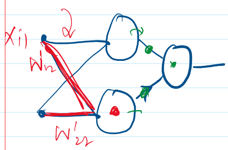
w^2_{11} $\quad O_{11}(1-O_{11})$ $\quad x_{i1}$

$L \rightarrow \hat{y} \rightarrow z_f \rightarrow O_{11} \rightarrow z_{prev} \rightarrow w'_{11}$

$$\boxed{\frac{\partial L}{\partial w'_{11}} = -(y-\hat{y})\, w^2_{11}\, O_{11}(1-O_{11})\, x_{i1}}$$ ③

$$\boxed{\frac{\partial L}{\partial w'_{21}} = -(y-\hat{y})\, w^2_{11}\, O_{11}(1-O_{11})\, x_{i2}}$$

$$\boxed{\frac{\partial L}{\partial b_{11}} = -(y-\hat{y})\, w^2_{11}\, O_{11}(1-O_{11})}$$



$\frac{\partial L}{\partial w'_{12}}$, $\frac{\partial L}{\partial w'_{22}}$, $\frac{\partial L}{\partial b_{12}}$

$\frac{\partial z_f}{\partial O_{12}} = w^2_{11} O_{11} + w^2_{21} O_{12} + b_{21}$

$$\frac{\partial L}{\partial w'_{12}} = \boxed{\frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_f}} \times \frac{\partial z_f}{\partial O_{12}} \times \frac{\partial O_{12}}{\partial z_p} \times \frac{\partial z_p}{\partial w'_{12}}$$

w^2_{21} $\quad O_{12}(1-O_{12})$ $\quad x_{i1}$

$L \rightarrow \hat{y} \rightarrow z_f \rightarrow O_{12} \rightarrow z_p \rightarrow w'_{12}$

$$\boxed{\frac{\partial L}{\partial w'_{12}} = -(y-\hat{y})\, w^2_{21}\, O_{12}(1-O_{12})\, x_{i1}}$$ ③

$$\boxed{\frac{\partial L}{\partial w'_{22}} = -(y-\hat{y})\, w^2_{21}\, O_{12}(1-O_{12})\, x_{i2}}$$

$$\boxed{\frac{\partial L}{\partial b_{12}} = -(y-\hat{y})\, w^2_{21}\, O_{12}(1-O_2)}$$

Backpropagation → The why?

$\boxed{\text{The intuition behind the algorithm}}$ ✓

Backpropagation Algorithm

epochs = 5

for i in range(epochs): — # epochs x rows

    for j in range(x.shape[0]):

        → select 1 row (random)

        → Predict (using Forward prop) — $\hat{y}$ y

        → Calculate loss (using loss function → mse) — bce

        → Update weights and bias using GD

            ↳ $W_n = W_0 - \boxed{\eta \frac{\partial L}{\partial W}}$ rule

            magic

    → Calculate avg loss for the epoch

    L_{avg}

→ Loss function is a function of all trainable parameters

data → target

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

cgpa | profile || lpa

$L = (y - \hat{y})^2$ → mathematical function

constant

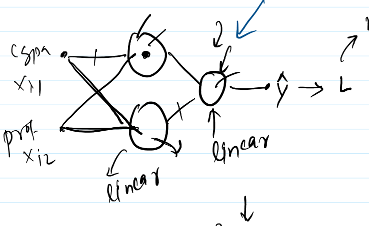
$L(?)$

$L(\hat{y}) =$

$y = f(x)$ 1 thing

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

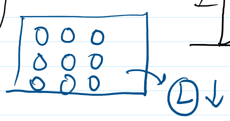
variable

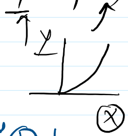


$L(w, b)$ ↓

9 things

w, b





$\hat{y} = W_{11}^2 O_{11} + W_{21}^2 O_{12} + b_{21}$

$\hat{y} = W_{11}^2 [W_{11}^1 x_{i1} + W_{21}^1 x_{i2} + b_{11}] + W_{21}^2 [W_{12}^1 x_{i1} + W_{22}^1 x_{i2} + b_{12}] + b_{12}$

→ Concept of Gradient → Gradient Descent

↳ fancy word for derivative

$y = f(x) = x^2 + x$

$\frac{dy}{dx} = \frac{d}{dx}(fx) = \frac{d}{dx}(x^2 + x) = 2x + 1$

derivative

$y \rightarrow x \rightarrow$ derivative $\frac{d}{dx}$

$z = f(x, y) = \boxed{x^2 + y^2}$

$\frac{\partial z}{\partial x} = 2x$

$\frac{\partial z}{\partial y} = 2y$

gradient $\boxed{\frac{\partial}{\partial x}}$

gradient complex → $\frac{\partial L}{\partial w}, \frac{\partial L}{\partial b}$

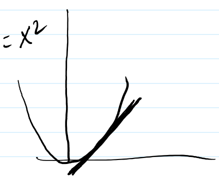
↳ 9 parameters

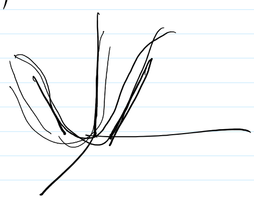
$L(W_{11}^1, W_{12}^1 \ldots \ldots W_{21}^2, b_{21}, b_{11}, b_{12})$

↳ 9D function → 9 diff slopes wrt each dim

3D → $z = x^2 + y^2$ $z = f(x, y)$

$y = f(x) = x^2$





→ Concept of Derivative → Derivative at a point

intuition

derivative

$\boxed{x=5}$ → deriv

$\frac{dy}{dx}$ → rate of change

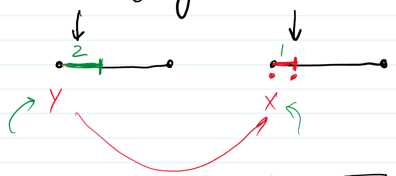
y func x → x

$\frac{dy}{dx} = 2\ (+ve)$ → mag, sign

$\frac{dy}{dx} = -2\ (-ve)$

$x \rightarrow$

$y \rightarrow$



$y = x^2 + 2x$

$\frac{dy}{dx} = (2x+1)_{x=5}$

$\left(\frac{dy}{dx}\right)_{x=5} = 11$ → slope

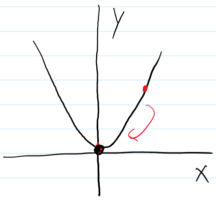
$\boxed{\frac{\partial L}{\partial w^1_{11}}}$

w^1_{11} — 1 unit

$L \rightarrow ?$ (mag) (sign)

$W_{new} = W_{old} - \eta \frac{\partial L}{\partial w}$

→ the concept of minima



$\boxed{y = x^2}$

$\frac{dy}{dx} = 2x = 0$

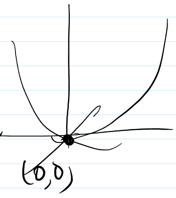
$\boxed{x=0}$

$z = x^2 + y^2$

$(x, y) \rightarrow z$ min

$\frac{\partial z}{\partial x} = 0 \quad x \rightarrow$

$\frac{\partial z}{\partial y} = 0 \quad y \rightarrow$



z min

(x, y) $(0,0)$

$2x = 0 \Rightarrow x = 0$

$2y = 0 \Rightarrow y = 0$

L (9 param)

(w, b) min

$L \downarrow$

$\frac{\partial L}{\partial w^1_{11}} \cdots \frac{\partial L}{\partial b_{12}} = 0$

9 dim → minima

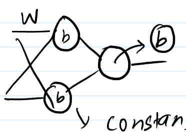
→ Backprop Intuition

$\eta = 1$

$\boxed{W_{new} = W_{old} - \eta \frac{\partial L}{\partial w}}$ ← 9 step

$\boxed{W_{new} = W_{old} - \frac{\partial L}{\partial w}}$

each parameters



constans

$L = L(w, b)$ → 9, 8

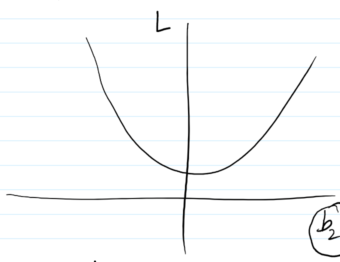
constant

$L(b_{21})$

$L = b_{21}^2$

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

$L(b_{21})$ $\hat{y} = (b_{21})$



$b_{21} = b_{21} - \boxed{\frac{\partial L}{\partial b_{21}}}$ → +ve, −ve

$b_{21} = 5$

b_{21} L min

$\boxed{\frac{\partial L}{\partial b_{21}}}$

smart

L derivative wrt b_{21}

? 1 unit

$L \downarrow$

$b_{21} \downarrow$

$b_{21} \downarrow$

$L \downarrow$

$\frac{\partial L}{\partial b_{21}} = \underline{+ve}$ → $b_{21} \uparrow$ $L \uparrow$

$b_{21} \uparrow$

$\frac{\partial L}{\partial b_{21}} = -ve$

$b_{21} \uparrow$ $L \downarrow$

$b_{21} \uparrow$

game

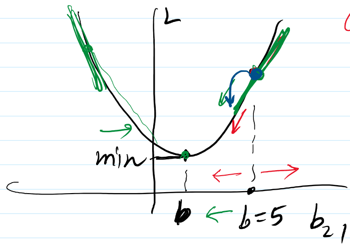
−ve of the gradient

$b_{21} = b_{21} - \frac{\partial L}{\partial b_{21}}$

game derivative b_{21}

−ve of the gradient

$$b_{21} = b_{21} - \boxed{\frac{\partial L}{\partial b_{21}}}$$



algo

$\frac{\partial L}{\partial b}$ = slope

+ve

b = −5

$\frac{\partial L}{\partial b}$ = −ve

−ve

slope (gradient) → −ve

$\boxed{\text{slope}}$

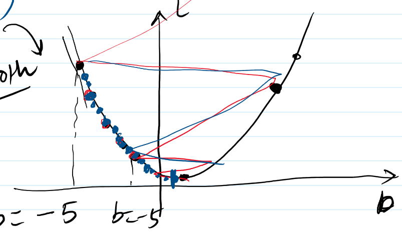
$$W_n = W_0 - \boxed{\frac{\partial L}{\partial w}} \qquad \boxed{\eta} \qquad \boxed{W_n = W_0 - \eta \frac{\partial L}{\partial w}}$$

→ Effect of Learning Rate (η)

smooth

$$W_n = W_0 - \eta \boxed{\frac{\partial L}{\partial w}}$$

0.1/0.01



train

$\boxed{\eta = 1}$

η = parameter → 0.01

η = 0.00001

$\frac{\partial L}{\partial b} = -50$

b = −5 − (−50) = 45

b = 45

$\frac{\partial L}{\partial b}$ = 50

b = 45 − 50

b = −5 $\quad \frac{\partial L}{\partial b} = -50 \quad \eta = 0.01$

b = −5 + (0.01 × 50) = 0.5

= −4.5

→ What is convergence ?

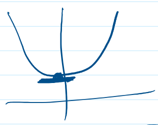
$$W_n = W_0 - \eta \frac{\partial L}{\partial w} \approx 0$$

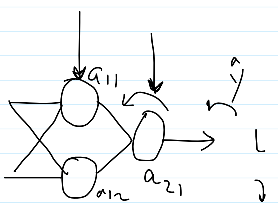
0

$\boxed{W_n = W_0}$

while → con → for 1000

$\boxed{\text{epoch} \to 100}$





i = 1, 0

$-(y - \hat{y})$

activation = ②

$\hat{y}(1 - \hat{y})\, O_{11}$

$-y \log(a_{21}) - (1-y) \log(1 - a_{21})$

$-(y - \hat{y})\, O_{11}$

$$\frac{\partial L}{\partial w_{11}^2} = \boxed{\frac{\partial L}{\partial a_{21}}} \times \frac{\partial a_{21}}{\partial w_{11}^2}$$

$$-\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}$$

$\frac{\partial L}{} =$

$$-\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}$$

$$\frac{\partial L}{\partial w^2_{21}} =$$

curr_activa = a[1]

$$\frac{\partial L}{\partial b_{21}} =$$

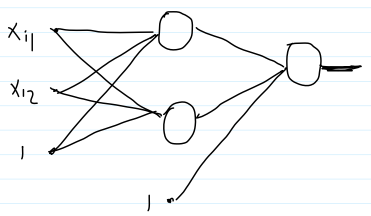
$$\sigma\left\{\begin{matrix} -\boxed{(y-\hat{y})\quad a_{21}(1-a_{21})\,a_{11}} \\ -(y-\hat{y})\,a_{21}(1-a_{21})\,a_{12} \\ -(y-\hat{y})\,a_{21}(1-a_{21}) \end{matrix}\right\}$$

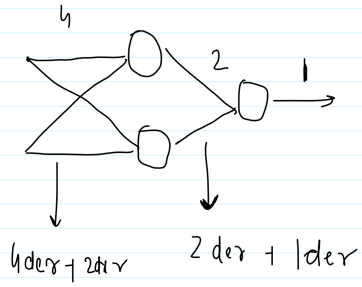
$$\text{delta} = (y-\hat{y})\ a^{(2)}(1-a^{(2)})$$

1×1 $\cancel{2\times1}$ $\boxed{1\times2}$

$\boxed{W[1]}$ $\boxed{1\times2}$

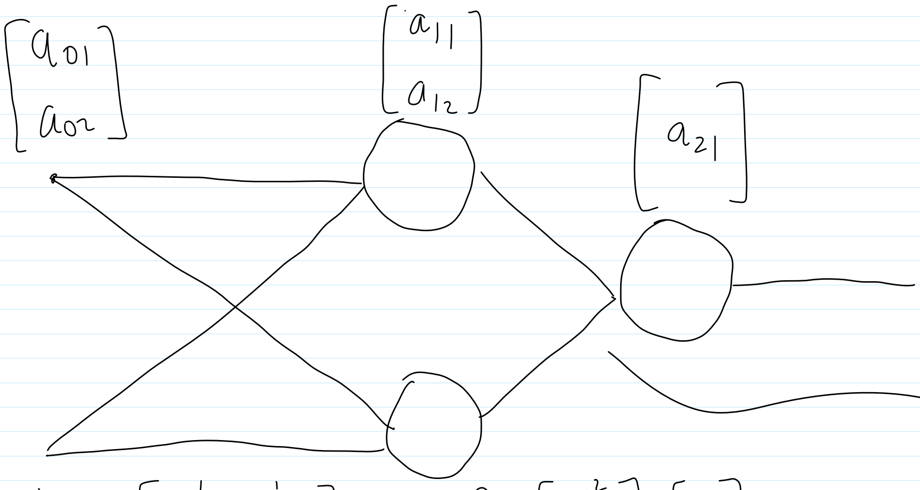
$$\boxed{-(y-\hat{y})\,a_{21}(1-a_{21})} * \underset{}{a[1]}\begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix}$$





1×1 $\quad 1\times2$

$1\times2 \quad \times \quad 2\times2$



L

$L \rightarrow a_2 \rightarrow a_1 \rightarrow$

$$\frac{\partial L}{\partial a_2} \times \frac{\partial a_2}{\partial a_1}\ \frac{\partial a_1}{\partial w}$$

$$\frac{\partial L}{\partial w} = \boxed{\frac{\partial L}{\partial a_2}} \times \boxed{\frac{\partial a_2}{\partial w}}$$

$$\begin{bmatrix} W^1_{11} & W^1_{12} \\ W^1_{21} & W^1_{22} \end{bmatrix} \quad \begin{bmatrix} b_{11} \\ b_{12} \end{bmatrix} \quad \begin{bmatrix} W^2_{11} \\ W^2_{12} \end{bmatrix} \quad \begin{bmatrix} b_{21} \end{bmatrix}$$

$$\begin{bmatrix} \partial L/\partial w^1_{11} & \frac{\partial L}{\partial w^1_{12}} \\ \frac{\partial L}{\partial w^1_{21}} & \frac{\partial L}{\partial w^1_{22}} \end{bmatrix} \qquad \begin{bmatrix} \frac{\partial L}{\partial w^2_{11}} \\ \frac{\partial L}{\partial w^2_{12}} \end{bmatrix}$$