Supervised Learning

$$x^{(i)} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \begin{array}{l} \leftarrow 1 \\ \leftarrow \text{sq. ft} \\ \leftarrow \text{\# Bed room} \end{array}$$

Training Data

$D = \{ (x^{(i)}, \underbrace{y^{(i)}}_{\text{Target}}),\ i = 1, 2, \dots n \}$

↓

Learning algo → Cost fun + optimizer

$J(\theta)$

↓

$x \rightarrow h_\theta \rightarrow y$

y → Cont: $y \in \mathbb{R}$ ✓ Regression

y → Dis: $y = \{0, 1\}$ ✓ Classification

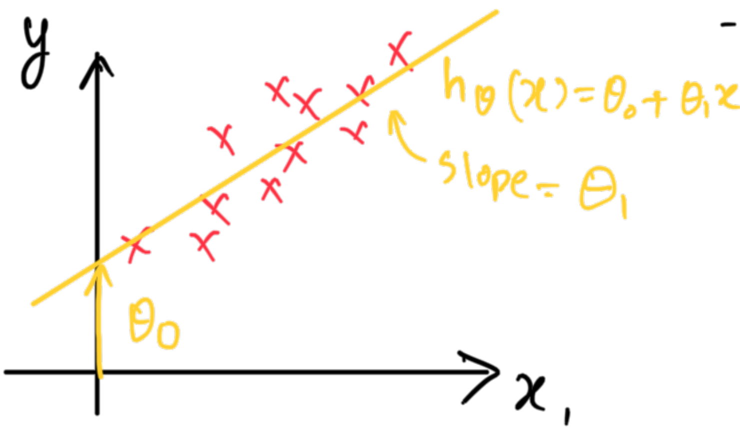
$y = h_\theta(x)$

parameters (θ)

model/hypothesis (h)

Linear Regression → Solving regression problem

✓ $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots \theta_d x_d$



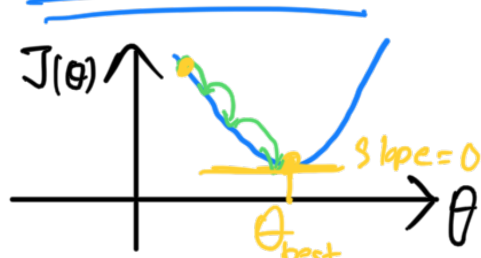
$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \quad x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$ ← bias $\equiv 1$ $\Rightarrow h_\theta(x) = \theta^T x$

Cost function: $J_\theta(x) = \overbrace{\frac{1}{n} \sum_{i=1}^{n}}^{\text{Total}} \overbrace{J_i(\theta)}^{\text{sample cost}}$

$$J_i(\theta) = \left[h_\theta(x^i) - y^i \right]^2$$

$$= \left[\theta^T x^i - y^i \right]^2$$

MSE Loss



Optimizer : Gradient Descent

θ^0 ← iteration no

$\theta^0 = \text{random}$

while not converged:

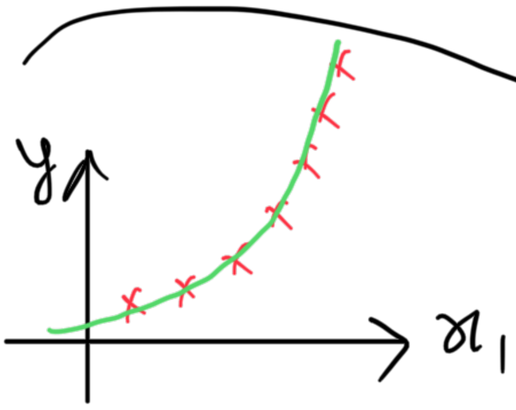
$$\theta^{k+1} = \theta^k - \underline{\alpha} \nabla_\theta J$$

→ gradient of J w.r.t θ

step size

if $\nabla_\theta J \approx 0$:

converged = True.



Feature Engineering

x_1	x_1^2	y
3	9	2
4	16	4
5	25	16

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \qquad \phi(x) = \begin{bmatrix} x_0 \\ x_1 \\ x_1^2 \end{bmatrix} = \underline{\underline{x'}}$$

$$\boxed{h_\theta(x) = \theta^T x'} = \theta_0 + \theta_1 x_1 + \theta_2 \underbrace{x_1^2}_{x_2}$$

$$J(\theta) = \frac{1}{n} \sum_{i=0}^{n} \left[\theta^T x^{(i)} - y^{(i)}\right]^2$$

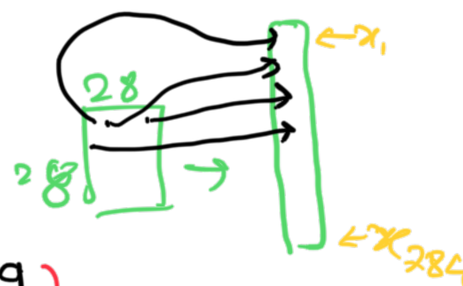
$$= \frac{1}{n} \sum_{j=0}^{n} \left[\theta_0 + \theta_1 x_1 + \theta_2 \underbrace{x_1^2}_{x_2} - y^{(i)}\right]^2$$

Classification Problem

y is discrete.

eg. Binary $\rightarrow y = \{0, 1\}$

x_1 (size)	x_2 (density)	x_3 (Age)	y (benign/malignant)
3	0.2	24	0
6	-0.5	50	1



Example

image pixels $\longrightarrow$ cat / dog

$x = [x_0, x_1, \dots x_d]$, $x_0 = 1$

$y = \{0, 1\}$

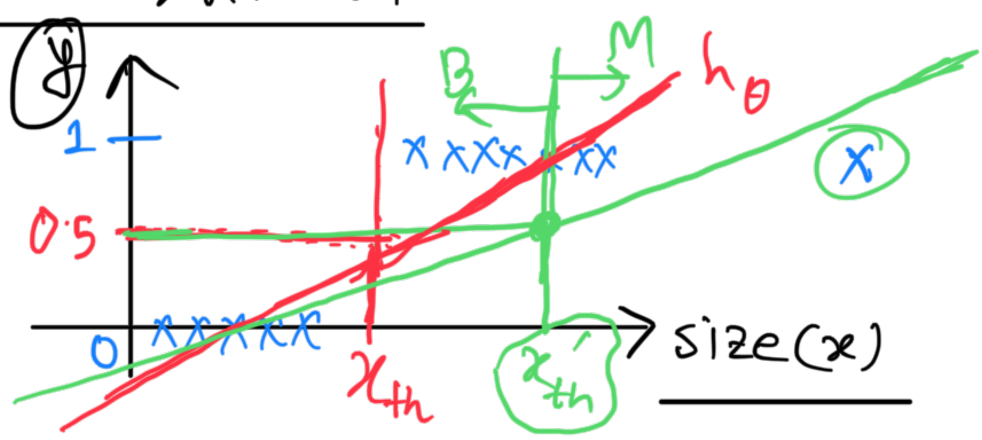
tumor size $\longrightarrow$ benign / malignant

$x = [x_0, x_1]$, $x_0 = 1$

$y = \{0, 1\}$

Email text $\longrightarrow$ spam / not spam

Tumor classification



$$h_\theta(x) = \theta^T x = \theta_0 + \theta_1 x$$

$x \geq x_{th} \equiv h_\theta(x) \geq 0.5 \Rightarrow$ malignant

$x \leq x_{th} \equiv \quad < 0.5 \Rightarrow$ benign

Classification problem $\Rightarrow$ Linear regression

① Sensitive to outliers

① $y = 0$ or 1

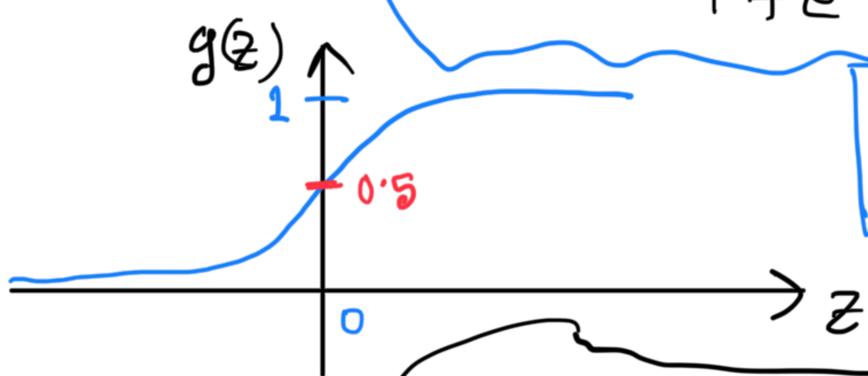
$h_\theta(x) = \theta^T x$ < 0 and > 1

Logistic Regression

Idea: Transform $\theta^T x$ so that $0 \leq h_\theta(x) \leq 1$

$$\boxed{h_\theta(x) = g(\underbrace{\theta^T x}_{z})}$$ model

for L.R $\rightarrow$ $g(z) = \dfrac{1}{1+e^{-z}}$



Sigmoid function
Logistic function

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

$$y = \begin{cases} 1 & \text{if } \dfrac{1}{1+e^{-\theta^T x}} \geq 0.5 \\ 0 & \text{if } \dfrac{1}{1+e^{-\theta^T x}} < 0.5 \end{cases}$$

interpret: $h_\theta(x) = P(y=1 \mid x, \theta)$

probability that $y=1$

eg

$$x = \begin{bmatrix} \overset{=1}{x_0} \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 0.54 \\ 0.1 \end{bmatrix}$$

$$\theta^T x = \theta_0 x_0 + \theta_1 x_1 = 0.54 \times 1 + 0.1 \times 3$$
$$= 0.84$$

$$g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} = \frac{1}{1+e^{-0.84}} = 0.7$$

$$h_\theta(x) = g(\theta^T x) = 0.2$$

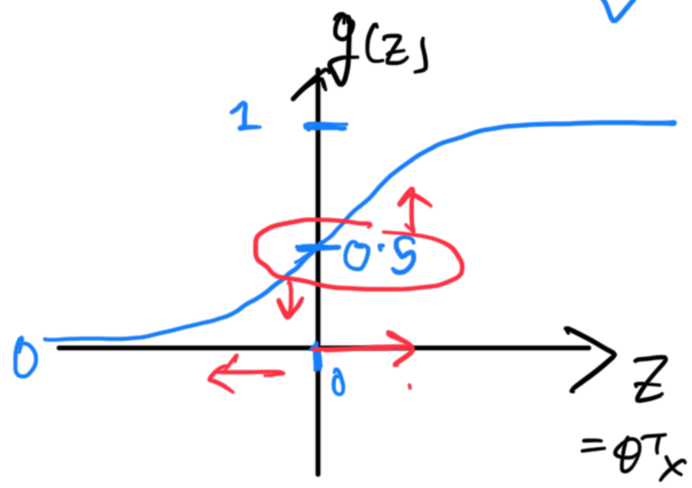
$P(y=1 \mid x, \theta)$

↳ 20% probability that the $y=1$ (tumor is malignant)

Decision Boundary

$$\boxed{h_\theta(x)} = g(\theta^T x)$$

sigmoid function ← $$= \frac{1}{1+e^{-\theta^T x}}$$



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

↳ when? $\Rightarrow z \geq 0$

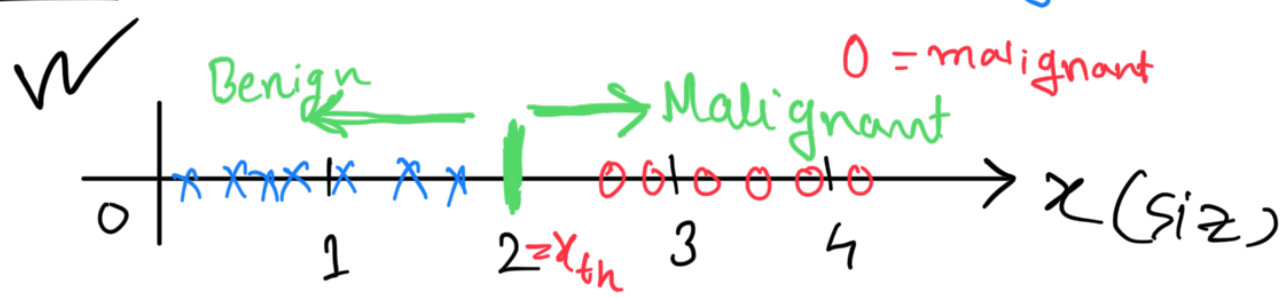
if $z < 0$

$\Rightarrow \theta^T x \geq 0$

eg. 1

X = benign

O = malignant



$$h_\theta(x) = g(\theta^T x) = g(\theta_0 + \theta_1 x)$$

$\rightarrow \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$

$\rightarrow x = \begin{bmatrix} \overset{=1}{x_0} \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ x \end{bmatrix}$

$\theta^T x = \theta_0 + \theta_1 x = -2 + x$

∴ Decision boundary:

$\theta^T x \geq 0$

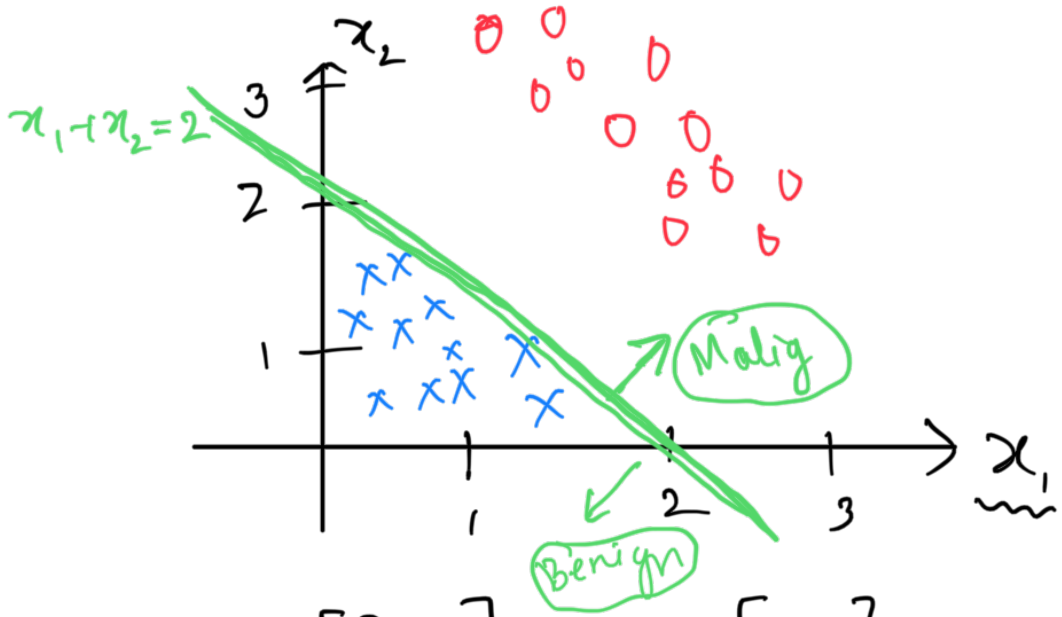
$\Rightarrow -2 + x \geq 0$

$\Rightarrow \boxed{x \geq 2}$

eg-2

X = benign

O = malignant



$$x = \begin{bmatrix} x_0 = 1 \\ x_1 \\ x_2 \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \Rightarrow \text{after optimization}$$

x_1 ← size

x_2 ← density

$$\theta_{best} = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

∴ Decision boundary ⇒ $\theta^T x \geq 0$

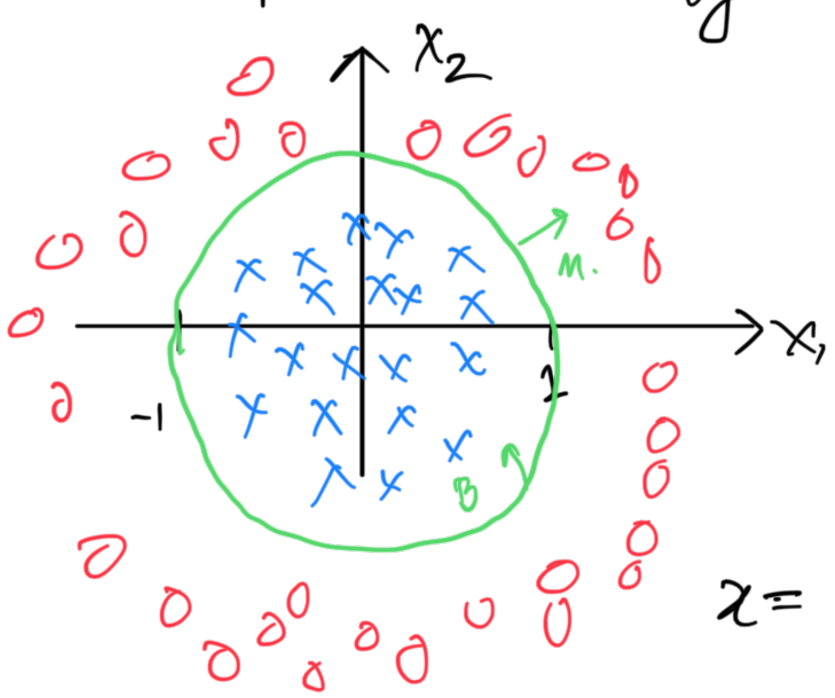
$\Rightarrow \theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0$

$\Rightarrow -2 + x_1 + x_2 \geq 0$

$\Rightarrow \boxed{x_1 + x_2 \geq 2}$

$x_1 + x_2 = 2$

eg-3 Complex boundary



x_1	x_2	x_1^2 (x_3)	x_2^2 (x_4)	y

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

x_1^2 ← x_3

x_2^2 ← x_4

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}$$

$$\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2$$

$$\theta_{best} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

$\therefore$ Decision boundary

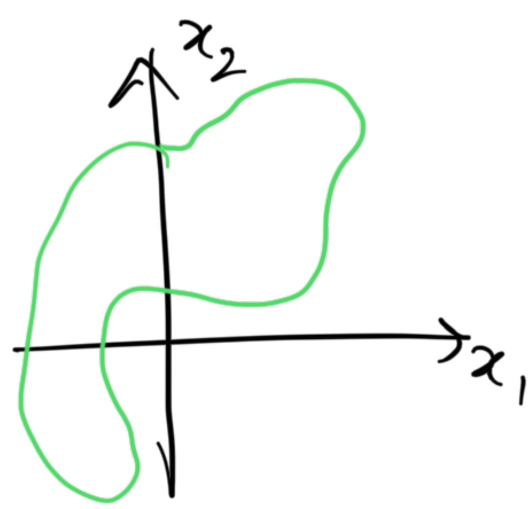
$$\Rightarrow -1 + 0 + 0 + x_1^2 + x_2^2 \geq 0$$

$$\Rightarrow x_1^2 + x_2^2 \geq 1$$

$$\boxed{x_1^2 + x_2^2 = 1}$$

$$h_\theta(x) = g\big(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1^3 x_2 + \theta_6 x_1^2 x_2^2 + \theta_7 \sin(x_1)\big)$$

$$x = \begin{array}{c} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1^3 x_2 \\ x_1^2 x_2^2 \\ \sin(x_1) \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \leftarrow x_3 \\ \leftarrow x_4 \\ \leftarrow x_5 \\ \leftarrow x_6 \\ \leftarrow x_7 \end{array} \Rightarrow \theta = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$



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

Fitting the Model ≡ Finding Best θ [Learning Algo]

$y = \{0, 1\}$

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots\ldots, (x^{(n)}, y^{(n)})\}$$

n-sample

$\rightarrow$ $x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$ $x_0 \equiv 1$ $\quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$

$g(z) = $ sigmoid

$\rightarrow$ $h_\theta(x) = g(\theta^T x) = \dfrac{1}{1 + e^{-\theta^T x}}$

Task: Find the "best" θ

$\rightarrow$ θ that performs best of D

$\Rightarrow$ minimizes some cost function $J(\theta)$

Cost function

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} J_i(\theta)$$

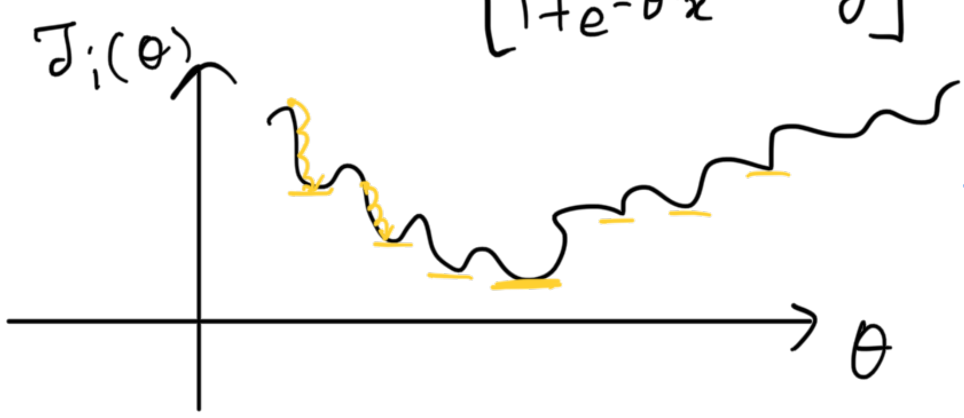
Cost for the i'th training sample

MSE

Opt 1: $J_i(\theta) = [\underbrace{h_\theta(x)}_{0 \text{ to } 1} - \underbrace{y}_{0 \text{ or } 1}]^2$

$x \equiv x^{(i)}$
$y \equiv y^{(i)}$

$$= \left[\frac{1}{1 + e^{-\theta^T x}} - y\right]^2$$



non-convex

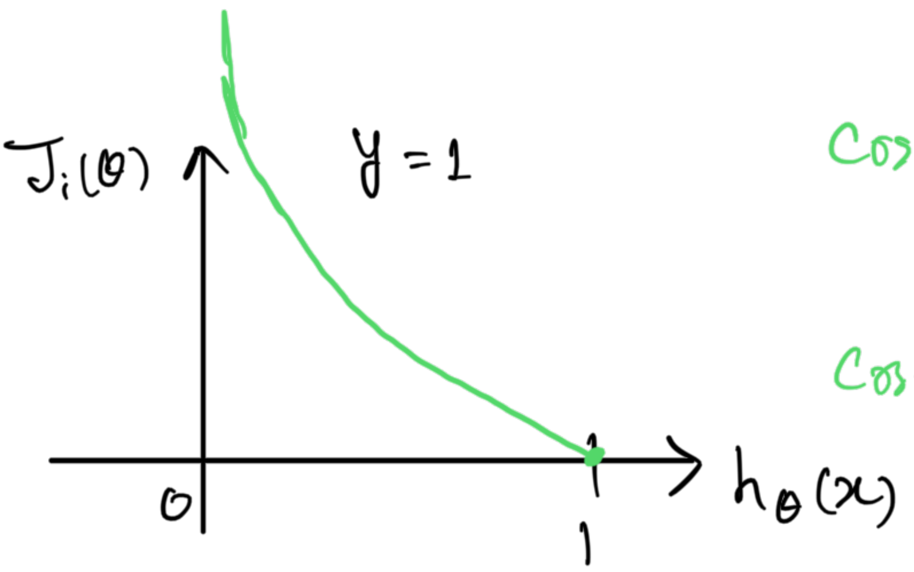
Opt-2: Binary Cross-Entropy Loss

log base = e

Cost function for ith sample

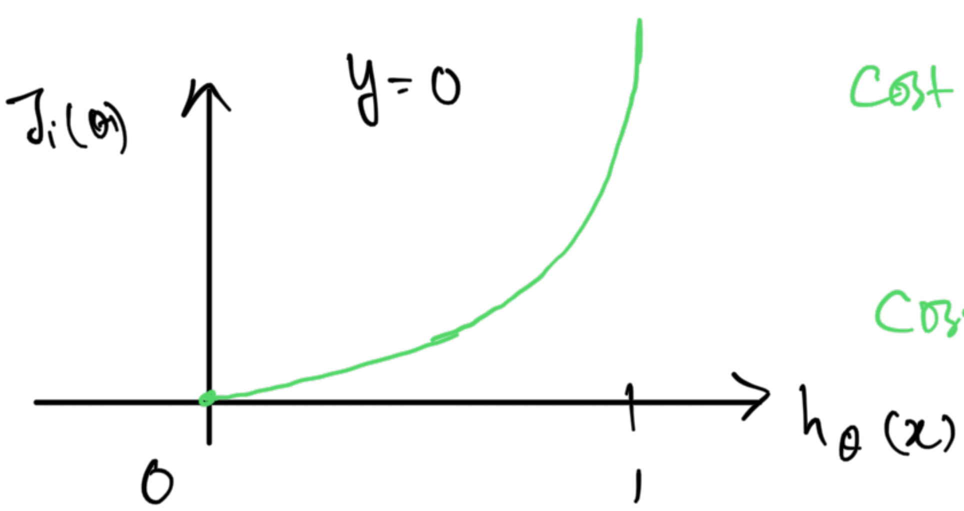
$$J_i(\theta) = \begin{cases} -\log(\overbrace{h_\theta(x)}^{\text{output of model}}) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\Rightarrow J_i(\theta) = -\Big[\underbrace{y \times \log(h_\theta(x))}_{\text{zero if } y=0} + \underbrace{(1-y)\log(1-h_\theta(x))}_{\text{zero if } y=1}\Big]$$



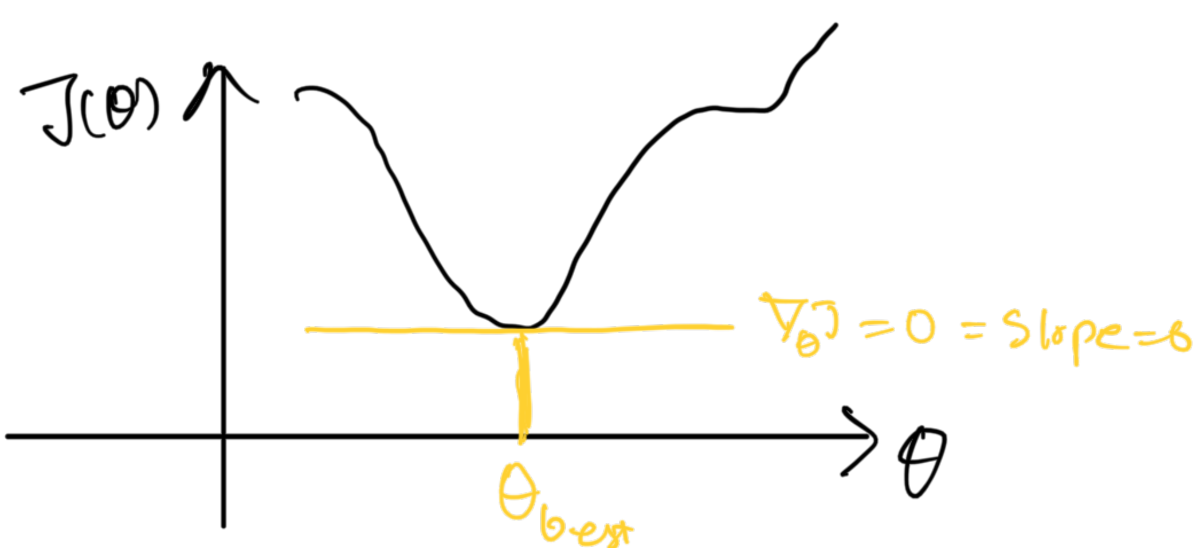
Cost = 0 if $y=1$ and $h_\theta(x) = 1$

Cost = ∞ if $y=1$ and $h_\theta(x) = 0$



Cost = 0 if $y=0$ and $h_\theta(x) = 0$

Cost = ∞ if $y=0$ and $h_\theta(x) = 1$



Minimizing the Cost

Optimizer

① Gradient descent

$\theta^0 = \text{random}$

while not converged:

$\theta^{k+1} = \theta^k - \alpha \nabla_\theta J$

if $\nabla_\theta J \approx 0$:

converged = True.

② Adam

③ Adaboost

④ LGBFS

complex but faster convergence

Logistic Regression Summary

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots\ldots, (x^{(n)}, y^{(n)})\}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

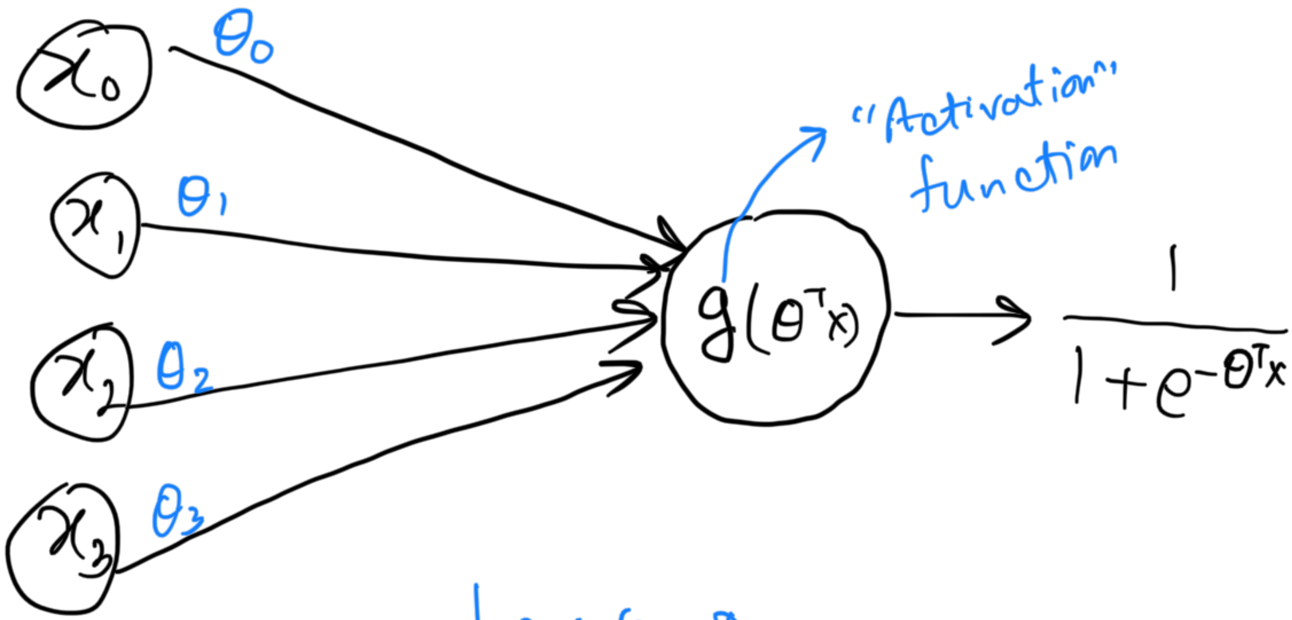
$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right]$$

Binary cross entropy loss

$$\theta_{best} = \arg\min J(\theta)$$

Optimizer
- → G.D.
- → S.G.D
- → Adam



Logistic Regression unit