Deep Learning

Activation Functions

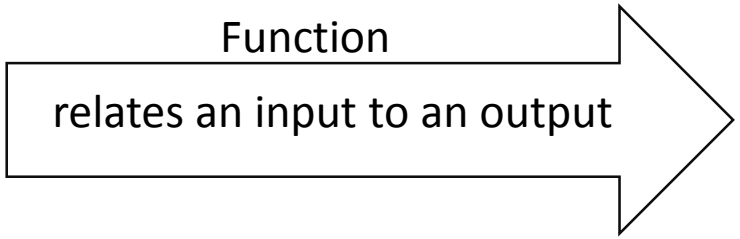
Rakshith

Walk Through

- Mathematical Functions
- Types of functions
- Activation function
- Laws of activation function
- Types of Activation functions
- Limitations of activation function

What is Function ?

Which takes some input and munch on it and generate some output



$f(x) = x^2$

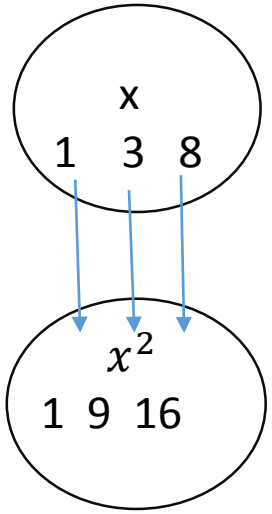
$f(q) = 1 - q + q^2$

$h(A) = 1 - A + A^2$

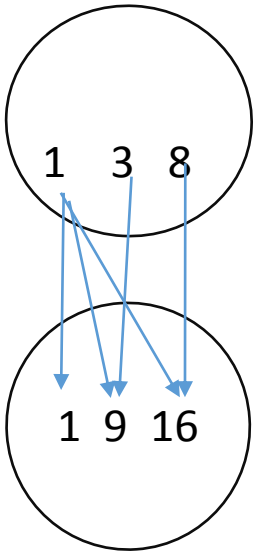
$w(\theta) = 1 - \theta + \theta^2$

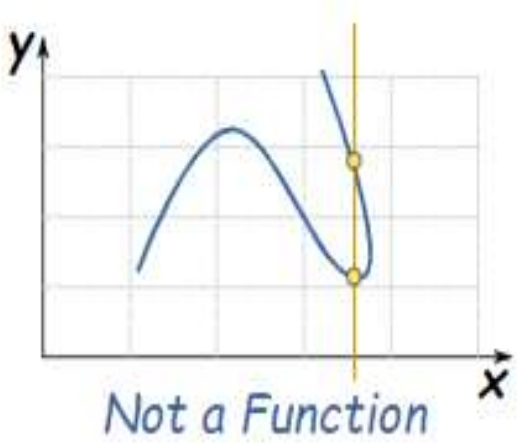
special rules:

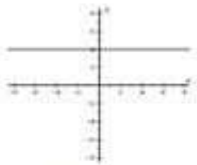
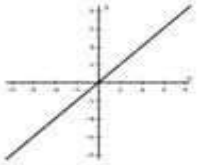
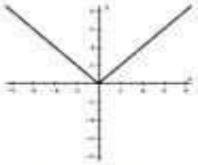
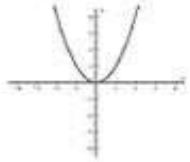
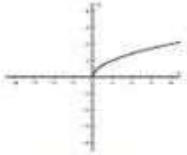
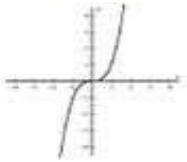
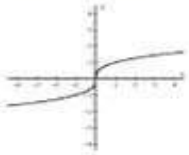
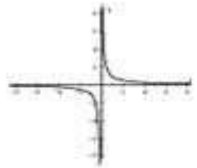
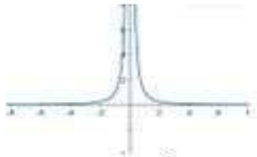
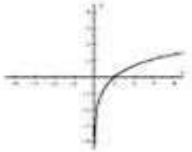
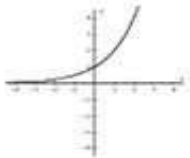
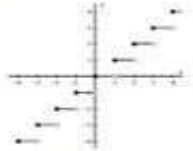
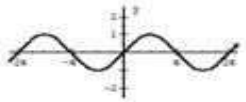
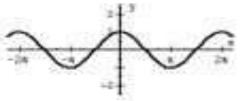
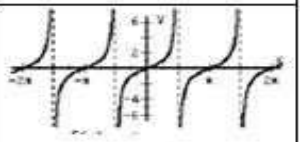
- It must work for every possible input value
- And it has only one relationship for each input value



The relationship $x \rightarrow x^2$

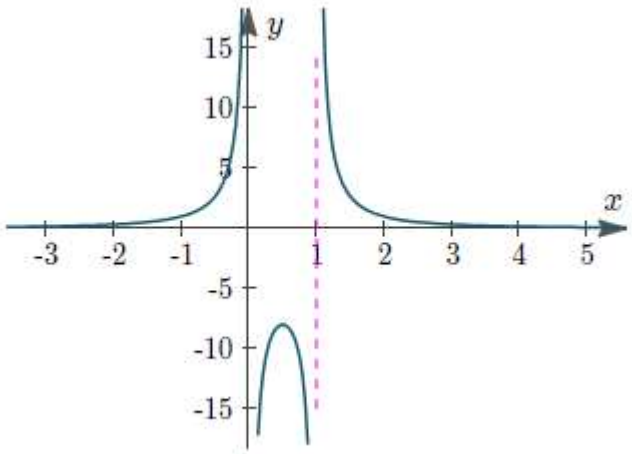




Constant  $f(x) = c$	Linear  $f(x) = x$	Absolute Value  $f(x) = \lvert x \rvert$	Quadratic  $f(x) = x^2$
Square Root  $f(x) = \sqrt{x}$	Cubic  $f(x) = x^3$	Cube Root  $f(x) = \sqrt[3]{x}$	Reciprocal/Inverse/ Rational  $f(x) = \frac{1}{x}$
Rational  $f(x) = \frac{1}{x^2}$	Logarithmic  $f(x) = \ln(x)$	Exponential  $f(x) = e^x$	Greatest Integer (Step Function)  $f(x) = [[x]]$
Trigonometric Functions →	 $f(x) = \sin(x)$	 $f(x) = \cos(x)$	 $f(x) = \tan(x)$

Kinds of Functions

A. Function With Discontinuities



plot(2/(x(x-1))

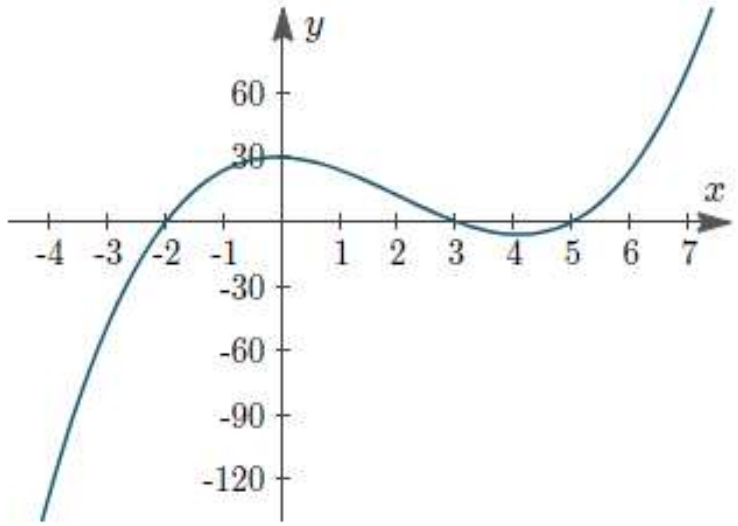
Consider the function $f(x) = \dfrac{2}{x^2 - x}$

Factoring the denominator gives: $f(x) = \dfrac{2}{x^2 - x} = \dfrac{2}{x(x-1)}$

We observe that the function **is not defined** for *x*=0 and *x*=1.
Here is the graph of the function.

We see that small changes in *x* near 0 (and near 1) produce large changes in the value of the function.
We say the function is **discontinuous** when *x* = 0 and *x* = 1.

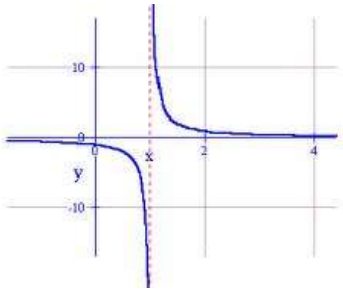
B. Continuous Functions

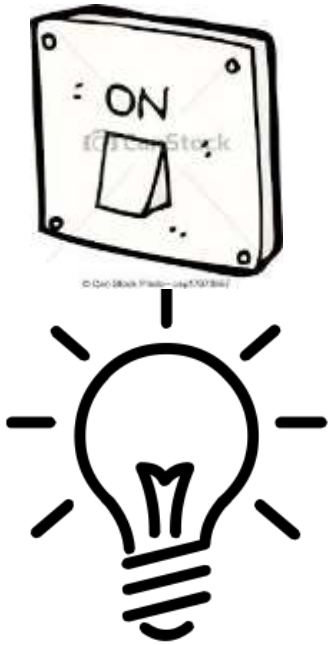


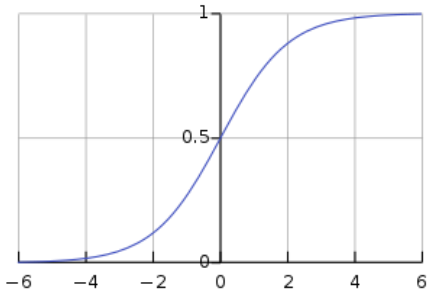
Graph of $y = x^3 - 6x^2 - x + 30$, a continuous graph.

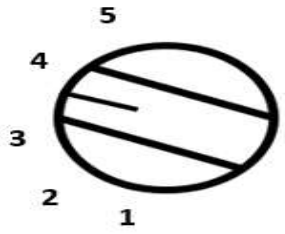
- Consider the graph of $f(x) = x^3 - 6x^2 - x + 30$ a continuous graph.
- We can see that there are no "gaps" in the curve. Any value of x will give us a corresponding value of y. We could continue the graph in the negative and positive directions, and we would never need to take the pencil off the paper.
- Such functions are called **continuous functions**.

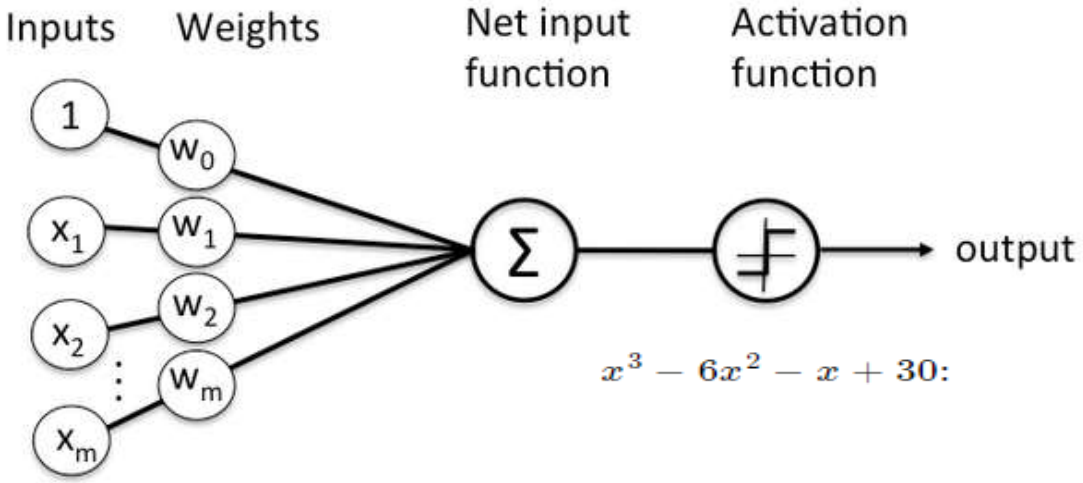
Activation function | Transfer Function





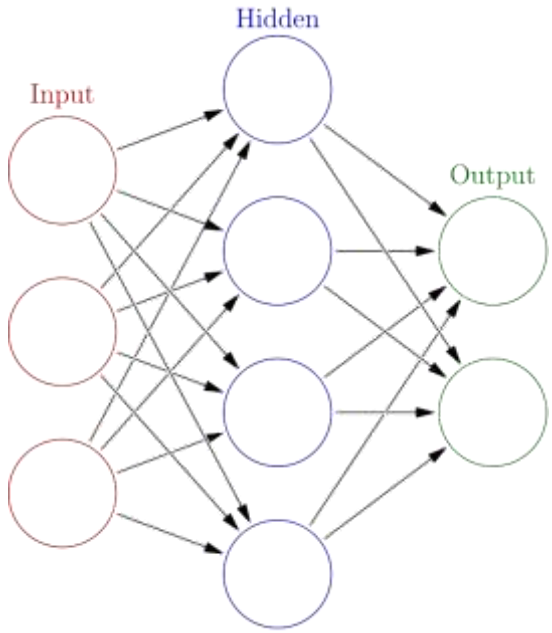






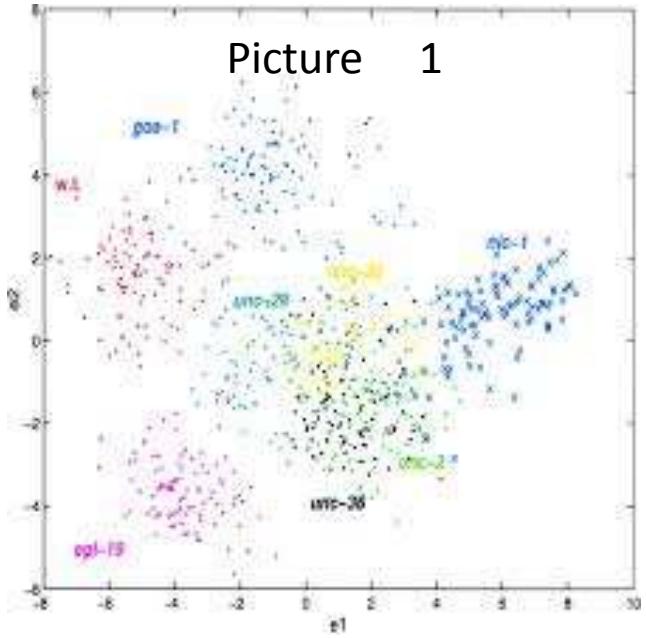
NN without activation function !

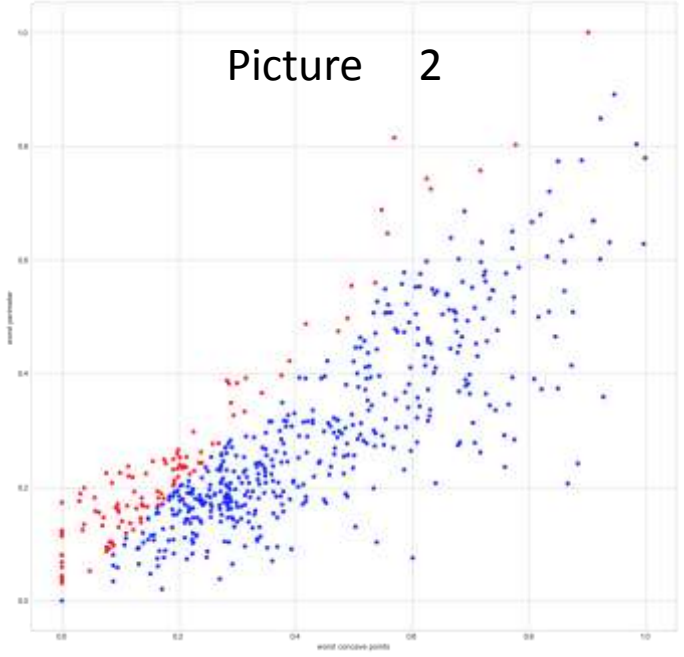
- A Neural Network without Activation function would simply be a **Linear regression Model**
- Linear function is simple polynomial of one degree
- Linear regression are easy to solve but they have less power to learn complex functional mappings from data.
- We want our Neural Network to not just learn and compute a linear function but something more complicated than that
- i.e complicated kinds of data such as images, videos , audio , speech etc.



Why activation function ?

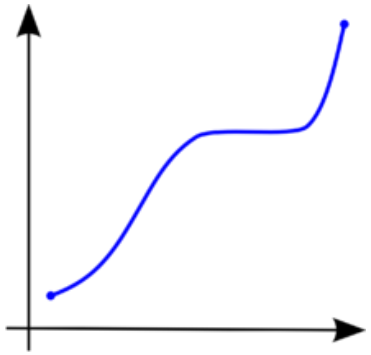
- we need to apply a Activation function f(x) so as to make the network more powerful
- Add ability to it to learn something complex and complicated from the data
- And represent non-linear complex arbitrary functional mappings between inputs and outputs. Output of one function is fed as input to next function so it can be approximated to function composition



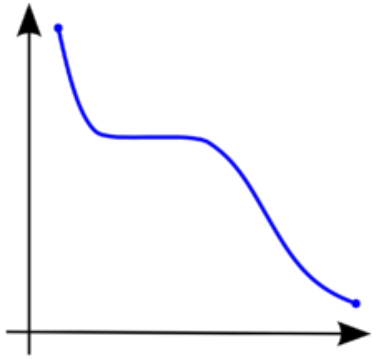


LAWS of Activation function

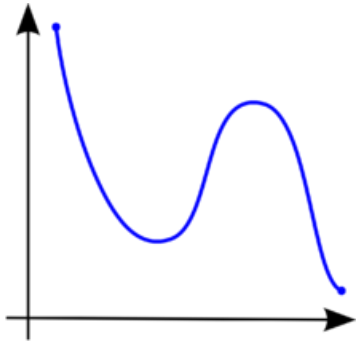
Monotonic : A function is said to be monotonically increasing if X ↑ it's corresponding Y ↑ also increases by some unit or remain constant for some time and then increases , monotonically decreasing function is vice versa it



Monotonic increasing



Monotonic decreasing

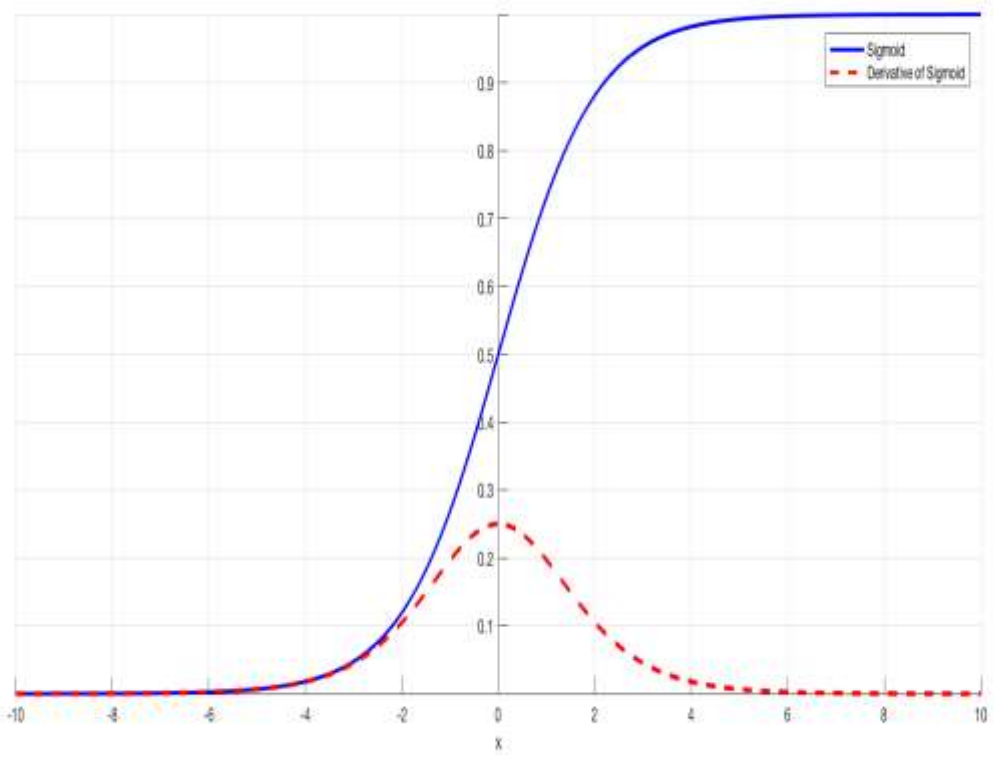


Non monotonic

Continues : A continues curve without any gape

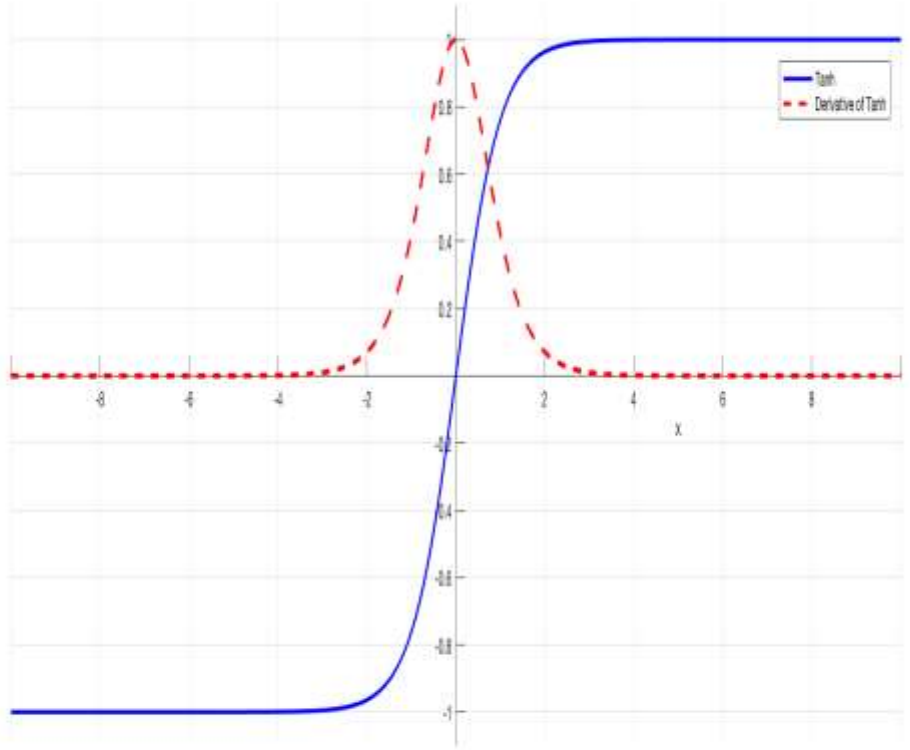
Easily differentiable : It should be smooth "**continues"** curve and differentiable at every point . Non continues .Curves can not be differentiated

Popular Activation functions



Sigmoid Function Range 0 -1

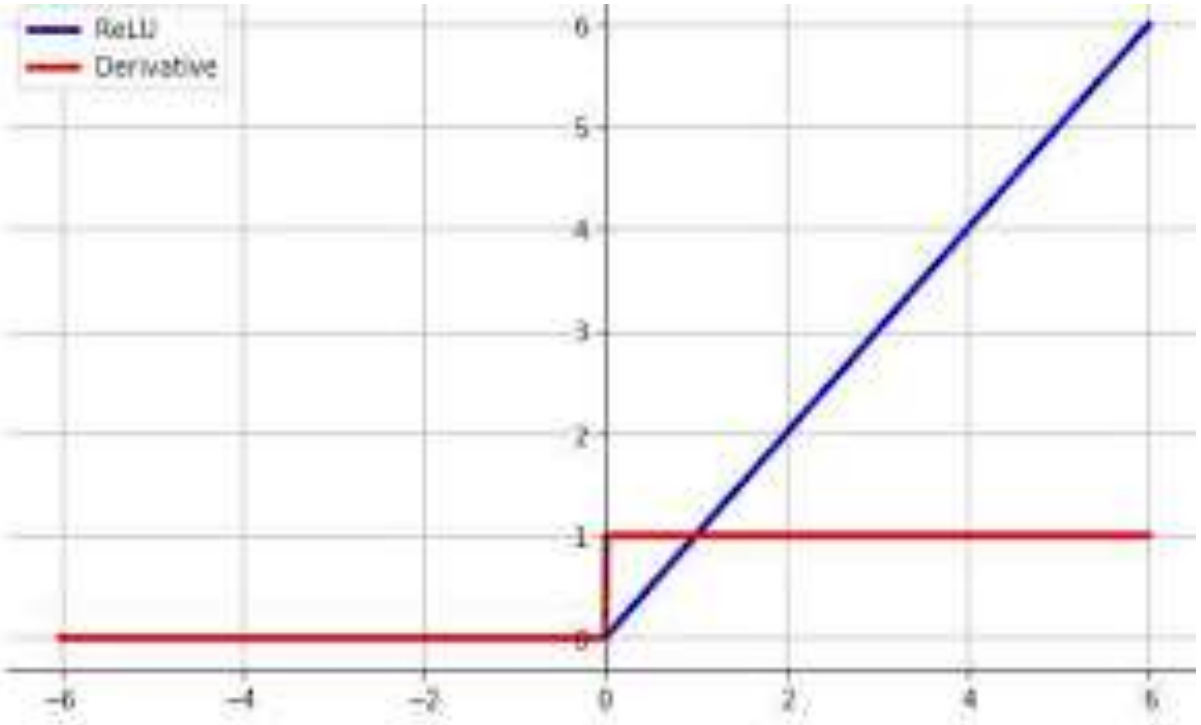
Differentiation Range 0 -0.25

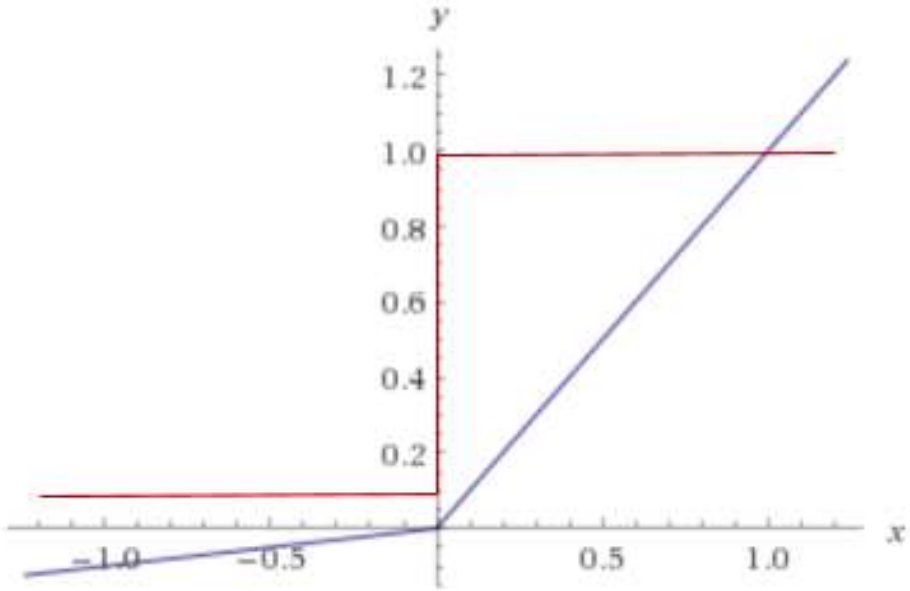


Tanh Function Range -1 + 1

Differentiation Range 0 -1

Continued…….





Rectified liner unit Range 0 , x

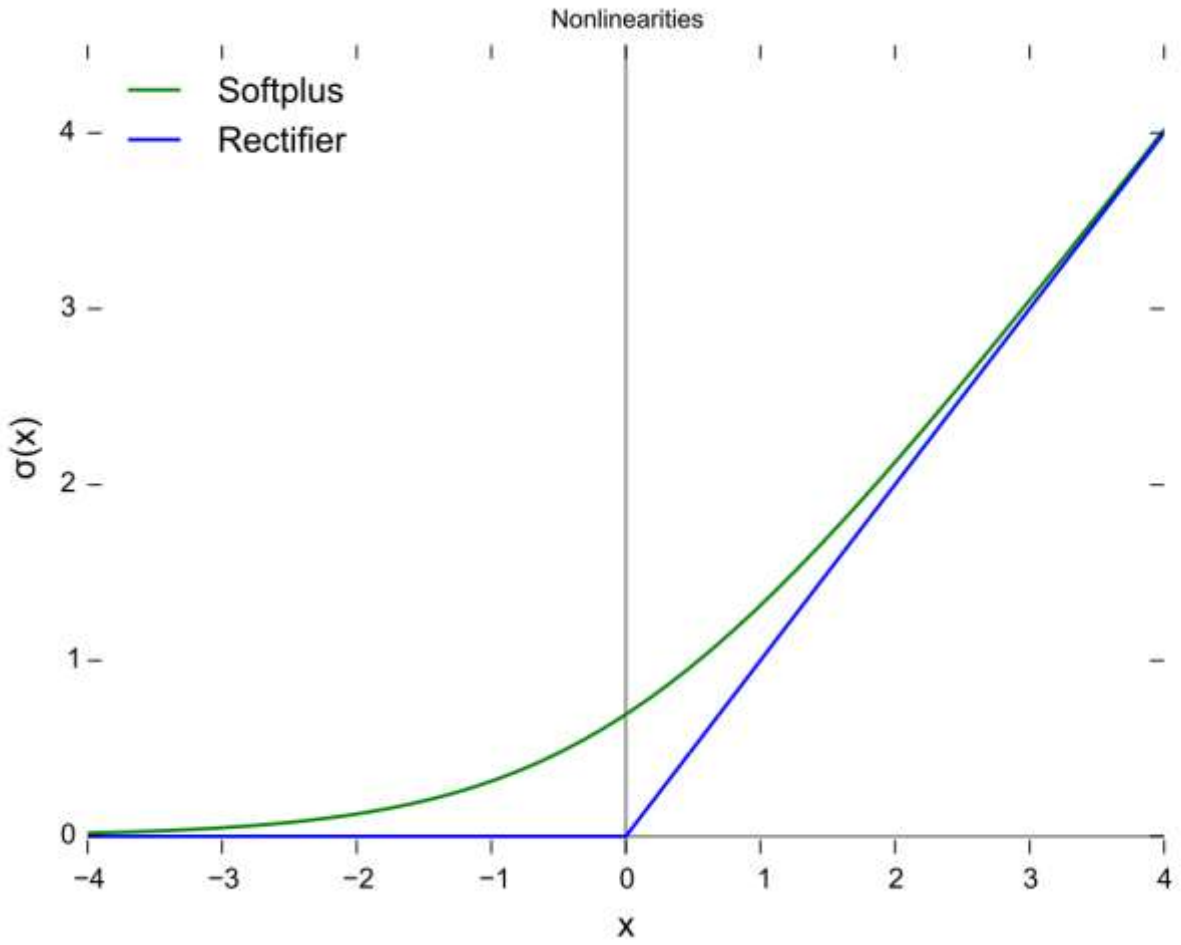
Differentiations range 0 or 1

[tan 45 = 1]

Liki Rectified liner unit Range -x , x

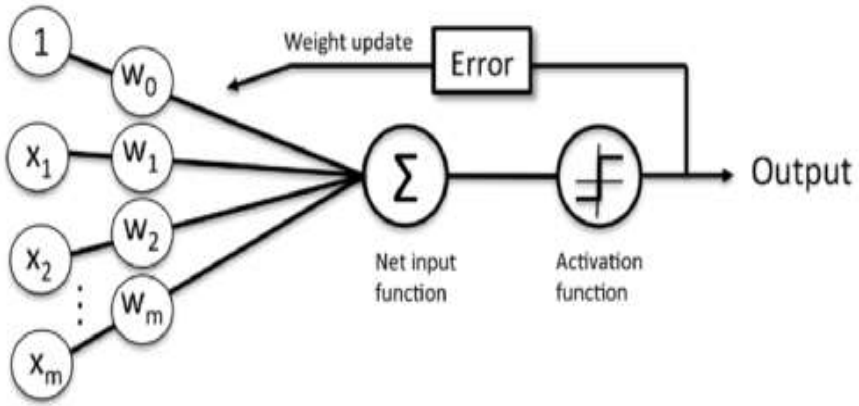
Differentiations range -x or 1

Continued…….



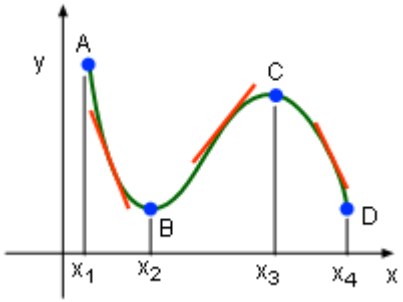
Additional info : https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html

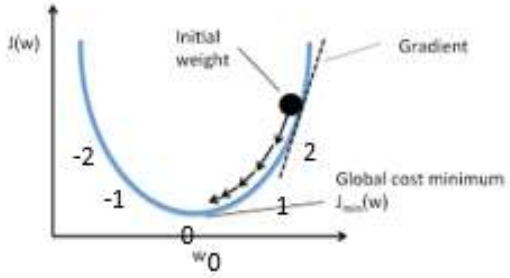
Weight update function



- Since we random initialize the weights we end up with some error.
- Now adjust your weights by differentiating your pervious output
- until you reach minimum error

When tangent is horizontal to x axis you may be at minima or maxima





Example : simple weight update

F(x) = x2 -3x + 2

df/dx = 0

df/dx = 2x – 3 + 0 = 0

x = 1.5 is this maxima or minima ?

put 1.5 in x2 – 3x + 2

f(1.5) = -0.25

f(1) = 0

Heart of gradient decent is chain rule in differentiation

Standard weight update functions

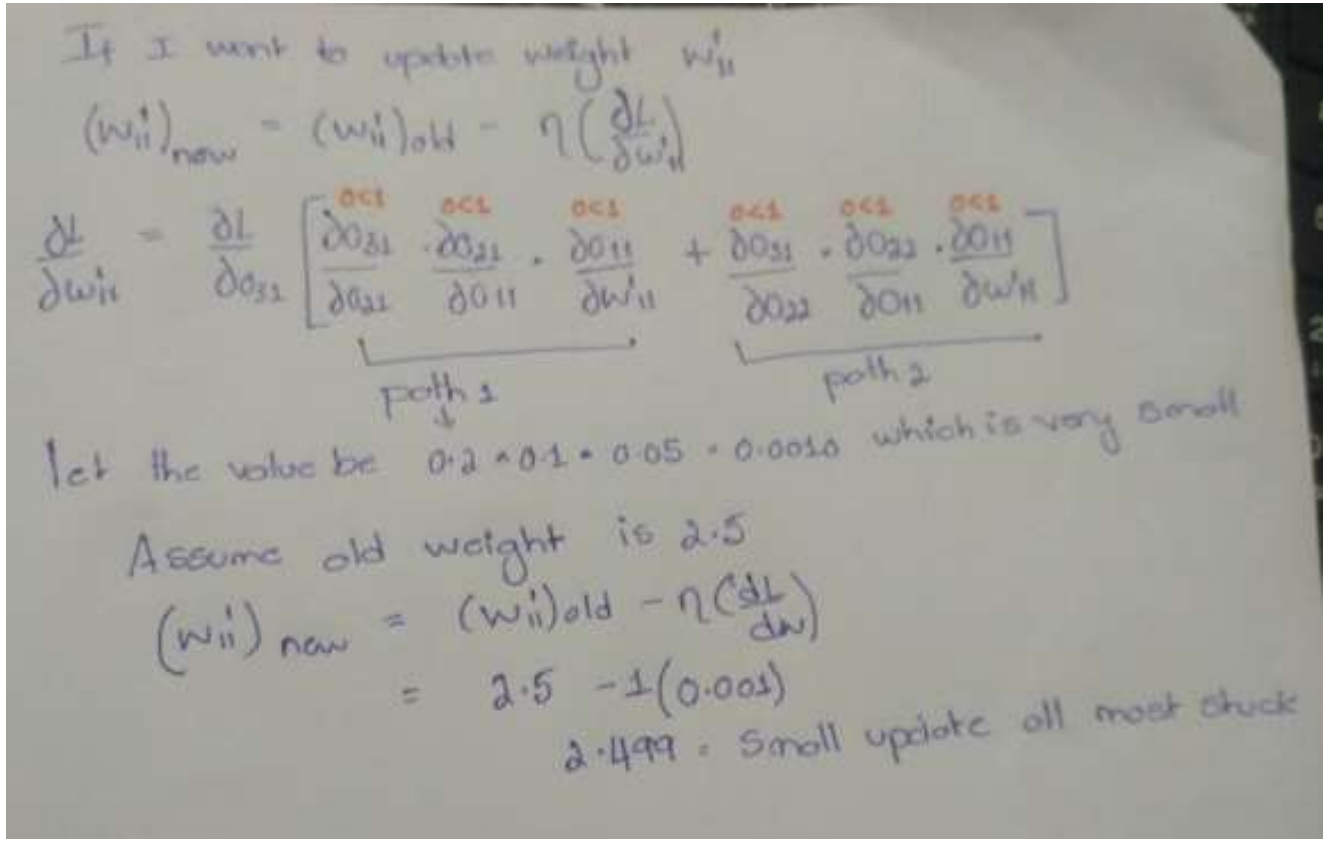
- SGD
- Min batch SGD
- SGD with momentum
- Ada grad
- Ada delta and RMS prop
- Adam [Adaptive momentum estimate]

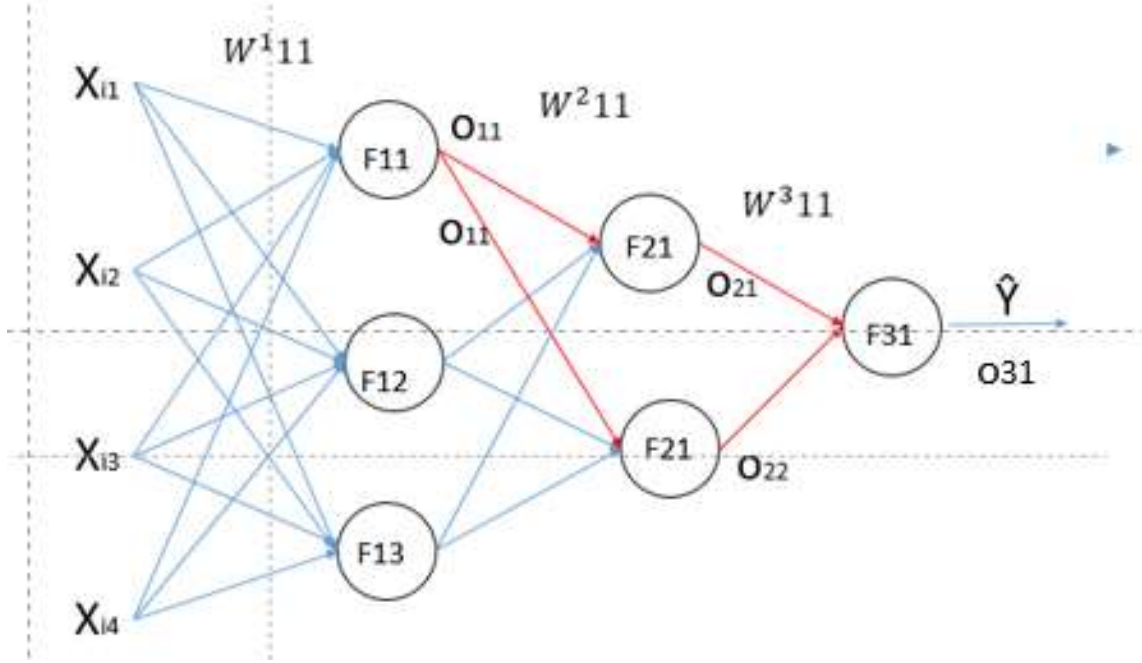
History of activation function

Dated from 1980 - 2012 people tried 2 - 3 layer neural networks

- Biggest problem faced is vanishing gradient [mathematical problem]
- Too little labelled data
- computation powers

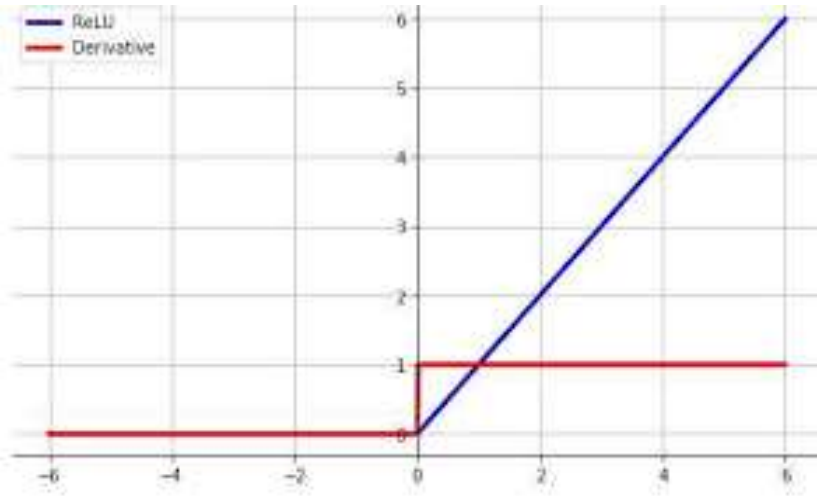
Vanishing Gradient in Sigmoid and Tanh functions





We can get stuck in **local minima** or at **saddle points**

How to avoid vanishing gradient ?
-- Rectified linear units ReLu



- Partial derivation of relu is either 0 or 1
- Since max value is 1 there is no problem of exploding gradient and again values are not in between 0 -1 there is no vanishing gradient
- Relu function converge faster then other function because value is either 1 or 0
- But it may lead into Dead activation

- If z is negative f(z) = 0
- df/dz = 0
- If one of the derivative is zero then complete = Dead activation
- Sigmoid and tanh may subjected to dead activation when you have very high negative values