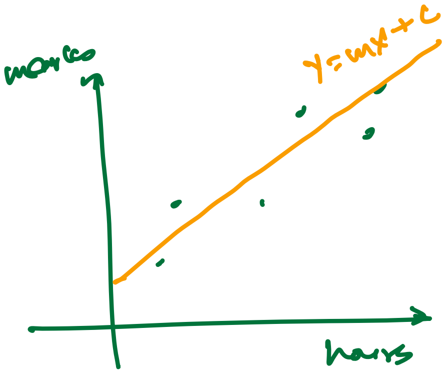
Linear Regression



m, c ?

Single feature

(x) hours	(y) marks

$y = \theta_1 x + \theta_0$ → bias

weight θ_1

θ_1, θ_0 ?

$J(\theta)$

$$\text{Loss} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y} - y)^2$$

(MSE)

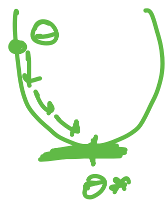
$$h_\theta(x) = \theta_1 x + \theta_0$$

→ Random θ_0, θ_1

→ How good θ_0, θ_1 is ?

→ Update θ_0, θ_1

GRADIENT DESCENT



$$\theta = \theta - \eta \frac{\partial J(\theta)}{\partial \theta} \qquad \nabla J(\theta)$$

$$\theta_0 = \theta_0 - \eta \cdot \frac{2}{m}\sum_{i=1}^{m}(\theta_1 x + \theta_0 - y)$$

$$\theta_0 = \theta_1 - \eta \cdot \frac{2}{m}\sum_{i=1}^{m}(\theta_1 x + \theta_0 - y)x^i$$

Multiple features

$$\begin{bmatrix} \text{——} \; x^1 \; \text{——} \\ \text{——} \; x^2 \; \text{——} \end{bmatrix} \rightarrow \begin{bmatrix} x^1_1 & x^1_2 & x^1_3 & \cdots & x^1_n \\ x^2_1 & x^2_2 & x^2_3 & \cdots & x^2_n \end{bmatrix}$$

$$\begin{bmatrix} \text{---}\ x^{j}\ \text{---} \\ \vdots \\ \text{---}\ x^{m}\ \text{---} \end{bmatrix} \qquad \begin{bmatrix} \vdots \\ x_1^m \quad x_2^m \quad x_3^m \quad \cdots \quad x_n^m \end{bmatrix}$$

x_1	x_2	x_3	x_4	y

$y \rightarrow$ Regression

$$h_\theta(x) = y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

bias

weights

$$h_\theta(x) = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$$

$$h_\theta(x) = \theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i \qquad x_0 = 1$$

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i \longrightarrow \text{feature!}$$

$$\left.\begin{array}{l} \theta_0 \underline{x_0} + \theta_1 x_1 \\ \quad 1 \\ \theta_0 + \theta_1 x_1 \end{array}\right\} \text{Single feature}$$

x_0	x_1	x_2	x_3	x_4	y

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$\theta^T x$

$$\begin{bmatrix} \theta_0 & \theta_1 & \theta_2 & \cdots & \theta_n \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n$$

$$\hat{y} = \boxed{h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x}$$

→ Hypothesis

→ Error / loss

→ Optimize

→ Gradient

Loss fxn:

MSE $\quad J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$

$$= \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

TASK: $\frac{\partial J(\theta)}{\partial \theta}$?

$\nabla_\theta J(\theta)$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{\partial}{\partial \theta_j} \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

$$= \frac{\partial}{\partial \theta_j} \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j} \left(\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \cdots \theta_j x_j^{(i)} + \cdots \theta_n x_n^{(i)} - y^{(i)}\right)^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} 2 \left(\underbrace{\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \cdots \theta_n x_n^{(i)}} - y^{(i)}\right) x_j^{(i)}$$

$h_\theta(x^{(i)})$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m}\sum_{i=1}^{m} 2\left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{i}$$

Final Gradient Update Rule

$$\theta_j = \theta_j - \eta \cdot \frac{1}{m}\sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) x_j^{(i)}$$

$$\hat{y}^{(i)} = h_\theta(x^{(i)}) = \sum_{i=0}^{n} \theta_i x_i = \theta^T . x$$

$[\theta_0 \quad \theta_1 \quad \cdots \quad \theta_n]$ Randomly

do

{

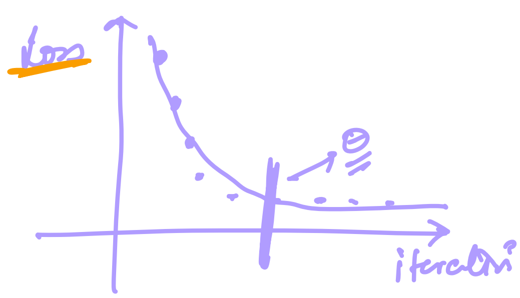
loss fn → how good our θ's are
(MSE)

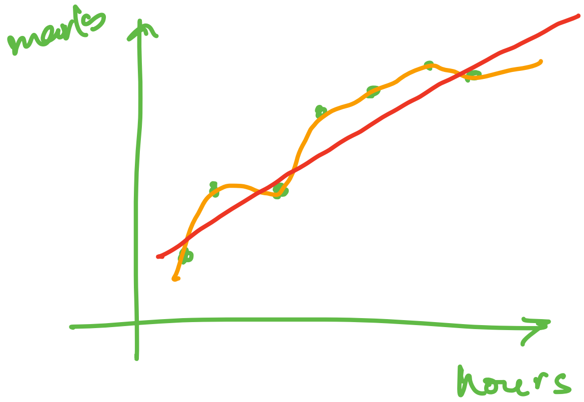
update θ

}

while (converge)

① 2500 times





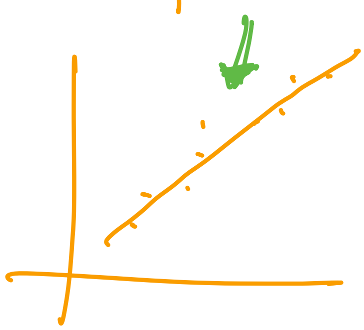
Interpolation

OVERFITTING

- not generalizable for test data points
- has info only about training data points.

ML Algo

- Supervised (y given)
  - Regression (y continuous) (Linear Regression)
  - Classification (y discrete) (Logistic Regression)
- Unsupervised
- Reinforcement



Logistic Regression

↳ Classification Algo

Eg:

→ wt, ht → Dog, cat

→ Spam or not Spam

→ Animal → Cat, Dog

} Binary Classification

Training Data:

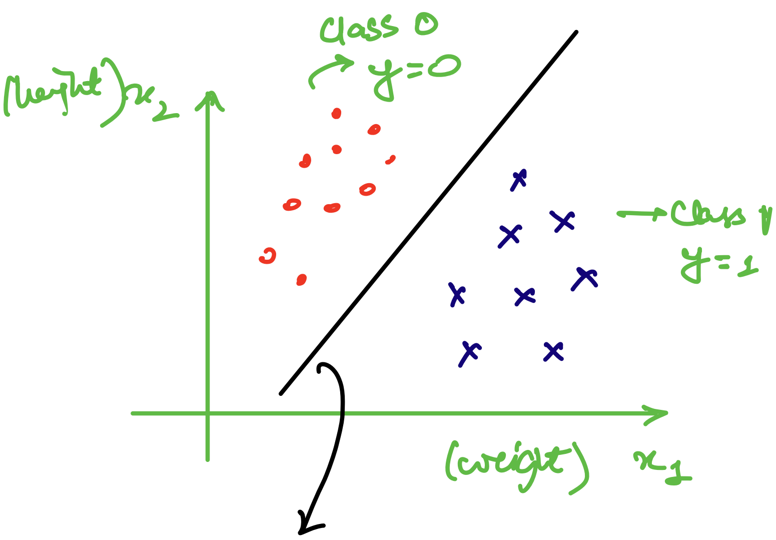
$\{x^{(i)}, y^{(i)}\}$

$x \in \mathbb{R}^n \rightarrow x$ has n features & all features are real nos.

Classification:
y should be a discrete value.

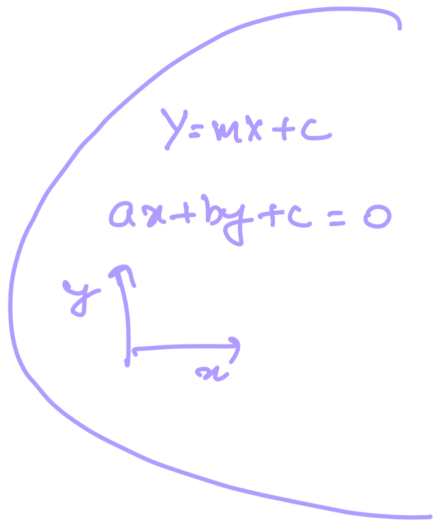
$y \in \{0, 1\}$

0 → Cat, 1 → Dog



x : cat : 1

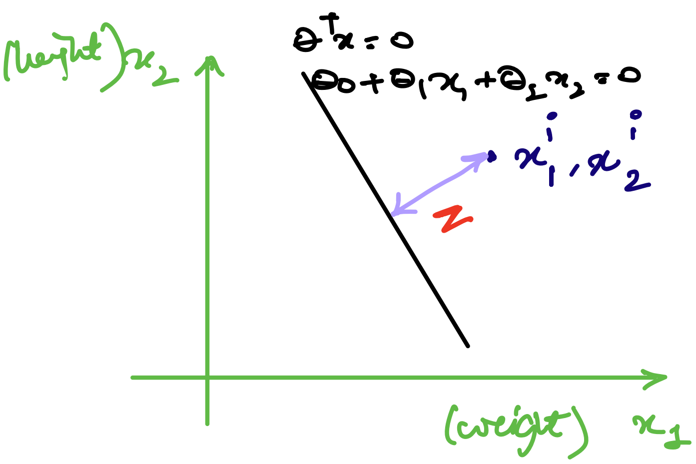
o : dog : 0

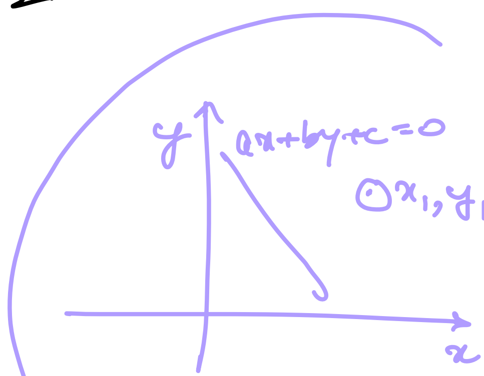


$$\underbrace{\theta_0 + \theta_1 x_1 + \theta_2 x_2}_{\theta^T x} = 0$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \quad x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \quad \theta^T x = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \qquad x_0 = 1$$

$$= \theta_0 + \theta_1 x_1 + \theta_2 x_2$$



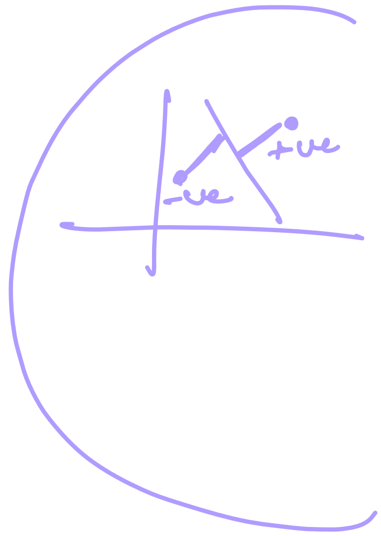


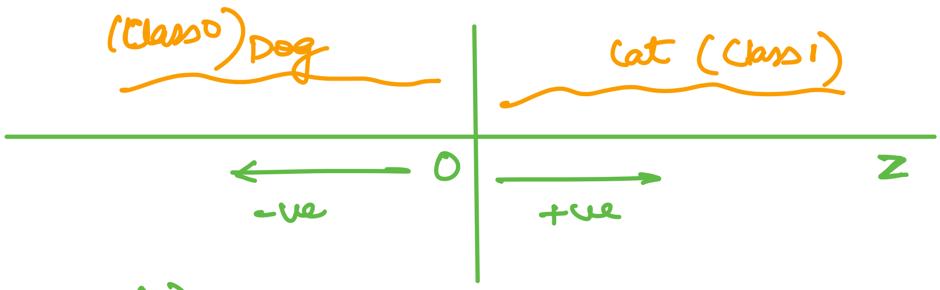
$$\frac{ax_1 + by_1 + c}{\sqrt{a^2+b^2}}$$

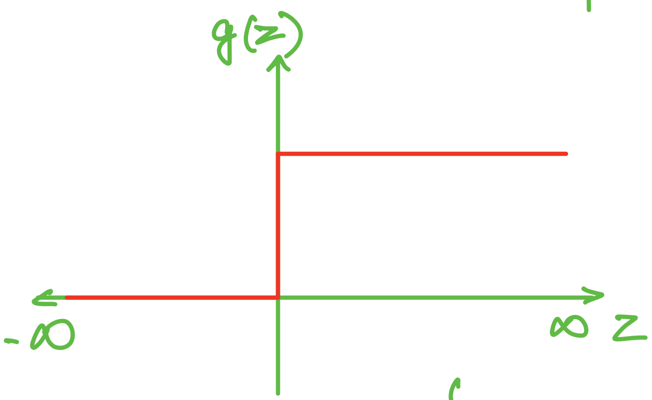
→ normalized distance

$\theta_0 + \theta_1 x_1^i + \theta_2 x_2^i$ → unnormalised distance b/w x^i & line $\theta_0 + \theta_1 x_1 + \theta_2 x_2$

$$\underbrace{[\theta_0 \quad \theta_1 \quad \theta_2]}_{\theta^T} \underbrace{\begin{bmatrix} 1 \\ x_1^i \\ x_2^i \end{bmatrix}}_{x^i} = \underbrace{\theta^T x^{(i)}}_{z}$$



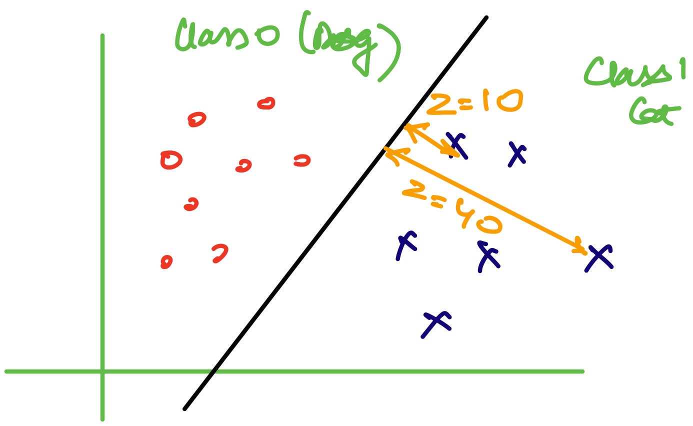




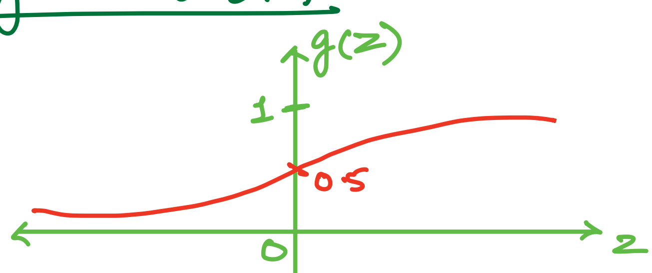
$g(z) = 1 \quad$ if $z \geq 0$

$g(z) = 0 \quad$ if $z < 0$

PROBLEM? Not able to distinguish b/w points which are close to line & which are far.



Sigmoid function



$$g(z) = \frac{1}{1+e^{-z}}$$ sigmoid fn

$z = \infty \qquad g(z) = \dfrac{1}{1 + \frac{1}{e^{z}}} = 1$

$e^{z} \to \infty, \quad \frac{1}{e^{z}} \to 0$

$z = 0 \qquad g(z) = \dfrac{1}{1 + \frac{1}{e^{0}}} = \frac{1}{2} = 0.5$

$e^{0} = 1, \quad \frac{1}{e^{0}} = 1$

$z = -\infty \qquad g(z) = \dfrac{1}{1 + \frac{1}{e^{-\infty}}} = \dfrac{1}{1 + e^{\infty}} = 0$

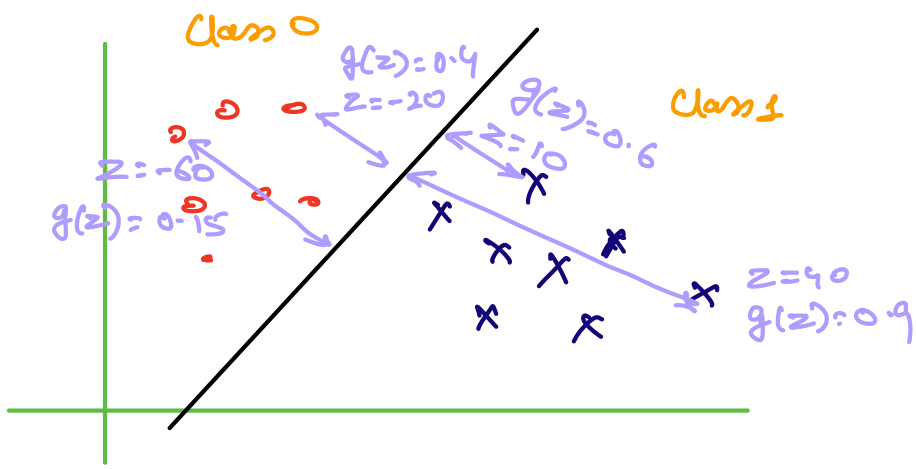
$e^{\infty} = \infty$

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

$$g(z) = \frac{1}{1 + e^{-z}} \qquad \text{where } z = \theta^T x$$

value b/w 0 & 1

Probability / Confidence with which you can say the point belongs to class 1



$g(z) = 0.4$

40% sure points belongs to class 1

60% sure point belongs to class 0

point lies on the line $g(z) = 0.5$

50% sure belongs to class 0

class 1

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

Probability point ∈ class 1

$$\hat{y} = 1 \quad \text{if} \quad h_\theta(x) \geq 0.5$$

$$= 0 \quad \text{if} \quad h_\theta(x) < 0.5$$

Loss fxn:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

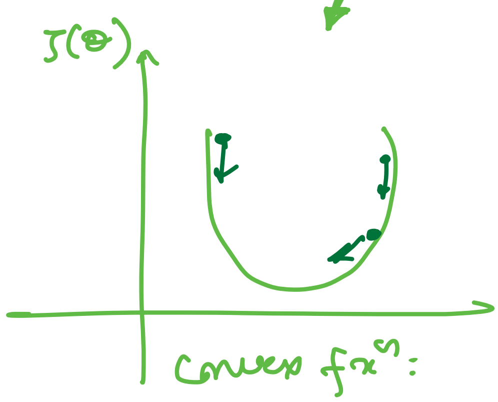
→ Hypothesis ✓
→ Error
→ Gradient

Linear Regression

$$h_\theta(x) = \theta^T x$$
$$= \theta_0 + \theta_1 x_1 + \dots \theta_n x_n$$

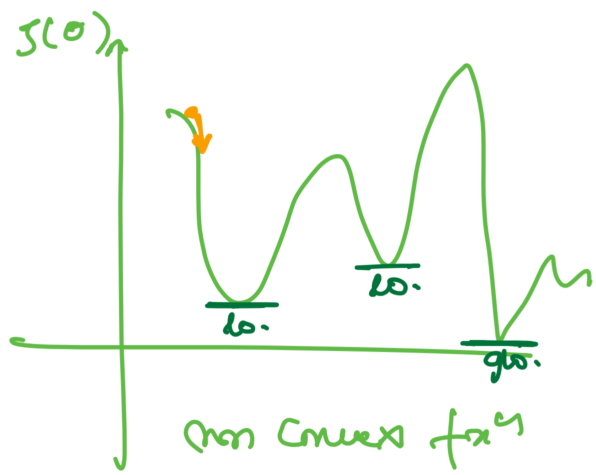
Logistic Regression

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



Convex fxn:

Local minima = Global minima



Non Convex fxn

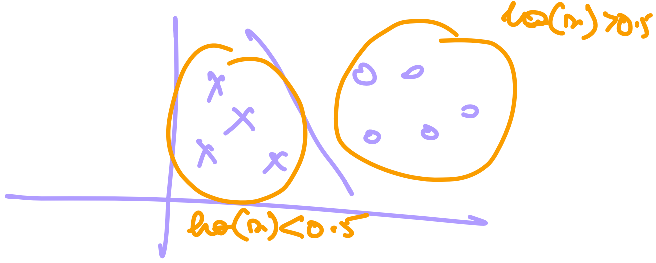
Binary Cross Entropy

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

Prob. that point belongs to class 1

$$P(y=1 \mid x;\theta) = h_\theta(x)$$

$$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$$



0.6 60% Class 1

40% Class 0

Bernoulli Distribution

Probability Mass function

$$P(y \mid x;\theta) = [h_\theta(x)]^{y}\,[1-h_\theta(x)]^{1-y}$$

$y=1$: $h_\theta(x)$

$y=0$: $1-h_\theta(x)$

$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

$P(A \text{ and } B) = P(A) \cdot P(B)$

Likelihood

$$P(y^{(1)} y^{(2)} \ldots y^{(m)} \mid x^{(1)} x^{(2)} \ldots x^{(m)}; \theta)$$

$$= \underline{P(y^{(1)} \mid x^{(1)};\theta)} \cdot \underline{P(y^{(2)} \mid x^{(2)};\theta)} \;\cdots\; \underline{P(y^{m} \mid x^{m};\theta)}$$

$$= \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)};\theta)$$

likelihood of the data y.

$$LL(\theta) = \prod_{i=1}^{m} [h_\theta(x)]^{y}\,[1-h_\theta(x)]^{1-y}$$

Loss minimize

LL maximize