9. Cost function for logistic regression.

I. Training set for our Model.

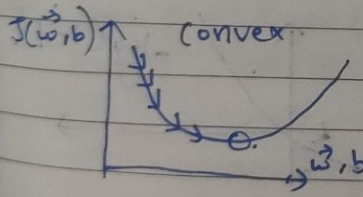
Let m = no. of training examples. $i = 1 \dots m$
n = no. of features $j = 1 \dots n$.
target y is 0 or 1.

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

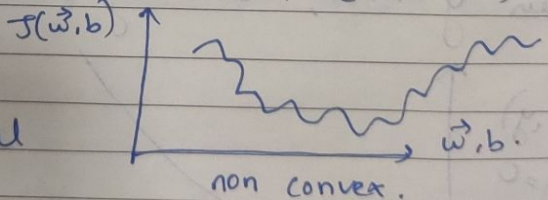
For linear regression, our Squarred Error cost function looks like this:

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

In case linear regression, $f_{\vec{w},b}(x) = \vec{w}\cdot\vec{x} + b$.



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



If we use gradient descent, it will have a lot of local minimas.
It won't guarantee global minima.

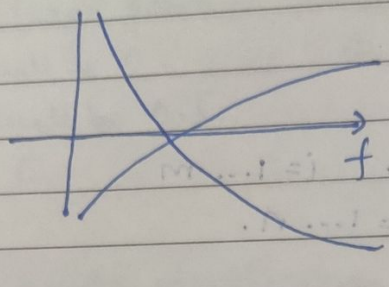
Let ~~loss~~ loss $L(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)})$. it is f^n of model $f(\vec{x})$ and y true label.

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

Logistic loss function.

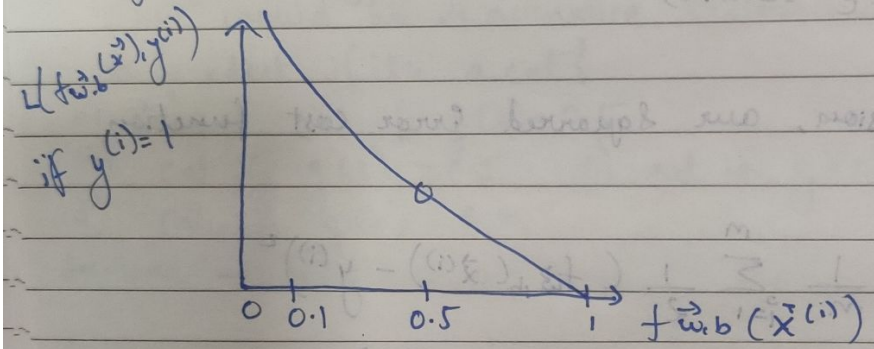
Why does this makes sense?

Now,

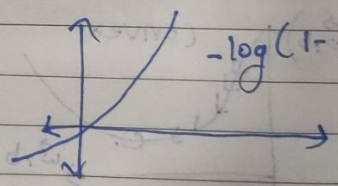


when $f=1$, $\log(f)=0$.

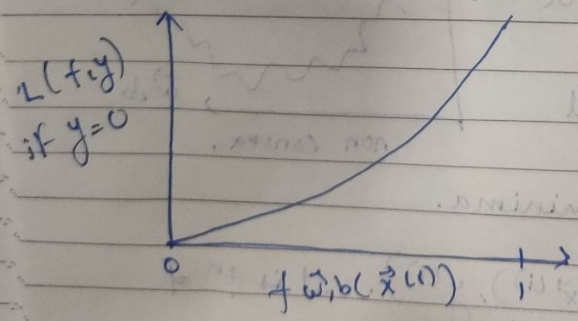
So, f is always between 0 and 1 for logistic regression.



As $f_{\vec{w},b}(x^{(i)}) \to 1$, then loss $\to 0$.
As $f_{\vec{w},b}(\vec{x}^{(i)}) \to 0$ then loss $\to \infty$



when $y^{(i)}=0$, our log function is



As $f_{\vec{w},b}(\vec{x}) \to 1$ then loss $\to \infty$
As $f_{\vec{w},b}(\vec{x}) \to 0$ then loss $\to 0$.

the further prediction $f_{\vec{w},b}(\vec{x})$ is from the target $y^{(i)}$, the higher the loss.

II. Cost.

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

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

Using this, our cost function can reach a global minimum.

:

Now, find w,b that minimize J.

Question: Why is the squared error cost not used in logistic regression?

a. The non-linear nature of the model results in a "wiggly", non convex cost function with many potential local minima.

b. The mean squared error is used for logistic regression.

Answer- a.

III. Simplified cost function.

As we know, our loss function looks like above.

Now,

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

This equation will be our simplified ~~cost~~ loss, since y can only take values of 1 and 0.

if $y^{(i)} = 1$

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

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

$$L(f(x), y) = -\log(1 - f(\vec{x})).$$

} loss values.

The simplified cost fn will be.

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

$$= -\frac{1}{m} \left[y^{(i)} \log(f_{\vec{w},b}(\vec{x}^{(i)})) + (1 - y^{(i)}) \log(1 - f_{\vec{w},b}(\vec{x}^{(i)})) \right]$$

This cost function is derived from Maximum Likelihood Parenthesis from Statistics.

This cost function above is Convex function.

Question: The above simplified cost function, if the target $y^{(i)} = 1$, the what does the expression simplify to?

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

Answer - b.