Day-3

June 17, 2024

Last Session:

Linear Regression:

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

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

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

Gradient Descent:

$$\theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \eta \cdot \nabla_{\theta_j} J(\theta)$$

$$\theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \eta \cdot \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

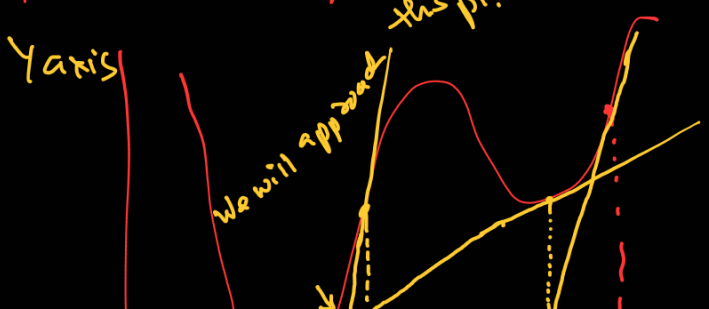
Normal Equations:

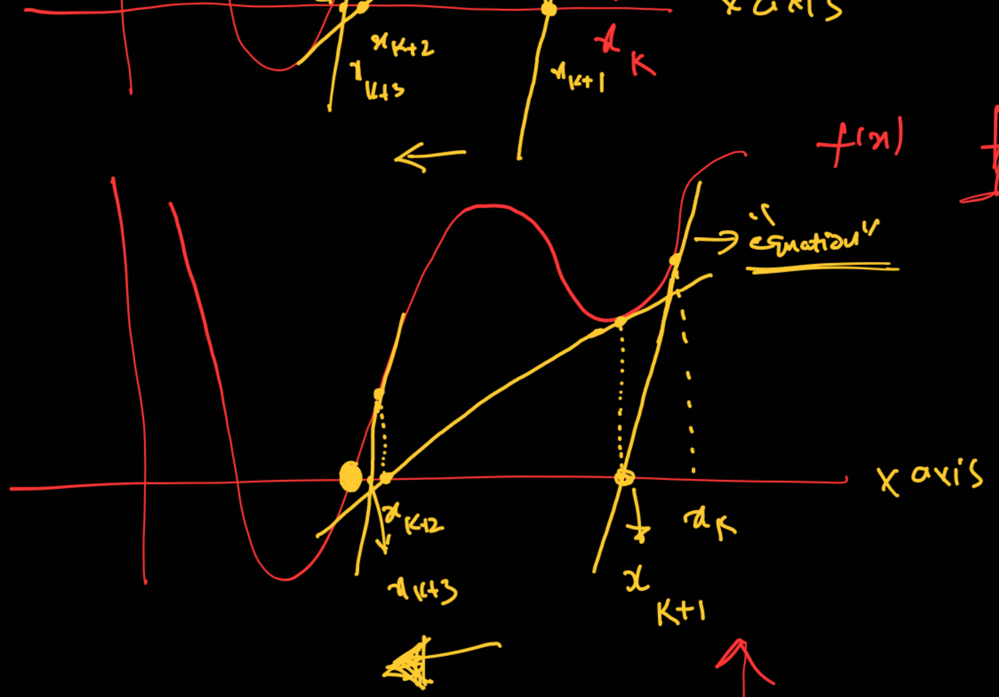
$$J(\theta) = \frac{1}{2m} (y - X\theta)^T (y - X\theta)$$

$$\boxed{\theta = (X^T X)^{-1} X^T y}$$

Newton's Method (Iterative method)

If we have find roots of an eqn $f(x) = 0$





$f(x)=0$

$$x^{t+1} = x^{t} - \frac{f(x)}{f'(x)}$$

"Newton's Update Rule"

$f'(x)=0$

→ **Optimization:**

We are not interested in $f(x)=0$, rather we are interested in finding $f'(x)=0$

$$f(x)=0 \Rightarrow x^{t+1} = x^{t} - \frac{f(x)}{f'(x)}$$

$$f'(x)=0 \Rightarrow x^{t+1} = x^{t} - \frac{f'(x)}{f''(x)}$$

$f''(x)$ → 2nd order derivative term.

In general, 2nd order methods are faster than 1st order methods. !! ×××

Let $f: \mathbb{R}^n \to \mathbb{R}$

... "a" (for x) using

From given a, we can approximate ... (Taylor's Theorem.)

(Vector)

$$f(x) \approx f(a) + g^T(x-a) + \frac{1}{2}(x-a)^T H (x-a),$$

where $g = \nabla f(a)$, $H = \nabla^2 f(a)$

Vector's

$$\rightarrow f(x) \approx f(a) + \nabla f(x)(x-a) + \frac{\nabla^2 f(x)(x-a)^2}{2!} + \cdots$$

$f(x)$ ↓ Scalar

$$f(x) = f(a) + g^T(x-a) + \frac{1}{2}(x-a)^T H (x-a)$$

$$= f(a) + g^T(x-a) + \frac{1}{2}(x^T - a^T) H (x-a)$$

$$= \frac{1}{2}\left[x^T H x - 2a^T H x + a^T H a\right] + g^T(x-a) + f(a)$$

$$= \frac{1}{2} x^T H x + b^T x + c, \quad \text{where } b = (g - H^T a), \quad c = \text{left terms}$$

I have to bring this eqn this form

$$= \frac{1}{2}\left[x^T H x - 2a^T H x + a^T H a\right] + g^T(x-a) + f(a)$$

$$= \frac{a^T H x}{2} - \frac{a^T H x}{2} + \frac{a^T H a}{2} + g^T x - g^T a + f(a)$$

$ax^2 + bx + c$

$$= \frac{x^T H x}{2} + g^T x - a^T H x + \frac{a^T H a}{2} - g^T a + f(a)$$

$$= \frac{x^T H x}{2} + (g^T - a^T H)x + \frac{a^T H a}{2} - g^T a + f(a)$$

$(a^T H)^T$ $(H^T a)$

$$g(x) = \frac{x^T H x}{2} + (g - H^T a)^T x + \frac{a^T H a}{2} - g^T a + f(a)$$

$\frac{1}{2}$

$\nabla q(x) = 0$

$$\frac{\partial}{\partial x}\left(\frac{1}{2}Hx + (g - H^T a)\right) = 0$$

$$\frac{\partial}{\partial x}\left(\frac{1}{2}x^T H x\right) = Hx$$

$$\nabla_x a^T x = a$$

$$Hx = H^T a - g$$

$$x = H^{-1}(H^T a - g)$$

$$x = H^{-1}(Ha - g)$$

$$\boxed{x = a - H^{-1}g}$$

$H = H^T$ → Symmetric matrix

$I\,a = a$

$$\tilde{x} = a - H^{-1}g$$

x_{t+1}

$\nabla^2 q(x)$

$\nabla_x q(x) = Hx - (g - H^T a)$ ← Not dependent on x (0)

$\nabla_x^2 q(x) = H$ → matrix "Hessian Matrix"

Hessian matrix → That must be "Positive Semi" Definite (PSD)

Vec: z → $\boxed{z^T H z \geq 0}$ ← Fact !!

Algorithm

$$x_{t+1} = x_t - f'(x) \to \nabla_\theta J(\theta)$$

$$f''(x) \to \nabla^2_\theta J(\theta)$$

Diff Gradient

Newton: $\theta^{(t+1)} = \theta^{(t)} - (H^{-1}) \nabla_\theta J(\theta)$

H^{-1}: $\nabla^2 J(\theta)$ — $\nabla_\theta J(\theta)$

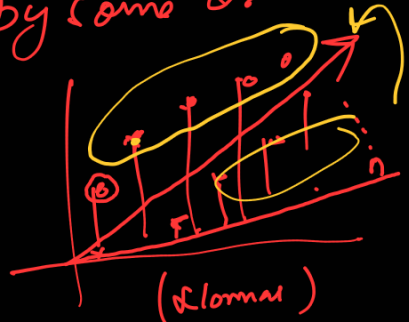
descent.

GD: $\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla_\theta J(\theta)$

Probabilistic Interpretation of Linear Regression

labels (target)

In probabilistic interpretation, we are concerned how $y^{(i)}$ are generated given by $x^{(i)}$ parameterized by some θ.

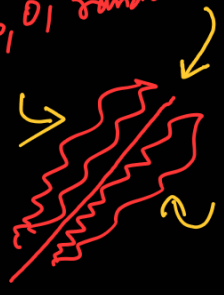


For Linear Regression,

$$y^{(i)} \approx \theta^T x^{(i)}$$

"I am expecting" that my data is distributed θ_0, θ_1 random

assume (along) my model (i.e. $\theta_0 + \theta_1 x$) here.



$\approx$ ← approximation sign

$y^{(i)} \approx \theta^T x^{(i)}$ ← x_0

$y^{(i)} = \theta^T x^{(i)} + \epsilon$ ← "noise"

$\epsilon \sim N(0, \sigma^2)$

mean variance

$\theta^T x$ ($m \times 1$)

matmul

θ: $(n+1) \times 1$

x: $m \times (n+1)$

$X\theta$

$(n+1) \times 1$

$y^{(i)} = \theta^T x^{(i)} + \epsilon$

$\epsilon \sim N(0, \sigma^2)$

$(m \times (n+1))$ $((n+1) \times 1)$

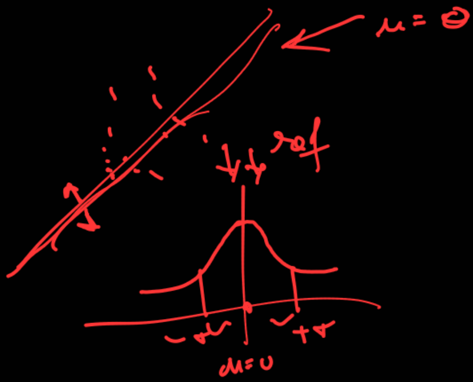
$(m \times 1) \leftarrow Y$ dim

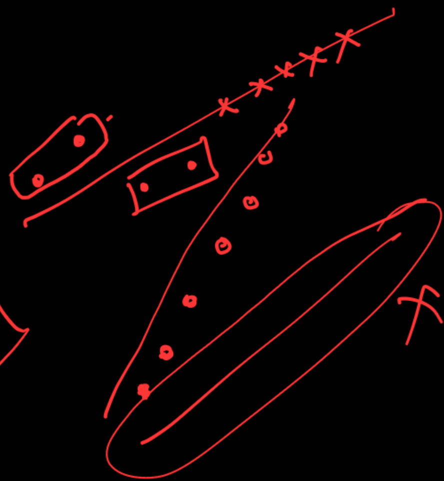
$y^{(i)} = \theta^T x^{(i)} + \epsilon \rightarrow \epsilon \sim N(0, \sigma^2)$

mean = 0

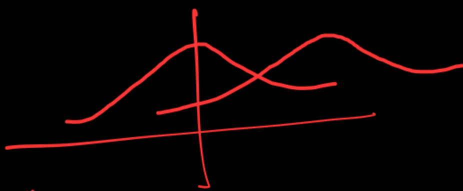
Now this becomes the eqn of how data is generated given some $x^{(i)}$

This is 0, because we want on average, the point must follow the line.





$\mu = 0$



$y^{(i)} = \theta^T x^{(i)} + N(0, \sigma^2)$

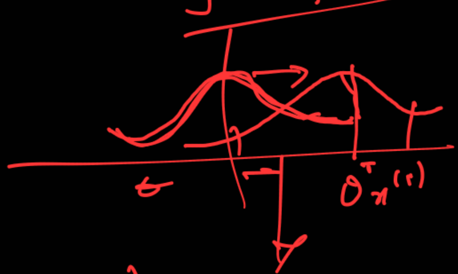
$\mu = 0$

+5

noise $\mu = 0$

This is interpreted as: -

$P(y^{(i)} | x^{(i)}; \theta) \sim N(\theta^T x^{(i)}, \sigma^2)$



(... distribution of labels) Noise

I am assuming that distribution of $y^{(i)}$ is Normal with avg $\theta^T x^{(i)}$ and some std dev. This also means that on an average, my $y^{(i)}$'s should fall on the line $\theta^T x^{(i)}$.

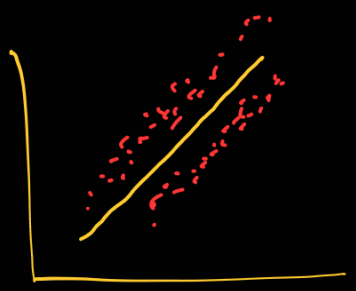
$$p(y^{(i)} | x^{(i)}; \theta) \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$

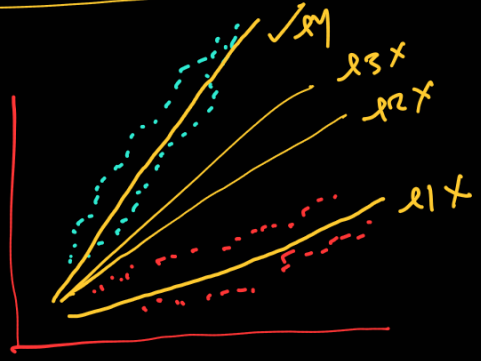
$$p(y^{(i)} | x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

→ Now we define the "LIKLIHOOD" of the data, given the model.

→ We are asking that "how likely" is this data generated by the given model.





We can perform "MLE" (maximum likelihood estimation) to find the model that best generates the data.

$$\theta_{MLE} = \underset{\theta}{\text{argmax}}\ P(y^{(1)} \dots y^{(m)} \mid x^{(1)} \dots x^{(m)}; \theta)$$

$y^{(1)} \dots y^{(m)}$: dependent data. $x^{(1)} \dots x^{(m)}$: given data.

$f(x) = 2x^2 + 3x + 4$

max/min value of $f(x)$

argmax/argmin → I am trying to find out that x value which optimizes the given $f(x)$

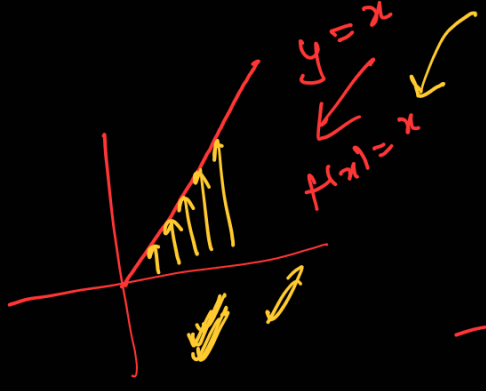
Any distribution

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}; \theta\right) \Leftarrow P\left(y^{(1)}, y^{(2)} \ldots y^{(m)}\right) \Leftarrow \text{Joint dist}$$

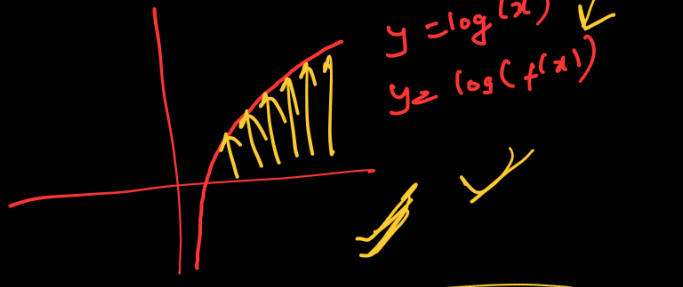
$$= p\left(y^{(1)}\right) \cdot p\left(y^{(2)}\right) \cdot p\left(y^{(3)}\right) \ldots$$

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ - \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)}{2\sigma^2} \right\}$$



(f(x) → to optimize)
(log(f(x)) →)



x → maximize
log(x) → maximize

$$\theta_{MLE}^{\max} \longrightarrow \log\left(\theta_{MLE}\right)^{\max}$$

$$\theta_{MLE} = L(\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}; \theta\right) \rightarrow X$$

$$LL(\theta) = \underset{\theta}{\operatorname{argmax}} \log \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

Log Likelihood

$$L(\theta) = \arg\max_{\theta} \log \prod_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta\right)$$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \log p\left(y^{(i)} | x^{(i)}; \theta\right)$$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}\right)$$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \left(\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right] + \left[\log\left(e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}\right)\right]\right)$$

$f(x)$, $\frac{f(x)}{4}$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \log e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

$$\arg\max_{\theta} \sum_{i=1}^{m} -(y^{(i)} - \theta^T x^{(i)})^2$$

$2\sigma^2 \rightarrow$ constant

$$\arg\max_{\theta} \frac{1}{2\sigma^2 m} \sum_{i=1}^{m} -(y^{(i)} - \theta^T x^{(i)})^2$$

$$\arg\max \frac{m}{\sigma^2} \times \frac{1}{2m} \sum_{i=1}^{m} -(y^{(i)} - \theta^T x^{(i)})^2$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$ "Im gradient descent"

$f(x)$, $f(x)$, $\frac{f(x)}{4}$

$J(\theta) \rightarrow$ minimize

$$\arg\max \frac{m}{4\sigma^2} \times (-1) \times \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

$$\sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

$\text{argmax} \; (-1) \frac{1}{2m} \sum_{i=1}^{m} (\quad)$

$\text{argmax} \; (-1) \; J(\theta)$

$\text{argmax} - J(\theta)$

If I max $-f(x)$, then what I am effectively doing $\rightarrow$ $\min f(x)$

$\text{argmax} - J(\theta)$

$\text{argmin} \; J(\theta)$

$\text{argmax} \; LL(\theta) = \text{argmin} \; J(\theta)$

" We were right about Probabilistic Int of Lin Reg. & that both approaches are interchangable "

Logistic Regression (Classification)

Cat vs Dog

L.R $\rightarrow$ continuous

[00.28, 79.67, 1821.224

Log. Reg $\rightarrow$ Labels depending how many Categories I have.

70.1. $\rightarrow$ Cat

30.1. $\rightarrow$ Dog

50.1.9 $\rightarrow$

(100-50).1.

Binary Classification: 2 categorical

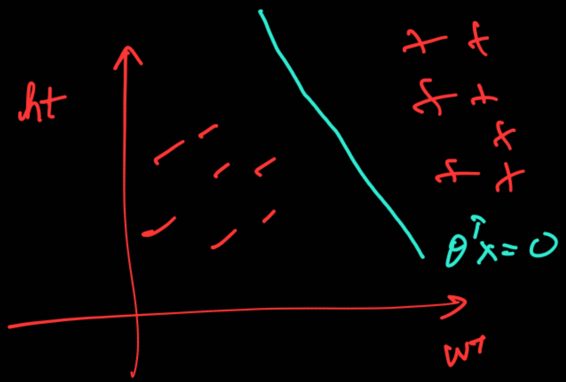
0.78 $\rightarrow$ 1/0

2 classes

0.23 → 0 | 1

$y = \{0, 1\}$

$y^{(i)} = \{0, 1\}$ → ϕ, $p, (1-p)$ dist



then $y^{(i)} | x^{(i)}; \theta \sim \text{Bernoulli}(p)$

Binomial

Bernoulli ϕ

↓ repeated bernoulli trials

Multinoulli (ϕ) = 1

$P < 1$

$0 \to p_1$
$0 \to p_2$
$0 \to p_3$
$0 \to p_4$
$0 \to 1 - (p_1 + p_2 + p_3 + p_4) = p_5$

$(p \,|\, (1-p))$ = 1

p, $(1-p)$

$^nC_x \, p^x q^{n-x}$

↓ $(1-p)$

$^nC_x \, p^x (1-p)^{n-x}$

W-18 teams → Top 8

T1 . . . T8

P1 - [. . . . . - P7 - 1 (1 - ...)

ϕ_m

(1) 100 chocolates

40 (0.4) | 10 (0.1) | 50 (0.5)

$(y^{(i)} | x^{(i)}; \theta) \sim \text{Bernoulli}(p)$

Let $d^{(i)}$ be the dist b/w the line and the point (i), then ideally to have a high prob for a pt that is farther from the line & vice-versa,

D is effectively giving me probability

$-\infty < d^{(i)} < \infty \xrightarrow{\text{map}} g(d) \in (0,1)$

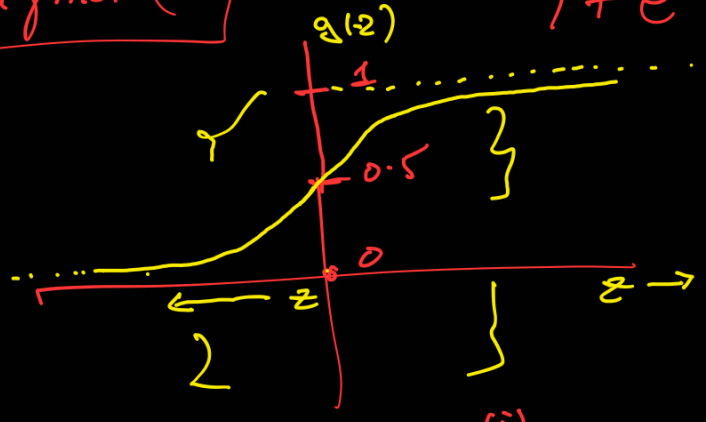
differentiable, ✓

$(0,1)$ ✓

$z \to \infty \Rightarrow g(z) = 1$ ✓

$z \to -\infty \Rightarrow g(z) = 0$ ✓

Sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$

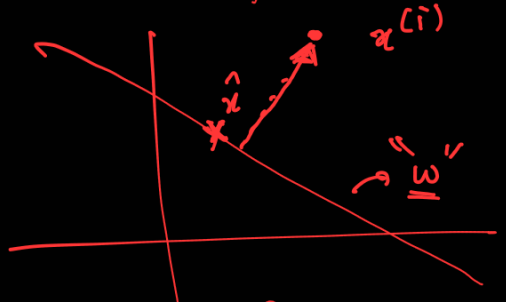


$d^{(i)} = w^T(x^{(i)} - \hat{x})$

$d^{(i)} = w^T x^{(i)} - w^T \hat{x}$

$d^{(i)} = w^T x^{(i)} + b$

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



$\theta_0 x_0 + \theta_1 x_1 + \dots \theta_n x_n$

$[\theta_0][\theta_1\ \theta_2 \dots \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$

$\theta_1 x_1 + \dots + \theta_n x_n + \theta_0 (x_0 = 1)$

$d^{(i)} = w^T(x^{(i)} - \hat{x})$

Fact ↑

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

$\downarrow$

$g(\theta^T x^{(i)})$

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

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

$\sigma'(z) = \sigma(z)(1-\sigma(z))$

Derive ↑↑↑ H.W

Fact.

$\sigma(z) = \frac{1}{1+e^{-z}}$

$P(y^{(i)} = 1 \mid x^{(i)}; \theta) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$

$P(y^{(i)} = 0 \mid x^{(i)}; \theta) = \left(1 - \frac{1}{1+e^{-\theta^T x^{(i)}}}\right)$

$p(y^{(i)}=1 \mid x^{(i)};\theta) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$ $\quad 1 - \frac{1}{1+e^{-\theta^T x^{(i)}}}$

Bernoulli: $p(y=0)^{(y)}\,(1-p(y=0))^{(1-y)}$ $\quad \phi^y \cdot y=\phi$ $\quad (1-\phi)^{1-y}$ $\quad y=0$

Var (p):

ϕ

$= \phi^y (1-\phi)^{1-y}$

Fact

log Likelihood: → log of prob was sum → either 1 or 0 "Binary classification"

$LL(\theta) = \sum_{i=1}^{m} \log p(y^{(i)} \mid x^{(i)};\theta)$ $\quad p(y^{(i)}=c \mid x^{(i)};\theta)$

$\log(L(\theta))$ $\quad \sum_{i=1}^{m} \log p(y^{(i)} \mid x^{(i)};\theta) = L(\theta)$ $\quad \log\left(\phi^y (1-\phi)^{1-y}\right)$

$LL(\theta) = \sum_{i=1}^{m} y^{(i)} \log p(y^{(i)}=1 \mid x^{(i)};\theta) + (1-y^{(i)}) \log\left(1 - p(y^{(i)}=1 \mid x^{(i)};\theta)\right)$

$\theta \quad z = B(p) = \phi^y (1-\phi)^{1-y}$

$\log(L) = \log \phi^y (1-\phi)^{1-y}$

$= y \log\phi + (1-y)\log(1-\phi)$

$h_\theta(x^{(i)}) = p(y^{(i)}=c \mid x^{(i)};\theta) = \frac{1}{1+e^{-\theta^T x^{(i)}}} = \sigma(\theta^T x^{(i)})$

$LL(\theta) = \sum_{i=1}^{m} y^{(i)} \log \underbrace{p(y^{(i)}=1 \mid x^{(i)};\theta)}_{\sigma(\theta^T x^{(i)})} + (1-y^{(i)}) \log \underbrace{\left(1 - p(y^{(i)}=1 \mid x^{(i)};\theta)\right)}$

$LL(\theta) = \sum_{i=1}^{m} y^{(i)} \log \sigma(\theta^T x^{(i)}) + (1-y^{(i)}) \log\left(1-\sigma(\theta^T x^{(i)})\right)$

$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} \left(\frac{y^{(i)}}{\sigma(\theta^T x^{(i)})} \cdot \sigma(\theta^T x^{(i)}) \cdot (1-\sigma(\theta^T x^{(i)})) \cdot x^{(i)} \right.$ $\quad$ Vector

$+$

$\left. (1-y^{(i)}) \cdot \frac{0 - (\sigma(\theta^T x^{(i)}) (1-\sigma(\theta^T x^{(i)}) \cdot x^{(i)})}{1-\sigma(\theta^T x^{(i)})} \right)$

$(1-\sigma(\theta^T x^{(i)}))$

$g(x) = \log f(x)$

$g'(x) = \frac{1}{f(x)} \cdot f'(x)$

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} y^{(i)}(1-\sigma(\theta^T x^{(i)}))x^{(i)} + (-1)(1-y^{(i)})\sigma(\theta^T x^{(i)})x^{(i)}$$

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} x^{(i)}\left(y^{(i)} - y^{(i)}\sigma(\theta^T x^{(i)}) + \left((y^{(i)}-1)\sigma(\theta^T x^{(i)})\right)\right)$$

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} x^{(i)}\left(y^{(i)} - y^{(i)}\sigma(\theta^T x^{(i)}) + y^{(i)}\sigma(\theta^T x^{(i)}) - \sigma(\theta^T x^{(i)})\right)$$

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} x^{(i)}\left(y^{(i)} - \underbrace{\sigma(\theta^T x^{(i)})}_{h_\theta(x^{(i)})}\right)$$

$$\boxed{\nabla_\theta LL(\theta) = \sum_{i=1}^{m}\left(y^{(i)} - h_\theta(x^{(i)})\right)x^{(i)}}$$

$\leftarrow \nabla D$

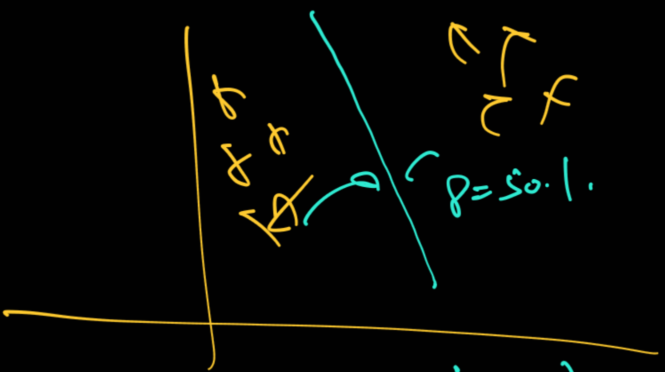
MSE loss derivative

Logistic Regression and Linear Regression have similar properties

$$\theta_j \leftarrow \theta_j^{(t+1)} = \eta \cdot \nabla_\theta J(\theta)$$

$\max LL(\theta) = \min J(\theta)$

Generalized Class of Linear Models (GLM)



$$p(y^{(i)}=1) = \frac{1}{2} = p(y^{(i)}=0)$$

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

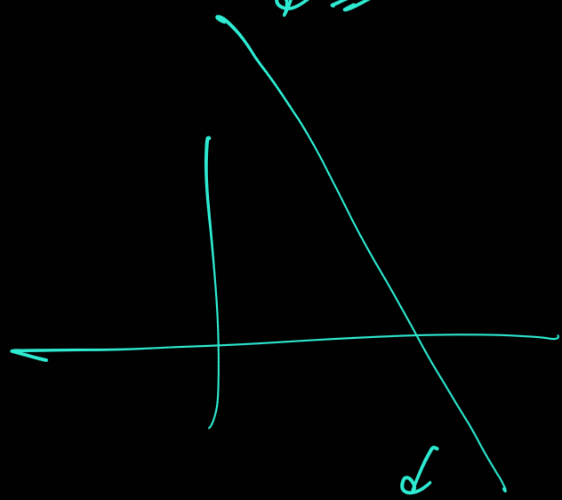
$$\frac{2}{1+e^{-\theta^T x^{(i)}}} = 1$$

$$1+e^{-\theta^T x^{(i)}} = 2$$

$$e^{-\theta^T x^{(i)}} = 1 = e^{0}$$

$$\boxed{\theta^T x^{(i)} = 0}$$

$\theta_0 \dots \theta_i$



2D

Class 1

Class 0

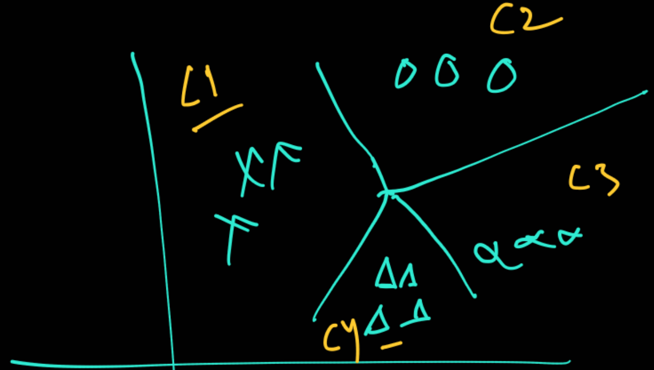
And this the eqn of sep line. ← Seperator line

hone kii prob = 50%.

1, 2, 3, 4

One vs All

(2, 3, 4)



[vs (1,3)

2 vs

$$P(y=1) = \frac{\binom{100}{4}}{100} = P(y=4)$$

Multi-class

log regression $P(y=1) = 0.25 = P(y=4)$

$$P(y=2) = \frac{\binom{100}{4}}{(100)} = P(y=3)$$