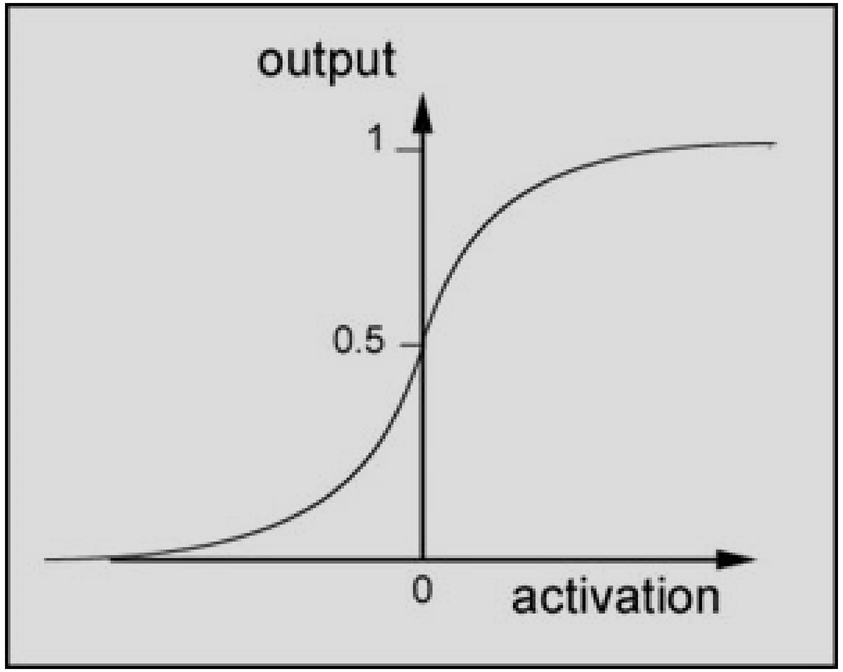
Activation Functions

Sigmoid Activation Function

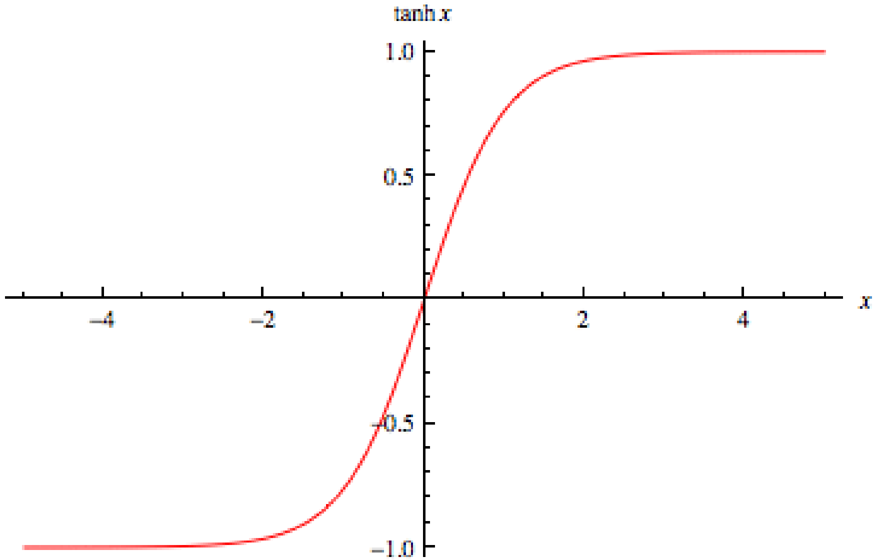
- $\hat{y}=\sigma\ (w^T x + b)$ where $\sigma(z)= \frac{1}{1+e^{-z}}$
- If z is very large then e^{-z} is close to zero and $\sigma(z) = \frac{1}{1+0} \approx 1$
- If z is very small then e^{-z} is large and $\sigma(z) = \frac{1}{1+Large\ Number} \approx 0$



Activation Functions: tanh

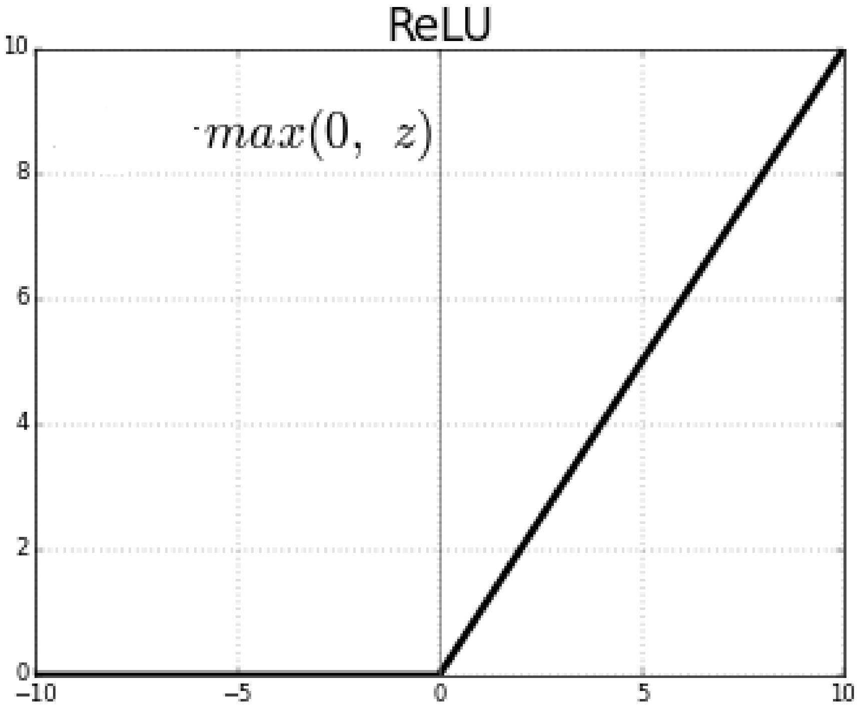
Tanh activation function is more preferred than the sigmoid function.
It is the shifted version of sigmoid function with a mean value of zero.
It has better centering effect for the activation function to be used on the hidden layer.
For binary classification problem at the output layer we use the Sigmoid function.

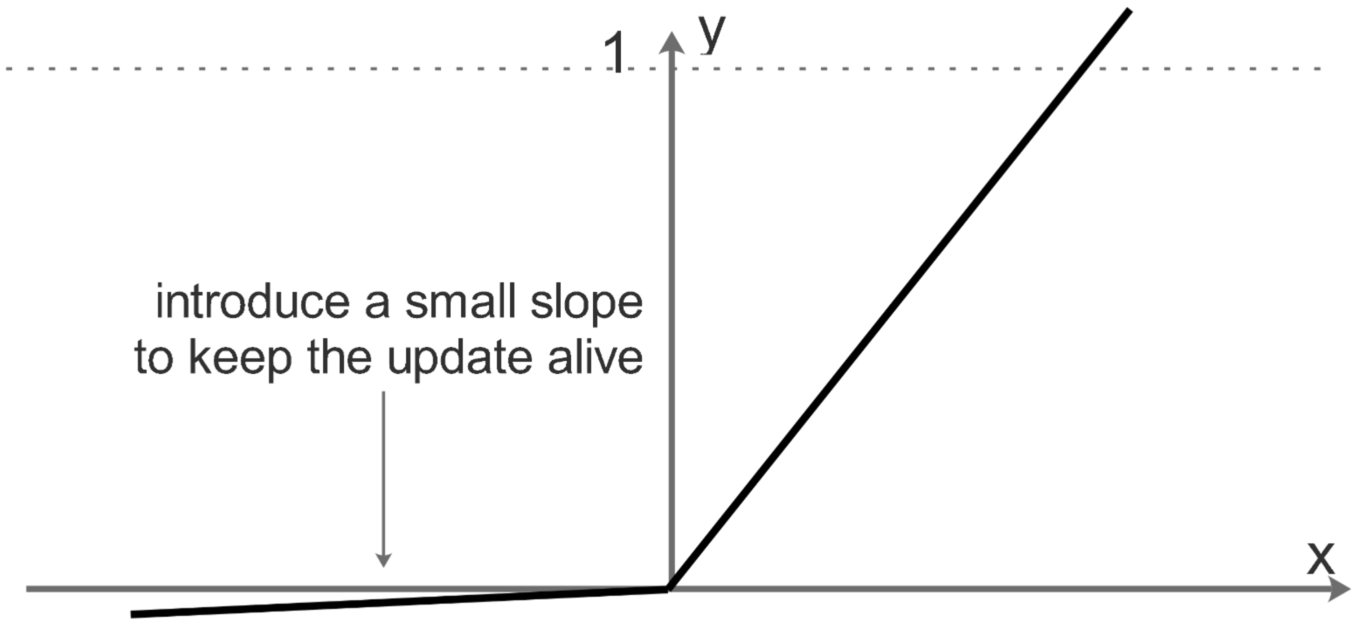
$$\tanh = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$



Problem in both Sigmoid and tanh is that the slope of the curve except the middle region is too small and goes close to zero. This create a serious issue with gradient descent and learning becomes very slow.

ReLU and Leaky ReLU





Leaky ReLU max(0.1z,z)

Rectified Linear Unit is the default activation function being used now. If you see the left part of the function, you can see that it is not zero, but almost zero. To resolve the issue of dead neurons people use Leaky ReLU

Softmax Function

When we have to classify in multiple categories then softmax function is useful. For example if you want to categorize pictures into

A) scene b) Animals c) Humans d) Vehicles then in that case we will have four outputs from the softmax function which will give us the probabilities of each of these categories.

Sum of the probabilities will be one and that with the highest probability will be shown as the answer.

Understanding softmax

$$z^{[L]} = \begin{pmatrix} 5 \\ 2 \\ -1 \\ 3 \end{pmatrix} \qquad t = \begin{pmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{pmatrix}$$

$$a^{[l]} = g^{[L]}(z^{[L]}) = \begin{pmatrix} e^5/(e^5+e^2+e^{-1}+e^3) \\ e^2/(e^5+e^2+e^{-1}+e^3) \\ e^{-1}/(e^5+e^2+e^{-1}+e^3) \\ e^3/(e^5+e^2+e^{-1}+e^3) \end{pmatrix} = \begin{pmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{pmatrix}$$

"Hard Max"

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Softmax regression generalizes logistic regression to C classes.
If c=2, softmax reduces to logistic regression.

Discussion on Activation Functions

Sigmoid vs Tanh

Sigmoid vs Relu

Sigmoid vs Softmax

What Activation Function to use

- Sigmoid functions and their combinations generally work better in the case of classifiers
- Sigmoids and tanh functions are sometimes avoided due to the vanishing gradient problem
- ReLU function is a general activation function and is used in most cases these days
- If we encounter a case of dead neurons in our networks the leaky ReLU function is the best choice
- Always keep in mind that ReLU function should only be used in the hidden layers
- As a rule of thumb, you can begin with using ReLU function and then move over to other activation functions in case ReLU doesn't provide with optimum results

https://www.analyticsvidhya.com/blog/2017/10/fundamentals-deep-learning-activation-functions-when-to-use-them/