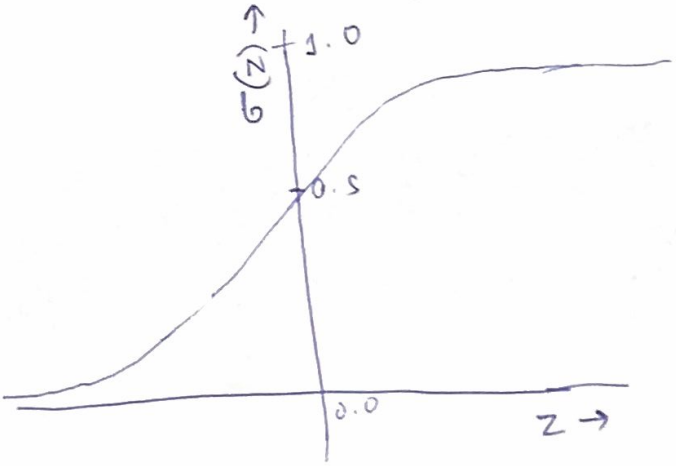
2. $\sigma(z) = \frac{1}{1+e^{-z}}$

We know that $z_i \in [0, 1000]$

$\Rightarrow \bar{z} = w\bar{x} + \bar{b}$

will have a higher magnitude since $\bar{z} \propto \bar{x}$

The graph of $\sigma(z)$ is as follows



$\therefore \forall z$ with high magnitude

$\sigma(z) \rightarrow 1$

$\Rightarrow \sigma'(z) = \sigma(z)(\sigma(z) - 1)$

approaches zero.

We know that $\delta^L = \nabla_a C_x \odot \sigma'(z)$

$\delta^l = (w^{l+1})^T \delta^{l-1} \odot \sigma'(z)$

$\Rightarrow$ Since $\sigma'(z)$ approaches zero it stagnates learning.

Another problem this poses is the overflow problem.

for high z, e^z is very large and we might not be able to handle it.

Using ReLu function in the hidden layer

ReLu being a linear function

does not stagnate learning.

Thus it definitely betters the problem.

However we see $f'(z)$ for ReLu

$$= \begin{cases} 1 & \forall\ z > 0 \\ 0 & \text{else} \end{cases}$$

$\Rightarrow$ ReLu will not be able to handle the big range of [0, 100] since gives the same derivative for the complete range.

Pre-processing required:

To overcome sigmoid problem:

① We can scale the data to a smaller range $\rightarrow$ like [0, 1]

To overcome ReLu problem:

② We can normalize the data about its mean to a range like [-1, 1]

3. In quadratic cost

$$C_x = \frac{1}{2}\|A_x^L - y\|^2$$

$$\Rightarrow \nabla_a C_x = \|A_x^L - y\|$$

So. according to BP1,

$$\delta^L = \nabla_a C_x \odot \sigma'(z)$$

$$= \|A_x^L - y\| \odot \sigma'(z)$$

In this expression, $\sigma'(z)$ slows down the learning as $\sigma'(z)$ near 0 as z approaches maxima/minima.

$\Rightarrow$ If we are able to remove this term, we will be able to fasten the learning process.

This is exactly what the cross entropy cost function does.

Cross-entropy cost:

$$C_x = -(y \ln a_x^L + (1-y)\ln(1-a_x^L))$$

$$\Rightarrow \nabla_a C_x = -\left[\frac{y}{a_x^L} + \frac{(1-y)}{(1-a_x^L)}\cdot(-1)\right]$$

$$= \frac{1-y}{1-a_x^L} \bar{\odot} \frac{y}{a_x^L} = \frac{a_x^L(1-y) - y(1-a_x^L)}{a_x^L(1-a_x^L)}$$

$$\Rightarrow \nabla_a C_a = \frac{a_x^L - y a_x^L - y + y a_x^L}{a_x^L (1 - a_x^L)}$$

$$= \frac{a_x^L - y}{\sigma(z)[1 - \sigma(z)]}$$

$$= \frac{a_x^L - y}{\sigma'(z)} \qquad \left[\because \sigma'(z) = \sigma(z)(1 - \sigma(z))\right]$$

$$\Rightarrow \delta^L = \nabla_a C_a \odot \sigma'(z)$$

$$= \frac{a_x^L - y}{\sigma'(z)} \cdot \sigma'(z)$$

$$= a_x^L - y$$

Hence, cross-entropy fastens the learning process