Data Science & AI

Machine Learning

Artificial Neural Network

One Shot

By- Krish Naik Sir

Recap of Previous Lecture

Topic

Topic

Topic

Topic

Topic

Deep Learning

Topics to be Covered

- Topic — Optimizers
- Topic — Drop outliers
- Topic
- Topic
- Topic

Deep Learning

① Optimizers ✓

② Weight Initialization Technique ✓ } {BATCH NORMALIZATION}.

③ DROPOUT LAYERS ✓ .

① Optimizers

① Gradient Descent ✓

② Stochastic Gradient Descent ✓

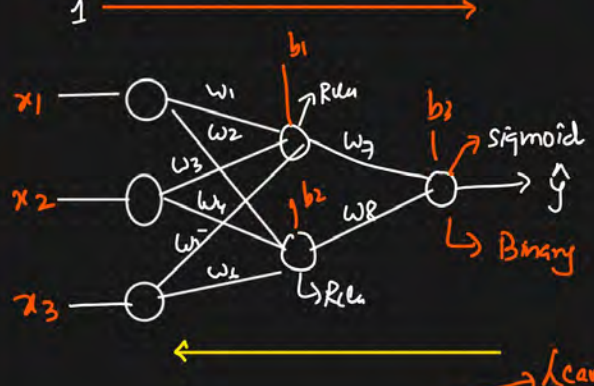
③ Mini batch SGD ✓

④ SGD with Momentum ✓ .

⑤ Adagrad and RMSprop

⑥ Adam Optimizers . ← Best

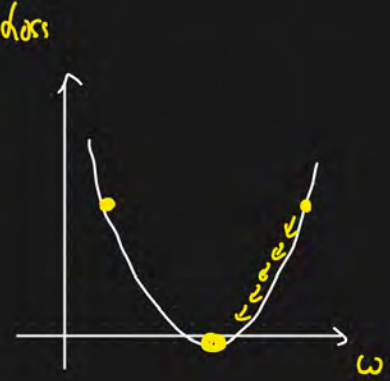
① GRADIENT DESCENT Optimizer



MSE, RMSE, MAE

Loss = [Error.] ↓↓

Gradient Descent



Learning Rate.

$$W_{new} = W_{old} - \eta \left[\frac{\partial L}{\partial W_{old}} \right]$$

DATASET = 1000 datapoints

Batch size = $1000/10 = 100$

1000 iteration

10 Iteration

1000 datapoints → TRAINING

Laptop

Epoch 1 → 1 record → Iteration

Epoch 1000

200

Epoch 1 — 100 record — $\sum_{i=1}^{100} (y_i - \hat{y}_i)^2$

Epoch 2 — 100 record

Time Training ↑↑.

TRAINING SET = 1 Million Records.

System Hang ⇒ Nvidia GPU data center.

Cost ↑↑ ⇓ Highly Expensive.

Gradient Descent ⇒ Convergence.

1 Epoch = 1 iteration

100 Epochs.

Advantages

① Convergence will happen

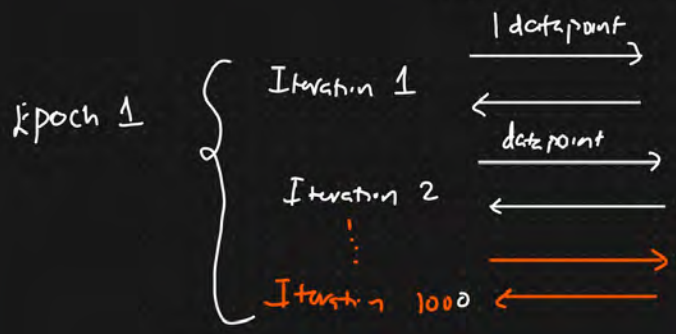
Disadvantage

① Huge Resource, RAM, GPU

② Stochastic Gradient Descent

batch_size = 1

Epoch & Iteration



Training = 1000 records

Epoch = 50, 100, 150, 200

Optimizers → Changing weight And bias.

Advantage

① Solves Resource Issue

Disadvantage

① Time Complexity ↑↑

② Convergence will take more time

③ Noise get introduced

③ Mini batch SGD

Batch size = 100 Training = 1000 iteration = $\frac{1000}{100}$ = 10 iteration

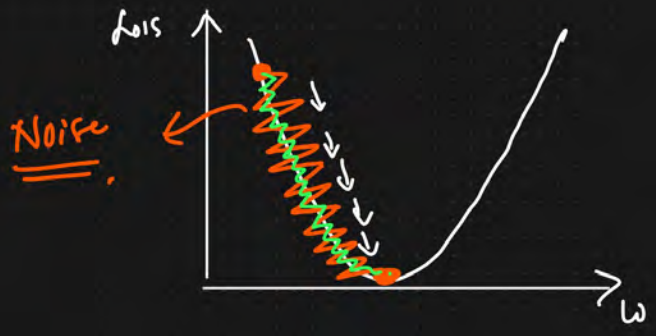
Epoc 1

Iteration 1 ⇄ Loss (100)

Iteration 10 ⇄

50, 100,

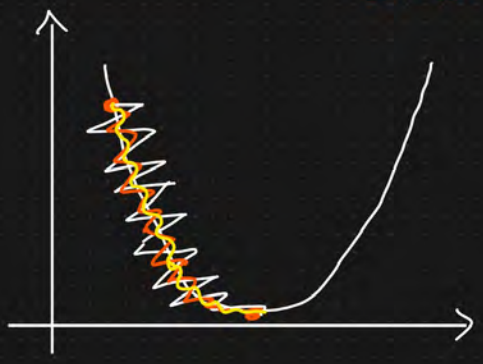
Early Stopping loss ↓ X



④ SGD With Momentum

Noise {Smoothening}

Exponential Weighted Average = ARIMA



Time	t_1	t_2	t_3	t_4	...	t_n
Values	a_1	a_2	a_3	a_4	...	a_n

β = Influence

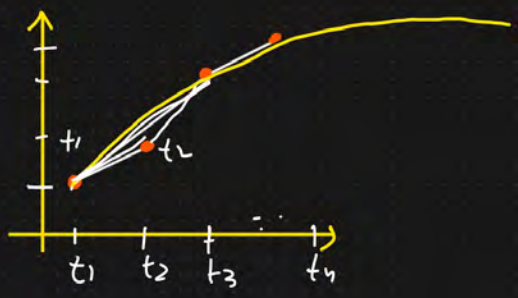
$V_{t_1} = a_1$

0.4 0.6

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

$= [0.95 * a_1 + (0.05) * a_2]$

$V_{t_3} = \beta * V_{t_2} + (1-\beta)\, a_3$



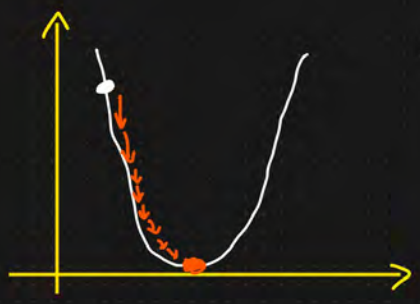
$$W_{new} = W_{old} - \eta \frac{\partial L}{\partial W_{old}} \Rightarrow \boxed{W_t = W_{t-1} - \eta \frac{\partial L}{\partial W_{t-1}}}$$

④ Adagrad → Adaptive Gradient Descent

$$\boxed{W_t = W_{t-1} - \eta \frac{\partial L}{\partial W_{t-1}}}$$

→ Learning Rate fixed.

η = fixed ⇒ Dynamic Learning Rate.



⇒ As the convergence happens the learning rate should change

$$\boxed{W_t = W_{t-1} - \eta' \frac{\partial L}{\partial W_{t-1}}}$$

$$\eta' = \frac{\eta}{\sqrt{\alpha_t + \epsilon}} \Rightarrow \text{Small value}$$

$\eta' \downarrow\downarrow\downarrow$ $\quad \alpha_t \uparrow\uparrow\uparrow$

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

$\overrightarrow{t=1}$ $\overrightarrow{t=2}$ $\overrightarrow{t=3}$

$\eta = 0.01$ $\eta = 0.005$ $\eta = 0.003$ - - - →

0.000009

Disadvantage

① η' → Possibility to become a very small value ≈ 0.

(6) Adadelta OR RMSPROP

Exponential Weight Average

$$\eta' = \frac{\eta}{\sqrt{S_{dw_t} + \epsilon}}$$

$\beta = 0.95$

$S_{dw_t} = 0$

$$S_{dw_t} = \beta * S_{dw_{t-1}} + (1-\beta)\left(\frac{\partial L}{\partial w_{t-1}}\right)^2$$

($1-\beta = 0.05$)

η' slowly reduce η' ↓

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

(7) Adam Optimizer

SGD with Momentum + RMSPROP [Dynamic LR + Smoothening]

$$w_t = w_{t-1} - \eta' V_{dw}$$

$$b_t = b_{t-1} - \eta' V_{db}$$

$$\eta' = \frac{\eta}{\sqrt{S_{dw_t} + \epsilon}} \qquad \eta' = \frac{\eta}{\sqrt{S_{db_t} + \epsilon}}$$

$$V_{dw_t} = \beta * V_{dw_{t-1}} - (1-\beta)\frac{\partial L}{\partial w_{t-1}}$$

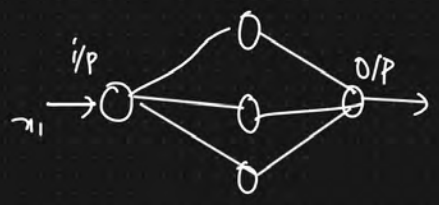
$$V_{db_t} = \beta * V_{db_{t-1}} - (1-\beta)\frac{\partial L}{\partial V_{db_{t-1}}}$$

Weight Initializing Technique

1. Uniform Distribution
2. Xavier/Glorot Initialization
3. Kaiming He Initialization

Key Points

1. Weights should be small
2. Weights should not be same
3. Weight should have good Variance



i/p = 1
o/p = 1.

1. Uniform Distribution

$$W_{ij} \sim \text{Uniform Distribution}\left[\frac{-1}{\sqrt{input}}, \frac{1}{\sqrt{input}}\right]$$

2. Xavier/Glorot Initialization

Researcher → Xavier Glorot

1. Xavier Normal Init

$$W_{ij} \sim N(0, \sigma)$$

$$\sigma = \sqrt{\frac{(2)}{(input + output)}}$$

2. Xavier Uniform

$$W_{ij} \sim \text{Uniform Distr}\left[\frac{-\sqrt{6}}{\sqrt{i/p + o/p}}, \frac{\sqrt{6}}{\sqrt{i/p + o/p}}\right]$$

③ Kaiming He Initialization

① He Normal

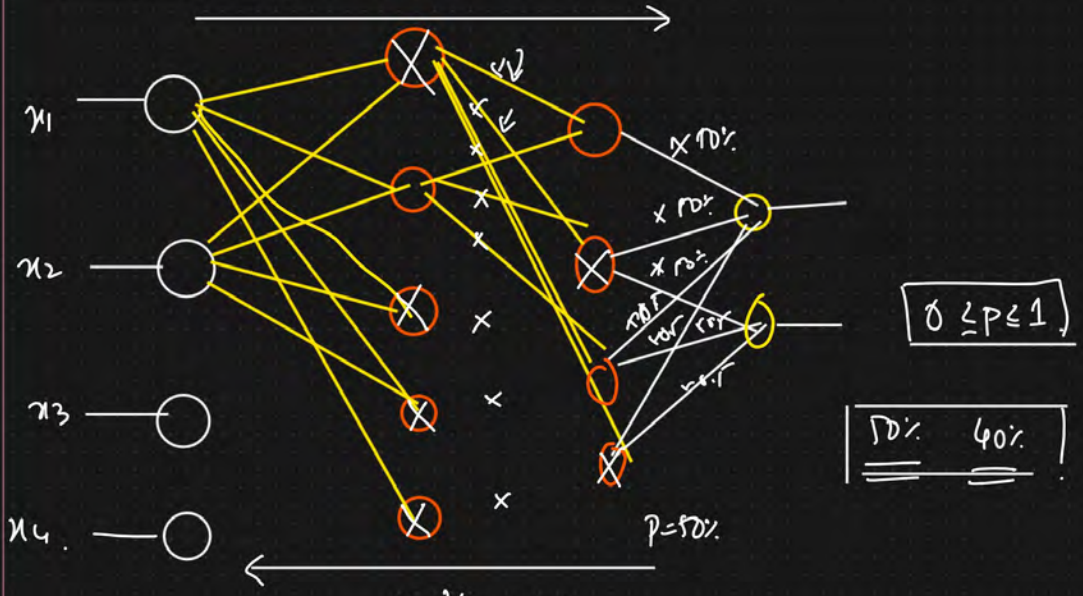
$$W_{ij} \approx N(0, \sigma)$$

$$\sigma = \sqrt{\frac{2}{i/p}}$$

② He Uniform

$$W_{ij} \approx \text{Uniform Dist}\left[-\sqrt{\frac{6}{i/p}}, \sqrt{\frac{6}{i/p}}\right].$$

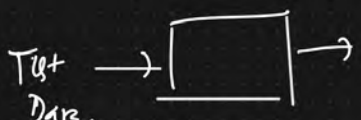
Drop Out Layer



$0 \leq p \leq 1$

50% 40%

$P = \frac{4}{5} = 40\%$



2 mins Summary

Topic One

Topic Two

Topic Three

Topic Four

Topic Five

THANK - YOU