Day 3 - Deep Learning.

Agenda

① Optimizers

1) Gradient Descent
2) SGD (Stochastic Gradient Descent)
3) Mini Batch SGD
4) SGD with Momentum
5) Adagrad
6) RMSPROP
7) Adam optimizer

Batch, Epochs, Iterations

⇓

ANN

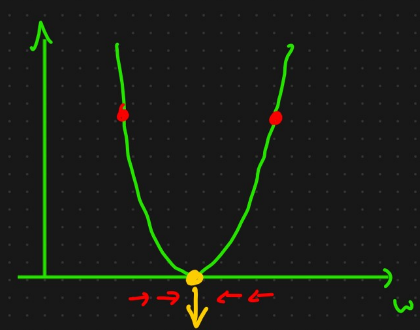
① GRADIENT DESCENT → Optimizer

Weight Updation Formula

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

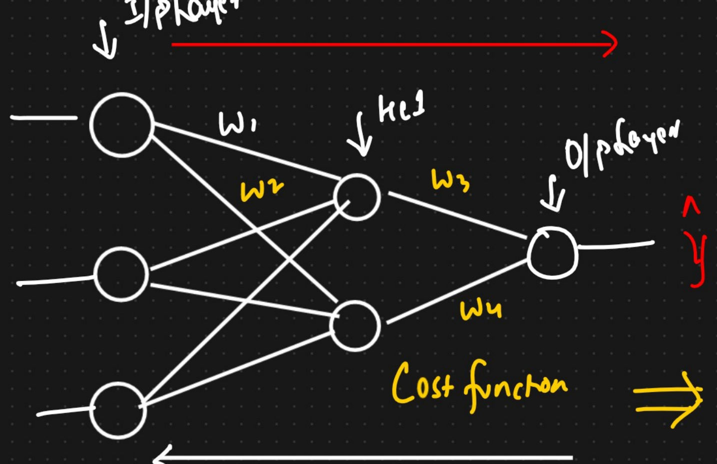
Learning Rate

Loss | Cost



Global Minima.

I/P layer



MSE ⇓ Optimizers

Cost function $\Rightarrow \dfrac{1}{2n} \sum_{i=1}^{n} (y - \hat{y})^2 \downarrow\downarrow$ }

Epoch 1000000 $\rightleftarrows$ } 1 Epoch { 1000000 }

Disadvantage

① Resource Extensive {Huge RAM}

1000000

② Stochastic Gradient Descent

Epoch 1 → 100 epochs

1 record → $\hat{y}$ ← Loss ← Update weights } → Iteration 1

2 record → ← } Iteration 2

⋮

} Million Iteration

$\frac{1000000}{1000}$

① RAM ↓↓

Disadvantage

① Convergence will be very slow

② Time complexity will also be high

③ Mini batch SGD

1000000 batch_size = 1000

Epoch-1

1000 → ← } Iteration 1

1000 → ← } Iteration 2

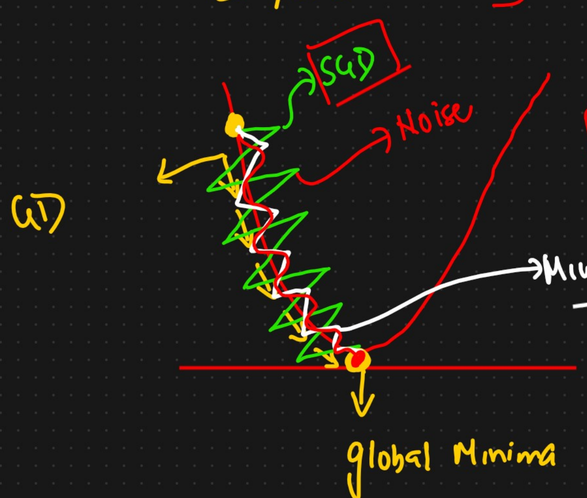
⋮

→ ← } Iteration 1000

① Resource Intensive

② Convergence will be better

③ Time Complexity will Improve



Remove Noise ?? ⇓ Momentum

(4) SGD With Momentum

{Exponential Weighted Average}

⇓

Time Series

⇓

ARIMA, ARMA,

$$W_{new} : W_{old} - \eta \frac{\partial L}{\partial w_{old}}$$

$$b_{new} : b_{old} - \eta \frac{\partial L}{\partial b_{old}}$$

$$\boxed{w_t = w_{t-1} - \eta \frac{\partial L}{\partial w_{t-1}}}$$

Exponential Weighted Average

{Forecasting}

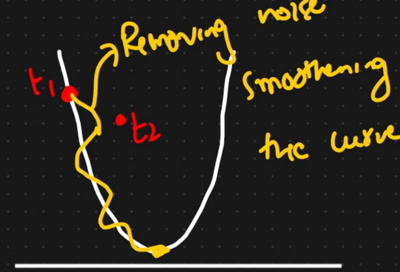
t_1	t_2	t_3	t_4	$\cdots$	t_n
a_1	a_2	a_3	a_4	$\cdots$	a_n

β ⇒ Hyper parameter

$\beta = 0$ to 1

⇓

0.95



$$V_{t_1} = a_1$$

$$V_{t_2} = \beta * V_{t_1} + (1-\beta) * a_2$$

$$= (0.95) * V_{t_1} + (0.05) * a_2$$

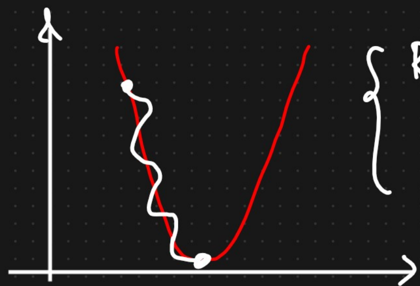
$$V_{t_3} : \beta \times V_{t_2} + (1-\beta) * a_3$$

Exponential Weighted Avg

⇑

$$W_t = W_{t-1} - \eta V_{dw}$$

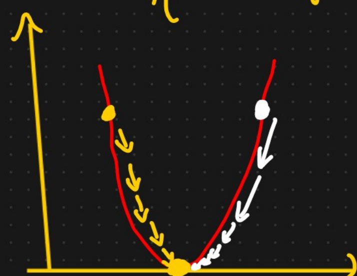
$$\boxed{V_{dw_t} = \beta \times V_{dw_{t-1}} + (1-\beta) * \frac{dL}{\partial w_{t-1}}}$$



Reduce the noise
Minibatch
Quicker Convergence

Recap

{ fixed } ⇒ optimizer
⇑
η = Learning Rate



① Gradient Descent

② SGD

③ Mini batch SGD

④ SGD with Momentum

⑤ Adagrad → Adaptive Gradient Descent

η = fixed ⇒ adaptive ⇒ Learning Rate ⇒ Decreasing → Global Minima

Global Minima
⇓
$\eta \rightarrow$
↓
↓ Decreasing ↓↓↓

$$w_t = w_{t-1} - \boxed{\eta} \frac{dL}{dw_{t-1}}$$

⇓

$$\boxed{w_t = w_{t-1} - \eta' \frac{\partial L}{\partial w_{t-1}}}$$

$$\boxed{w_t \approx w_{t-1}}$$

$$\downarrow\downarrow\downarrow \ \eta' = \frac{\eta}{\sqrt{\alpha_t + \epsilon}}$$

0.01

ϵ → Small number

$$\alpha_t = \sum_{i=1}^{t} \left(\frac{\partial L}{\partial w_t}\right)^2 \uparrow\uparrow\uparrow$$

Huge number

$t=1$ $\quad$ $t=2$ $\quad$ $t=3$

$\eta = 0.01$ $\quad$ $\eta = 0.005$ $\quad$ $\eta = 0.002$

⑥ Adadelta And RMSprop

Exponential Weighted Average
⇓

$Sdw_t = 0$

$$\eta' = \frac{\eta}{\sqrt{Sdw + \epsilon}}$$

$$Sdw_t = \beta \, Sdw_{t-1} + (1-\beta)\left(\frac{\partial L}{\partial w_{t-1}}\right)^2$$

$\beta = 0.95$

$$\boxed{Sdw_t} = (0.95)\, Sdw_{t-1} + (0.05)\left(\frac{\partial L}{\partial w_{t-1}}\right)^2$$

(5) Adam Optimizer (Best Optimizer)

Momentum + RMSPROP (Adaptive Learning Rate)

{ (1) Smoothening (2) Learning Rate Adaptive }

$Vdw = 0 \quad Vdb = 0 \quad Sdw = 0 \quad Sdb = 0$

$$W_t = W_{t-1} - \eta' \, Vdw$$
$$b_t = Wb_{t-1} - \eta' \, Vdb$$

$$\eta' = \frac{\eta}{\sqrt{Sd_w + \epsilon}}$$

$$Vdw_t = \beta \times Vdw_{t-1} + (1-\beta)\frac{\partial L}{\partial w_{t-1}}$$

$$Vdb_t = \beta \times Vdb_{t-1} + (1-\beta)\frac{\partial L}{\partial b_{t-1}}$$