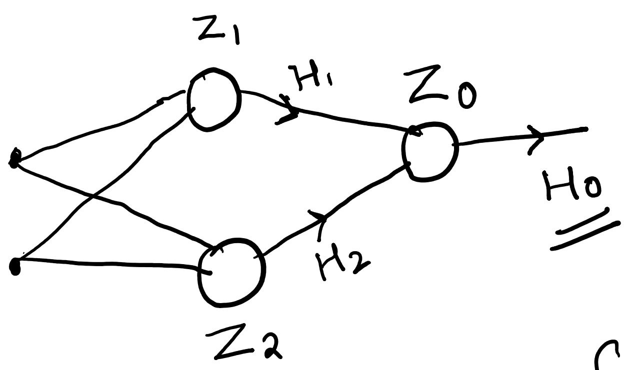
WHAT YOU WILL STUDY IN TODAY VIDEO ?

- What is a Activation Function?
- Types of Activation Functions
- Formula , Range , Use Case of Each Activation Function
- Drawbacks of Each Activation Functions

Introduction



$Z \rightarrow$ weighted sum

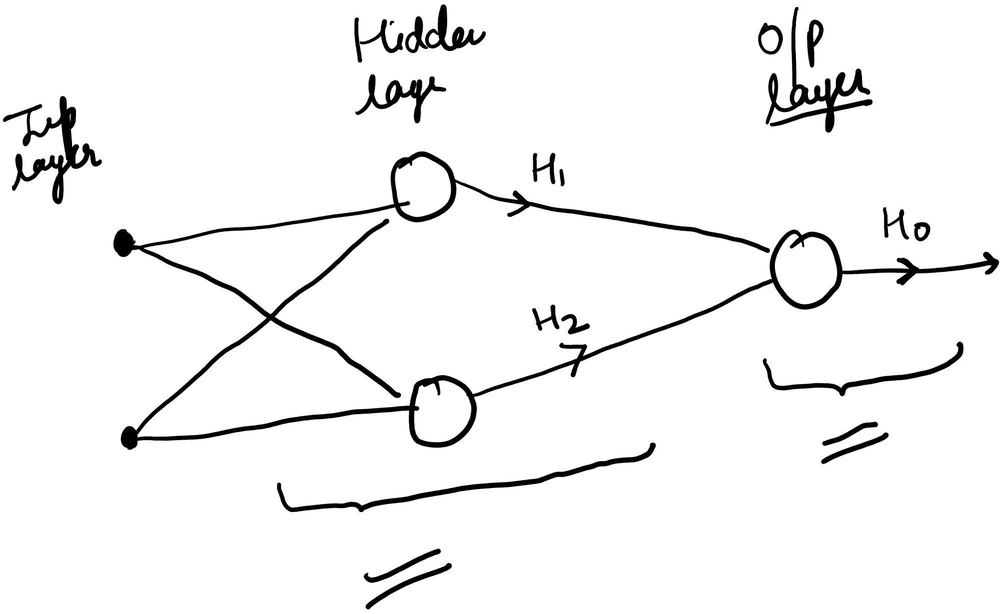
Convert weighted sum to output of neuron
$\rightarrow$ range.

$-\infty, \infty \rightarrow [0, 1]$

Types

Hidden Layer act. fns

Output layer act fn



Hidden layer

1. step fn ✓
2. Sigmoid fn ✓
3. Tanh fn ✓
4. ReLU fn ✓
5. Leaky ReLU fn ✓
6. Swish fn ✓
7. GeLU ✓

① Step fn

→ 1943

→ Perceptron only

formula

Z

$$f(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0 \end{cases}$$

Range = $\{0, 1\}$

{ } → discrete value

Use → Not used nowadays

Problem → Not useful for complex problem

Reason :- Step fn is not differentiable

Backpropagation ✗

② Sigmoid fn

↳ 1958

<u>ML</u>
{Logistic regression}

formula

$$f(z) = \frac{1}{1+e^{-z}}$$

( ) → Continous values

Range → $(0,1)$

Problem

① Vanishing Gradient Problem

② O/P is not zero-centered

O/P always +ve

↓

optimization problem

③ Tan h

→ 1965

(Hyperbolic Tangent)

formula

$$f(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

Range → $(-1, 1)$

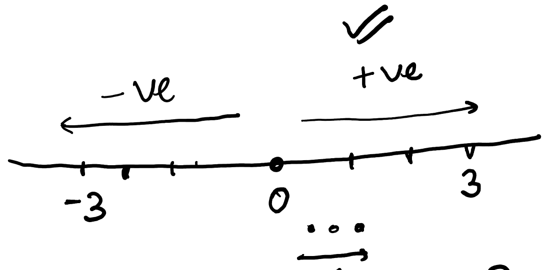
Use → Better than sigmoid
⇒ Zero centered

Problems → Vanishing Gradient Problem (still exist)

Zero Centered?

① Sigmoid → Not zero centered

Range = (0,1)
all positive



eg. $Z = -3$

$$f(z) = \frac{1}{1+e^{-(-3)}}$$

$= 0.047$

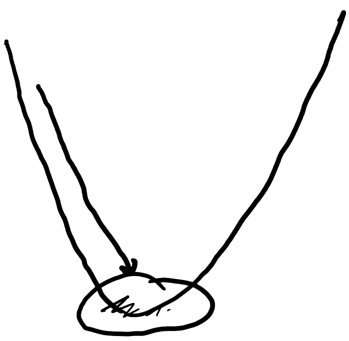
$Z = 0$

$f(z) = 0.5$

$Z = 3$

$f(z) = f(3)$

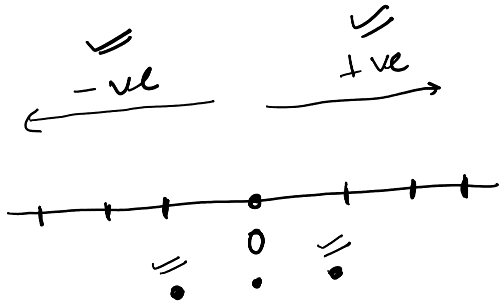
$= 0.952$



All o/p are positive
Not balanced
↓
Weight updation balanced X
↓
Too much time to converge
(optimizad X)
bad

② Tan h → Zero centered

Range = (-1, 1)



eg $Z = -3$

$Z = 0$

$f(z) = 0.$

$Z = 3$

$f(z) = 0.995$

eg
$\Rightarrow$

$z = -3$

$f(z) = -0.995$

$z = 0$

$f(z) = 0$

$z = 3$

$f(z) = 0.995$

O/p are centered around 0

↓

Gradient balanced

↓

Weight updation balanced

↓

Convergence faster
(optimization ✓)
efficient

More
efficient
optimization

(4) ReLu

ReLu → 2010

Rectified Linear Unit.

formula.

$$f(z) = \max(0, z)$$

eg.

$z = 3$

$f(z) = \max(0, 3)$

$= 3.$

$z = -1$

$= \max(0, -1)$

$= 0.$

$( ) = (0, 1)$

$[ ] = [0, 1]$

Range. → $[0, \infty)$

Use.
→ Popular (Most used)
→ Solve Vanishing Gradient Problem.

Problem → Dying ReLu Problem (neuron dead)

(... 0 (gradient = 0)}

$$\left\{ z < 0 \longrightarrow 0 \ (\text{gradient} = 0) \right\}$$

(5) Leaky ReLu

→ Advanced version ReLu
→ 2015.

formula.

$$f(z) = \begin{cases} z, & \text{if } z \geq 0 \\ 0.01z, & \text{if } z < 0 \end{cases}$$

Range. → $(-\infty, \infty)$

Use → fix dying ReLu problem

Problem → 0.01 → small no.
small gradient formed
↓
Convergence takes time
↳ optimization ✗

Parametric ReLU

ELU (Exponential Linear unit)

inspired Leaky ReLU

parameter $= (\alpha)$

Why not used ? $\longrightarrow$ Cost expensive (Computational)

⑥ Swish.

→ Google.
→ 2017.

formula.

$$f(z) = \frac{z}{1+e^{-z}}$$

Range = $(-\infty, \infty)$

Use = EfficientNet (famous architecture)
→ better accuracy

Problem → Computationall expensive than ReLu.

7 GELU → 2018. (Gaussian Error Linear Unit)

formula.

$$f(z) = 0.5z\left(1 + \tanh\left(\sqrt{2/\pi}\left(z + 0.044715z^3\right)\right)\right)$$

Range. → $(-\infty, \infty)$

Use. → BERT, GPT-3, GPT-2 → Transformers model GELU → complex architecture

Problem → Computationally expensive.

Output Layer

Classification

① Sigmoid → binary classification

② Softmax → multiclass classification

Regression.

③ Linear act fn → $f(z) \rightarrow z$ → $(-\infty, \infty)$

(for simple linear problem)

④ ReLu → $[0, \infty)$ → positive values.

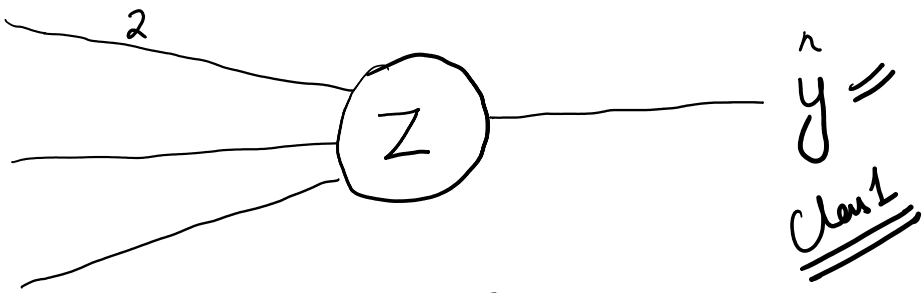
(Those regression problem where o/p is always positive)

⑤ Leaky ReLu. → $(-\infty, \infty)$ → both +ve & -ve values.

(All reg problems)

Softmax.

No of class = 3



$$Z = \underset{\text{class1}}{\left[2\right.}, \underset{\text{class2}}{1}, \underset{\text{class3}}{\left.0.1\right]}$$

① Exponent of each value

$e^2 = 7.389$ $\quad e^1 = 2.718$ $\quad e^{0.1} = 1.105$

② Sum of exponents

$\Sigma = 7.389 + 2.718 + 1.105$

$= \underline{\underline{11.212}}.$

③ Normalize ⟶ Probability

class 1 $= \dfrac{7.389}{11.212} = \underline{\underline{0.659}}.$ ✓

$\boxed{\text{O/P} \rightarrow \text{Class 1}}$

0.242.

1 11.21...

$$\text{class } 2 = \frac{2.718}{11.212} = 0.242.$$

$$\text{class } 3 = \frac{1.105}{11.212} = 0.098.$$