squared error cost function ≠ not good for classification

Training Set

	tumor size (cm) x_1	...	patient's age x_n	malignant? y
$i=1$	10		52	1
⋮	2		73	0
	5		55	0
	12		49	1
$i=m$	...		...	...

$$z = w_1 x_1 + \dots \, w_n x_n$$

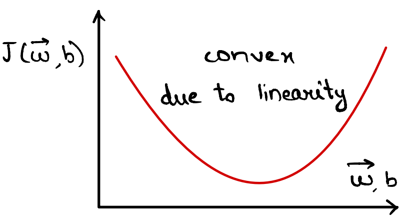
$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1+e^{-(\vec{w}\cdot\vec{x}+b)}}$$

In this case target $\hat{y}$ is 0 or 1

$$J(\vec{w},b) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - \hat{y}^{(i)}\right)^2$$

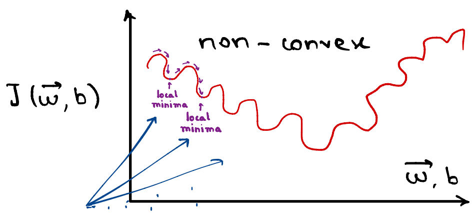
In linear regression

$$f_{\vec{w},b}(\vec{x}) = \vec{w}\cdot\vec{x} + b$$



In non-linear (logistics)

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1+e^{-(\vec{w}\cdot x+b)}}$$



multiple local minima, so the algorithm gets confused

$$\overbrace{\boxed{J(\vec{w}, b)}}^{\text{cost}} = \frac{1}{m} \sum_{i=1}^{m} \overbrace{\boxed{\frac{1}{2} \left(f_{w,b}(\vec{x}^{(i)}) - \hat{y}^{(i)} \right)^2}}^{\text{Loss}}$$

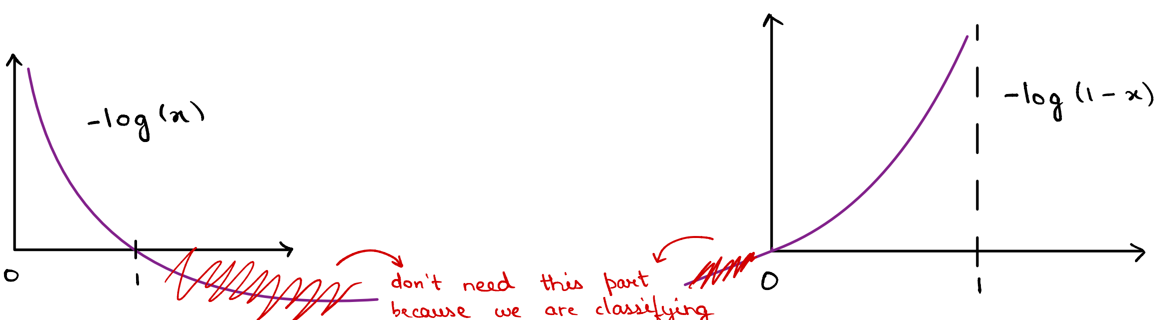
Loss :- measure of differences of a single example to that of its target value.

Cost :- measure of the losses over the whole training set.

Loss in a broad manner can be different for different examples.

$L(f_{w,b}(\vec{x}^{(i)}), y^{(i)})$ ← writing it as a function with $f_{\vec{w},b}(\vec{x}^{(i)})$ and $y^{(i)}$ as arguments

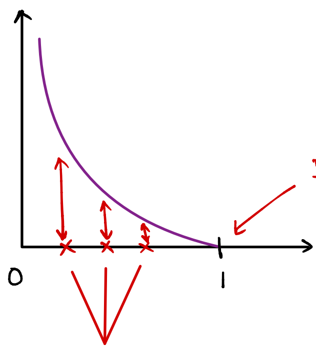
$$L(f_{w,b}(x^{(i)}), y^{(i)}) \begin{cases} -\log(f_{\vec{w},b}(\vec{x}^{(i)})), & \text{if } y^{(i)} = 1 \\ -\log(1 - f_{\vec{w},b}(\vec{x}^{(i)})), & \text{if } y^{(i)} = 0 \end{cases}$$



don't need this part because we are classifying only between 0 and 1.

Distance of the graph from the x-axis decides the error.

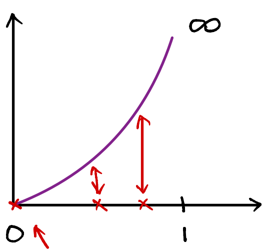
If $y^{(i)} = 1$:-



If point is 1, then error is least

The further the point from 1, the higher the error

Same for $y^{(i)} = 0$:-



perfect (loss = 0)

Two different functions for two different labels i.e. $y^{(i)} = 0$ and $y^{(i)} = 1$.

$$J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{L\left(f_{\vec{w},b}(\vec{x}), y^{(i)}\right)}_{\text{loss}}$$

$$\left.\begin{cases} -\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)})\right) & \text{if } y^{(i)} = 0 \end{cases}\right\}$$

plotting of $J(\vec{w})$ will be convex = easily finding $\vec{w}$ & b

Simplied version of loss function

$$\cancel{L\left(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)}\right) \begin{cases} -\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) \\ -\log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)})\right) \end{cases}}$$

$$L\left(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)}\right) = -y^{(i)} \log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) - (1 - y^{(i)}) \log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)})\right)$$

This is the same thing as the above conditional if $y^{(i)} = 1$:

$$L(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)}) = -1 \log(f_{\vec{w},b}(x^{(i)})) - \underbrace{(1-1)}_{=0} \log(1 - f_{\vec{w},b}(x^{(i)}))$$

$$= -\log(f_{\vec{w},b}(x^{(i)})$$

if $y^{(i)} = 0$:

$$L(f_{\vec{w},b}(x^{(i)}), y^{(i)}) = -0 \log(f_{\vec{w},b}(x^{(i)})) - (1-0) \log(1 - f_{\vec{w},b}(x^{(i)}))$$

$$= -\log(1 - f_{\vec{w},b}(x^{(i)})$$

$\therefore$ We can say cost function is :-

$$J(\vec{w},b) = \frac{1}{m} \sum_{i=1}^{m} -y^{(i)} \log(f_{\vec{w},b}(x^{(i)})) - (1-y^{(i)}) \log(1 - f_{\vec{w},b}(x^{(i)}))$$

This function is derived from the statistical principle " maximum likelihood principle "