Logistic Regression: classification method.

* binary classification algorithm
* linear algorithm

prerequisit:

* what is ml?
* what is classification?
* what is training & testing dataset?
* what is x, y in training data
* what is sigmoid?
* Gradient decent?

I assume, you have some knowledge on this and starting

training data --> data used to train the model.

x --> is the features - N dimentional array

y --> label / output / target value - single dimention array. class (0 or) 1 when binary classification)

Sigmoid → will convert real value to probabilites. (0-1)

need to find the best value for the weights & bias which best fits the line.

Cost function → say how close your predicted value to the original target value

Common formula ⇒ $\frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$

$h_\theta(x^{(i)})$ → predicted via hypothesis

$y^{(i)}$ → original value

parameter estimation:

mi try to minimus the weight & bias

here: for samples labelled "1":

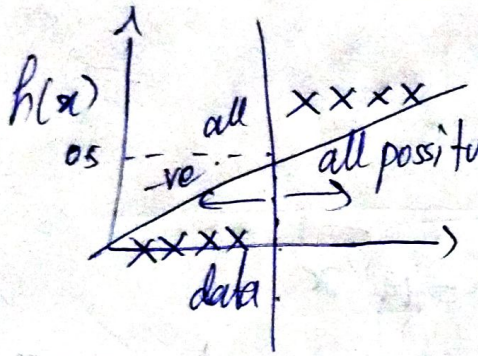
Estimate $(\hat{w}^T x) = \hat{\beta}$ such that $p(\hat{x})$ is as close to 1 as possible

for " " "0":

" " " $1-p(\hat{x})$ is as close to 1 as possible

(or)

$p(\hat{x})$ is as close to 0 as possible



* we need something to change the value b/w 0-1
* so we are using sigmoid fun to change the o/p to b/w 0-1
* then this is the probability

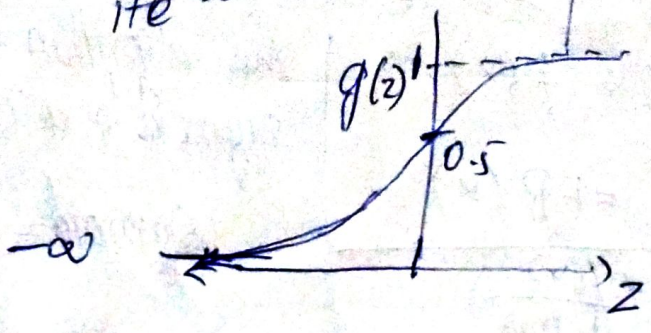
if ≥ 0.5 then 1

< 0.5 then 0.

hypothesis representation:

$h_\theta(x) = g(\theta^T x)$ $\quad g(z) = \frac{1}{1+e^{-z}}$ | sigmoid (or) logistic function

$= \frac{1}{1+e^{-(\theta^T x)}}$



so what is the meaning of it.... $h_\theta(x)$

let's say o/p 1 ⇒ ~~yes~~ cancer cell is malignant

0 = cancer cell is not malignent (dangerous)

something
if features $x = \begin{bmatrix} x_{11} & x_{12} \dots \\ x_{21} & x_{22} \dots \end{bmatrix}$

$h_\theta(x) = 0.7$ // let's say my hypothesis giving o/p as 0.7

then I can say that patient having 70% change to being malignant.

$h_\theta(x) = P(y=1 / x;\theta)$ = P (output / input)

↳ parameterized by. ↓ given

so as per statistics: $\boxed{P(1)+P(0)=1}$

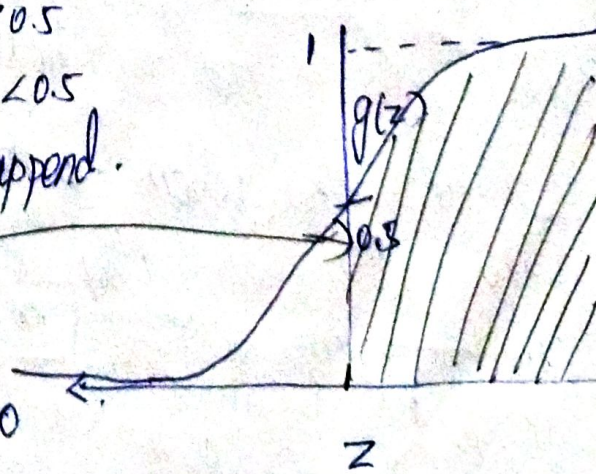
so $P(y=0/x;\theta) + P(y=1/x;\theta) = 1$

then $\boxed{P(y=0/x;\theta) = 1 - P(y=1/x;\theta)}$

this d info will help you later.

$y = \begin{cases} 1 & \text{if } h_\theta(x) \geq 0.5 \\ 0 & \text{if } h_\theta(x) < 0.5 \end{cases}$

so we can find when this will happend.



so, $g(z) \geq 0.5$ when $z \geq 0$

so, $h_\theta(x) = g(\theta^T x) \geq 0.5$

when

$\boxed{\theta^T x \geq 0}$

$$y = \begin{cases} 1 \text{ if } h_\theta(x) \geq 0.5 \Rightarrow \text{this will happen when } g(\theta^T x) \geq 0.5 \text{ or } \theta^T x \geq 0 \\ 0 \text{ if } h_\theta(x) < 0.5 \Rightarrow \text{this will happen when } g(\theta^T x) < 0.5 \text{ or } \theta^T x < 0 \end{cases}$$

simply $\Rightarrow$ $\theta^T x \geq 0$ o/p = 1
$\theta^T x < 0$ o/p = 0.

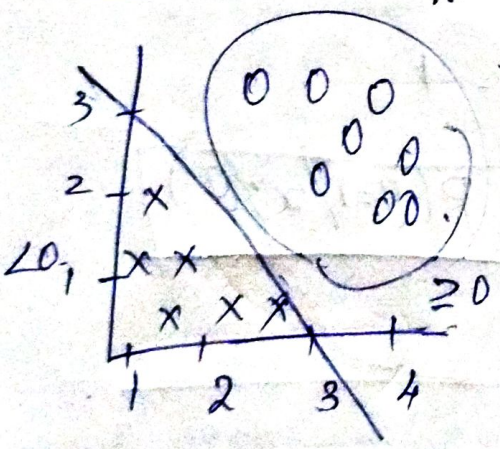
So what i need to find here:

I need to find the value of θ => which correctly classifies things.

(or) in other hand

θ which minimizes the cost function

so the start form linear example.



here $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

find the value for θ_1 which correctly classifies this.

so let take $\theta_0 = -3$ $\theta_1 = 1$ $\theta_2 = 1$

$h_\theta(x) = g(-3 + x_1 + x_2)$

so when $-3 + x_1 + x_2 \geq 0$ o/p is 1

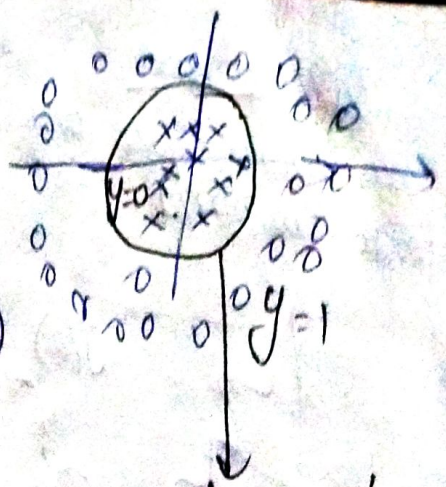
other hand $x_1 + x_2 \geq 3$. o/p is 1

lets plot this

Non linear example:

need to add non-linearity to the hypothesis so.

$h(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$



lets say $\theta_0 = -1$ $\theta_1 = 0$ $\theta_2 = 0$
$\theta_3 = 1$ $\theta_4 = 1$.

$-1 + x_1^2 + x_2^2 \geq 0$ then $y = 1$

$x_1^2 + x_2^2 \geq 1$ → equation of $x_1^2 + x_2^2 = 1$ circle with radius "1" centered to zero

---

Cost function: → used to fit the parameters (weights, θ)

Training set: $\{(x^1, y^1), (x^2, y^2), \ldots (x^m, y^m)\}$

m # training data $x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_m \end{bmatrix}$ $x_0 = 1$ $y \in \{0, 1\}$

$h_\theta(x) = 1/(1+e^{-\theta^T x})$ find θ?

Cost function:

linear regression = $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\frac{1}{2}(h_\theta(x^i) - y^i)^2}_{\text{cost}(h_\theta(x^i), y^i)}$

so $\text{cost}(h_\theta(x^i), y^i) = \frac{1}{2}(h_\theta(x^i) - y^i)^2$

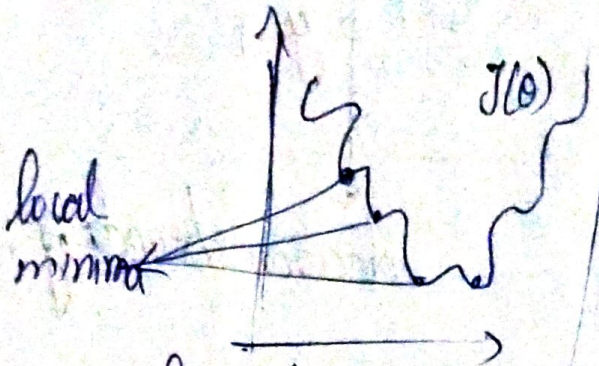
diff b/w predicted from original value.

but this $J(\theta)$ ~~will~~ won't work for logistic regression.

because log. reg is a **non-linear function** so the cost fun may ~~contain~~ come **non-convex** → we need **convex cost funct**

non-convex

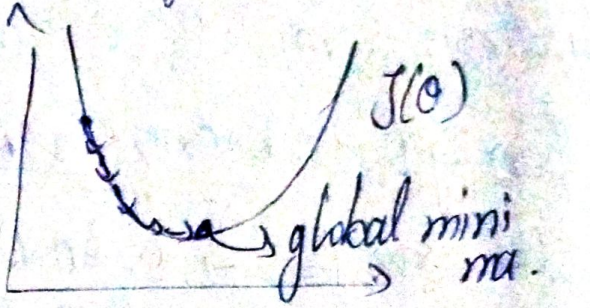
so it has more than one local minima



so, it has always converge to global minima.

convex

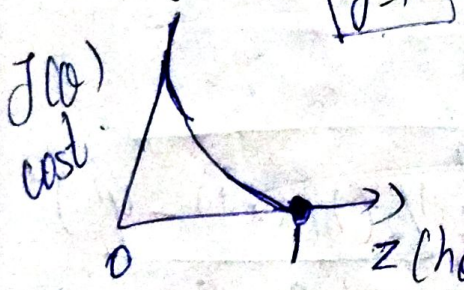
one local minima (global minima)



so we are using some complex function here.

$$\text{cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

$-\log(h_\theta(x))$ $\boxed{y=1}$

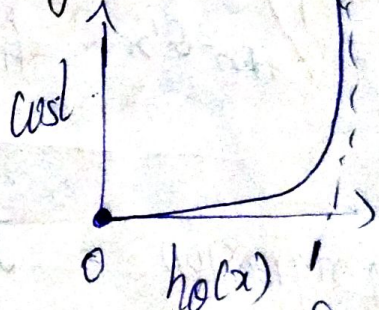


so if $h_\theta(x) = 1$ then cost is 0

$h_\theta(x) \to 0$ then cost ↑.

we are finding!

$-\log(1-h_\theta(x))$ $\boxed{y=0}$



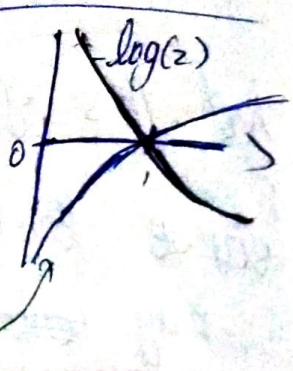
so if $h_\theta(x) = 0$ then cost is 0

$h_\theta(x) > 0$ then cost ↑ —

why $-\log(z)$ $z = h_\theta(x) = \theta^T x$

because we need a convex function $\log(z) =$

we need value $0 \le z \le 1$ so we choose $-\log(z)$



So finally coming to cost function for logi reg & Gradient decent

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^i), y^i)$$

$$\text{Cost}(h_\theta(x^i), y^i) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases} \quad \text{--- (1)}$$

we can write it in a single line:

Note $y = 0$ (or) 1 always.

$$\text{Cost}(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)(1-h_\theta(x)) \quad \text{--- (2)}$$

(1-1) 0

(1) = (2) => let's prove.

if $y=1$

$\text{cost}(h_\theta(x), y) = -1\log(h_\theta(x))$ - (same as first one)

if $y=0$

$\text{cost}(h_\theta(x), y) = -0 \quad -\log(1-h_\theta(x))$ same as second

(1) == (2)

$\Rightarrow$ (1)

So => $\boxed{J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))}$

why this cost fun?

* this has derived from statistics principle

Maximum likelihood estimation.

which is always used to find the best parameter.

* and it has the "convex" property.

To fit θ:

$\min_{\theta} J(\theta)$ and find θ feature scaling help to train faster for log reg

To make a prediction

o/p $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ with found θ last step.

using Gradient decent:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$$J(\theta) = y^i \log(h_\theta(x^i)) + (1-y^i) \log(1-h_\theta(x^i))$$

$\frac{\partial}{\partial \theta_j} J(\theta)$ = let's forgot i and do for one data then add i

$y^i \log(h_\theta(x)) \Rightarrow y \log\left(\frac{1}{1+e^{-\theta^T x}}\right) \Rightarrow$

to derivate this use chain rule

$\frac{d}{dx}\log(x) = 1/x$ ↳ $f(g(x))' = f'(g(x)) \cdot g'(x)$

$$\Rightarrow y\left(\frac{1}{1+e^{-\theta^T x}}\right) \frac{\partial}{\partial \theta_j}\left(\frac{1}{1+e^{-\theta^T x}}\right)$$

$1/g(\theta^T x) \Rightarrow g(z) = \frac{1}{1+e^{-z}}$

$$\Rightarrow y^i \left(\frac{1}{g(\theta^T x)}\right) \frac{\partial}{\partial \theta_j}(g(\theta^T x)) + (1-y^i) \frac{1}{1-g(\theta^T x)} \frac{\partial}{\partial \theta_j}(1-g(\theta^T x))$$

$$\Rightarrow \frac{\partial}{\partial \theta_j}\left[g(\theta^T x)\right]\left[y\left(\frac{1}{g(\theta^T x)}\right) - (1-y)\left(\frac{1}{1-g(\theta^T x)}\right)\right]$$

$\hookrightarrow x_j \cdot g(\theta^T x)(1-g(\theta^T x))$

$\Rightarrow x_j \left[y \left(\frac{1}{g(\theta^T x)} \right) - (1-y) \left(\frac{1}{1-g(\theta^T x)} \right) \right]$

- $g(\theta^T x)(1-g(\theta^T x)) x_j$ again chain rule

derivation for $\frac{\partial}{\partial \theta_j} g(\theta^T x)$: $\boxed{\frac{\partial}{\partial \theta_j} g(z) \cdot \frac{\partial}{\partial \theta_j} \theta^T x}$

$\frac{\partial}{\partial \theta_j} g(\theta^T x) \Rightarrow \frac{\partial}{\partial z} g(z) = \frac{d}{dz} \left(\frac{1}{1+e^{-z}} \right)$

$\Rightarrow \frac{1}{(1+e^{-z})^2} (e^{-z})$

$= \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{(1+e^{-z})} \right)$

$= g(z) \cdot (1 - g(z))$ //

so $\frac{\partial}{\partial \theta_j} [g(\theta^T x)) = \frac{\partial}{\partial \theta_j} [g(\theta^T x)) \cdot \frac{\partial}{\partial \theta_j} (\theta^T x)$

$\Rightarrow g(\theta^T x)(1 - g(\theta^T x) \cdot x_j$

---

so:

$\Rightarrow g(\theta^T x)(1-g(\theta^T x)) x_j \left[y \left(\frac{1}{g(\theta^T x)} \right) - (1-y) \left(\frac{1}{1-g(\theta^T x)} \right) \right]$

$x_j \left[y \cdot g(\theta^T x)(1-g(\theta^T x)) \cdot \frac{1}{g(\theta^T x)} - (1-y) \cdot g(\theta^T x)(1-g(\theta^T x)) \cdot \frac{1}{1-g(\theta^T x)} \right]$

$\Rightarrow x_j \left[y(1-g(\theta^T x)) - (1-y)(g(\theta^T x)) \right]$

$\Rightarrow x_j \left[y - yg(\theta^T x) - [g(\theta^T x) - yg(\theta^T x)] \right]$

$\Rightarrow x_j [y - g(\theta^T x)]$

$-1/m$

$\Rightarrow$ [redacted] $\left[x_j \left(y - g(\theta^T x) \right) \right]$

comes from $J(\theta) = -1/m \left[\sum_{i=1}^{m} \cdots \right]$ — 2 pages above ①

$\Rightarrow$ so $x_j \left(g(\theta^T x) - y \right)$

add for multiple values (add i)

$\Rightarrow -1/m \, x_j^{(i)} \left(g(\theta^T x^i) - y^i \right)$

$\Rightarrow \boxed{1/m \, x^i_j \left(h_\theta(x^i) - y^i \right).}$

So $\boxed{\theta_j = \theta_j - \alpha 1/m \sum_{i=1}^{m} (h_\theta(x^i) - y^i) \, x^i_j}$

for bias $\boxed{\text{bias} = \text{bias} - \alpha 1/m \sum_{i=1}^{m} (h_\theta(x^i) - y^i)}$