Machine Learning WS 2018
Solution to Assignment 8
Submitted By : Vindhya Singh **Enrollment Number :** 03693296
Collaborated By : Wasiq Rumaney **Enrollment Number :** 03694978

Solution 1:

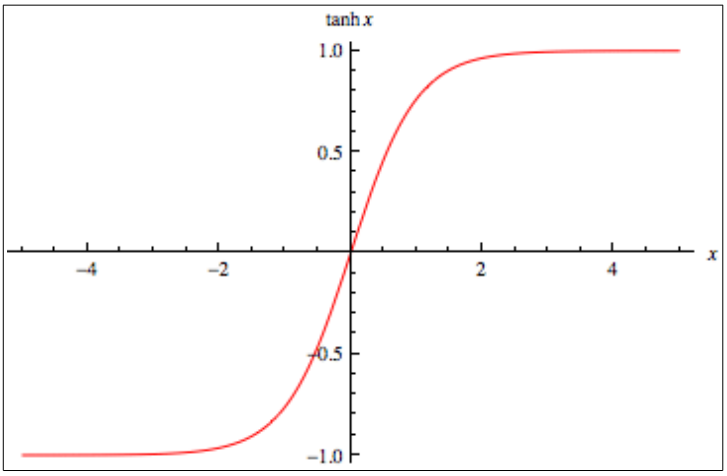
If we have a linear input, linear activation functions , the output is a linear function of the input variables. No matter how many hidden layers we use, it is only an output of a linear input. Thus, linear hidden layer is useless. We can keep adding linear hidden layers but it won't amount to much. Linear activation functions can be used as an output activation function, however.

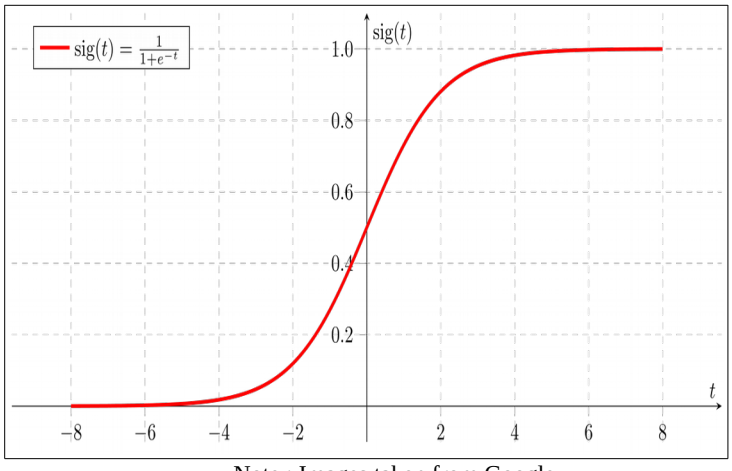
Solution 2:

We need to prove that "an equivalent network, which computes exactly the same function, but with hidden units using tanh(x) as activation functions." considering that a neural network where the hidden units use the sigmoid activation function is given.

Intuitively,

1) $\tanh(x)$ and $\sigma(x)$ are both S-shaped.





Note : Images taken from Google

2) $\tanh(x)$ and $\sigma(x)$, exists between -1 and 1, 0 and 1 respectively.

Mathematically,

We know that,

$$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \text{and} \quad \sigma(x)=\frac{1}{1+e^{-x}} \qquad \text{(a)}$$

Using, the formula for tanh(x),

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

$$= \quad \frac{2e^{x}}{e^{x}+e^{-x}}=\frac{2}{1+e^{-2x}}$$

$$= \quad 2(\sigma(2x)) \qquad \text{(Using the definition of sigmoid from (a))}$$

Therefore,

$$1+\tanh(x)=2(\sigma(2x))$$

$$= \quad \frac{1}{2}+\frac{\tanh(x)}{2}=\sigma(2x) \quad \Rightarrow \quad \tanh(x)=2\sigma(2x)-1$$

tanh near 0 is similar to the identity function, thus training a deep neural network resembles training a linear model as long as the activation of the network can be kept small. *[Goodfellow 6.3.2]*

Solution 3:

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

Using the Quotient Rule,

$$\frac{d(\tanh(x))}{dx}=\frac{(e^{x}+e^{-x}).(e^{x}+e^{-x})-(e^{x}-e^{-x}).(e^{x}-e^{-x})}{(e^{x}+e^{-x})^{2}}$$

$$= \frac{(e^{x}+e^{-x}).(e^{x}+e^{-x})}{(e^{x}+e^{-x})^{2}}-\frac{(e^{x}-e^{-x}).(e^{x}-e^{-x})}{(e^{x}+e^{-x})^{2}}$$

$$= 1-\tanh^{2}x$$

Hence, Proved that the the derivative of the tanh activation function can also be expressed in terms of the function value itself.

Like the sigmoid, the derivative of the tanh activation function can also be expressed in terms of the function value itself. It is a useful property because as output units they are compatible with the use of gradient-based learning.

Solution 4:

Given : $y=\log\sum_{i=1}^{N}e^{x_i}$

To Prove : The following identity holds,

$$y=\log\sum_{i=1}^{N}e^{x_i}=a+\log\sum_{i=1}^{N}e^{x_i-a}$$

Proof : Building on the Given information,

$$y = \log \sum_{i=1}^{N} e^{x_1}$$

$$= \quad e^{y} = \sum_{i=1}^{N} e^{x_i}$$

For getting the term *a* as in the To Prove section,

$$e^{y} . e^{-a} = \sum_{i=1}^{N} e^{x_i} . e^{-a}$$

$$= \quad e^{y-a} = \sum_{i=1}^{N} e^{x_i - a}$$

Taking log of the above equation,

$$y - a = \log \sum_{i=o}^{N} e^{x_i - a}$$

This implies that,

$$y = a + \log \sum_{i=1}^{N} e^{x_i - a}$$ which is the identity that we have to prove.

Hence, proved.

Solution 5:

Given : *Softmax function*

To Prove: $$\frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}} = \frac{e^{x_i - a}}{\sum_{i=1}^{N} e^{x_i - a}} \; where \; a = max_i \, x_i \qquad (1)$$

Proof:

Multiplying and dividing (1) by e^{-a} , we get,

$$\frac{e^{x_i} . e^{-a}}{e^{-a} . \sum_{i=1}^{N} e^{x_i}} \quad = \quad \frac{e^{x_i} . e^{-a}}{\sum_{i=1}^{N} e^{x_i} . e^{-a}} \quad = \quad \frac{e^{x_i - a}}{\sum_{i}^{N} e^{x_i - a}} \quad \text{where} \quad a = max_i \, x_i$$

Hence, Proved.

Solution 6:

Given : $-(ylog(\sigma(x))+(1-y)\log(1-\sigma(x)))$

To Prove: The following identity holds :

$$max(x,0)-x.y+\log(1+e^{-|(x)|})$$

Proof : Using the value of the sigmoid in the equation below

$-(ylog(\sigma(x))+(1-y)\log(1-\sigma(x)))$ where $\sigma(x)=\frac{1}{1+e^{-x}}$, we get

$$-(y\log(\frac{1}{1+e^{-x}})+(1-y)\log(1-\frac{1}{1+e^{-x}}))$$

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

$$= -y[\log(1)-\log(1+e^{-x})]-(1-y)[\log(e^{-x})-\log(1+e^{-x})]$$

$$= y\log(1+e^{-x})-\log(e^{-x})+\log(1+e^{-x})+y\log(e^{-x})-y\log(1+e^{-x})$$

$$= -\log(e^{-x})+y\log(e^{-x})+\log(1+e^{-x})$$

$$= x-x.y+\log(1+e^{-x})$$

Since the exponents is raised to the power of a negative number, therefore if the value of x < 0 then the chances of overflow will increase. Thus, it is better to use the maximum of (x,0)

$$x-x.y+\log(1+e^{-x})$$

$$= max(0,x)-x.y+\log(1+e^{|(x)|})$$ in order to prevent overflow.

Hence, it proves the required condition.