02 Oct 2024

Activation Functions

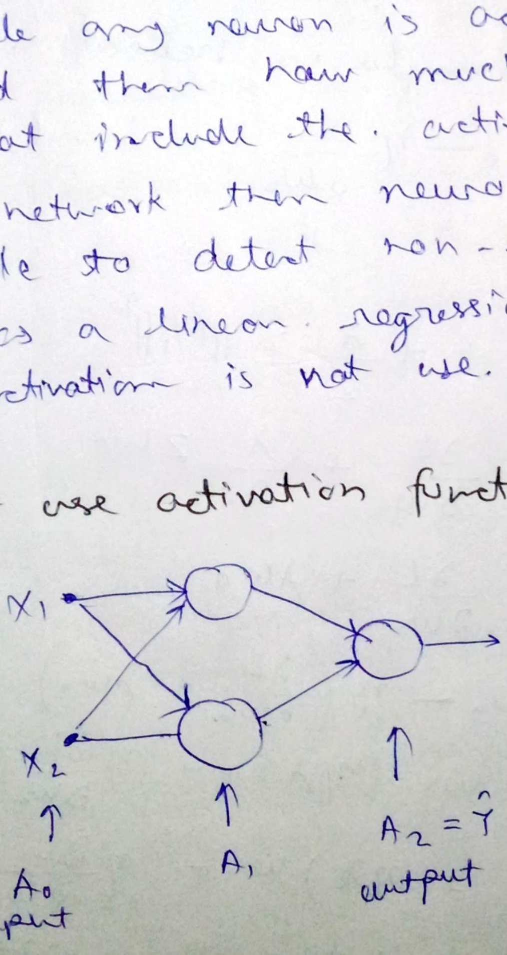
In ANN, each neuron form a weighted sum of its. inputs and. passes. the resulting scalar value through a function. referred to ~~as~~ as activation function or transfer function.. If a neuron has. 'n' inputs then the output or activation of. a neuron is

$$a = g(w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots w_n x_n + b)$$

This function g is referred to as the activation function.

- It decide any neuron is activated or not. if activated then how much?
- if we not include the. activation function in neural network then neural network ~~only~~, not able to detect non-linear features/data
- work as a linear regression or classification, if activation is not use.

if we not use activation function.



$$z_1 = w_1 A_0 + b_1$$

$$A_1 = g(z_1) = z_1$$

↘ linear function

$$A_2 = g(W_2 A_1 + b_2)$$

$$= W_2 A_1 + b_2 = W_2(W_1 A_0 + b_1) + b_2$$

$$= W_2 W_1 A_0 + W_2 b_1 + b_2$$

$$A_2 = W' A_0 + b' = \hat{Y}$$

A_0 → input

A_2 → output

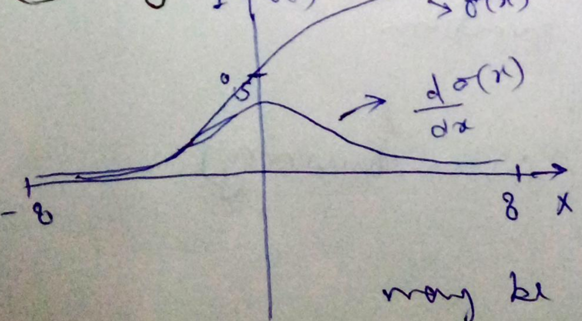
Here there is linear relationship b/w A_0, A_2

ideal Activation function:

1. Non-linear.
   example. sigmoid
2. Differentiable. ~~action~~ activation function.
   because may be we use GD
3. it should be computationally inexpensive.
4. it should be zero-centered (Normalize)
   example. tanh
5. it should be Non-saturating
   example saturating function sigmoid, tanh
   Non-Saturating function ReLU

vanishing Gradient problem.

(1.) Sigmoid function



$$\sigma(x) = \frac{1}{1+e^{-x}}$$

sigmoid is not used in hidden layer, it is always used in output.

may be used in hidden layer for binary classification.

advantages:

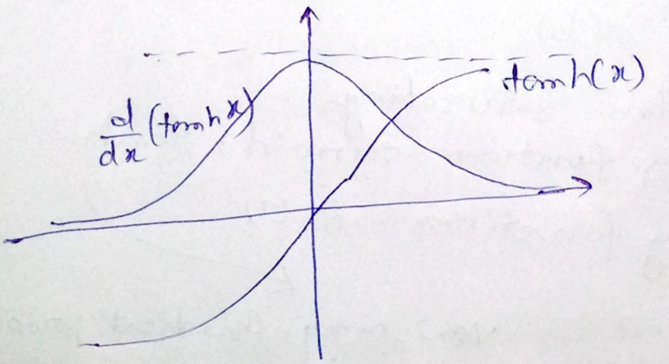
- range is [0, 1] → can be treated as probability. in output layer for binary classification.
- it is non-linear function.
- it is differentiable function.

Disadvantages:

- it is saturating function.
  input [-∞, ∞] → output [0, 1]
  due to this we face vanishing gradient problem.
- it is Non-zero centered.
- here due to this convergence time increase.
- Computationally expensive

(2.) Tanh Activation Function:

- it is also called tangent hyperbolic function.



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

$$f'(x) = (1 - \tanh^2(x))$$

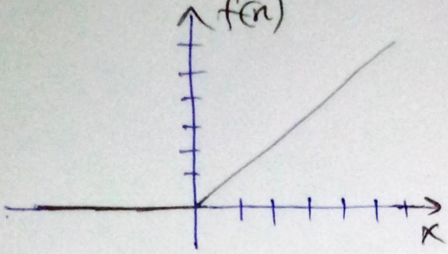
advantages:

- it is non-linear
- it is differentiable function
- it is zero centered.
- fast convergenece (faster training)

Dis advantages:

- it is saturating function.(vanishing gradient problem)
- Computationally expensive.

3. ReLU Activation function:



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

-Advantages:

- it is non-linear.
- it is not saturated in the positive region
- it is computationally. inexpensive
- convergence is faster. ~~a~~ as compared to sigmoid and tanh.

Disadvantages:

- Not differentiable at '0' (zero)

→ it is not zero centered, to resolve this issue we use batch Normalisation.

- Dying ReLU problem.