Day -3

June 17, 2024

Last Session:

Linear Regression:

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

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

$$\frac{\partial}{\partial \theta_j} J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)} \rightarrow \text{Grad of MSE-Loss}$$

Gradient Descent:

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

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

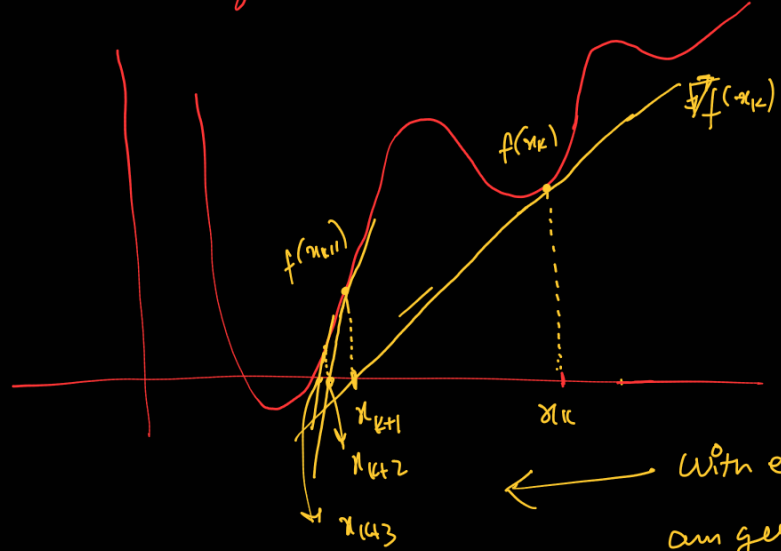
Normal Equations:

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

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

Newton's Method: (Iterative Method) of optimization.

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



← With each moving iteration, I am getting closer & closer to $f(x) = 0$

$x^{k+1} = x^k - \frac{f(x)}{f'(x)}$ → $f(x) = 0$ (Iterative process)

↪ "Newton's Update Rule"

→ Optimization

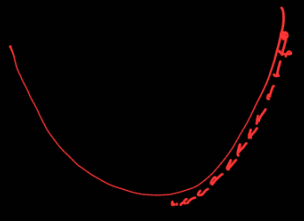
We are not interested in $f(x) = 0$, rather we are interested in finding $f'(x) = 0$

$f(x) = 0 \rightarrow x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$

$f'(x) = 0 \rightarrow x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$ → 2^{nd} order derivative term

Newton's Method involves 2^{nd} order derivatives.

→ In general 2^{nd} order methods perform faster than 1^{st} order methods.



→ SGD (GD)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$.

then given a, we can approximate the pts near "a" (let x) using Taylor's Theorem,

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

where $g = \nabla_x f(x)$, $H = \nabla_x^2 f(x)$

← Vector form of Taylor

← scalar

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

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

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

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

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

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

$$= \boxed{\frac{1}{2}x^T Hx + b^T x + c} \rightarrow \text{I want a quad eqn in } x.$$

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

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

$$\frac{x^T Hx}{2} + \boxed{b^T}x + c$$

$$b = g - Ha$$

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

$$q(x) = \frac{x^T Hx}{2} + \underbrace{(g - Ha)}_{A}x + \frac{a^T Ha}{2} - g^T a + f(a)$$

$$\nabla q(x) = \frac{2Hx}{2} + (g - Ha) + 0$$

$\nabla(\theta^T A\theta) = 2A\theta$

$\nabla_\theta(a^T\theta) = a$

$$\nabla q(x) = Hx + (g - Ha)$$

$$\nabla q(x) = 0$$

$$Hx = -(g - Ha)$$

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

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

(H: Symmetric)

$x = a - Hg$

$\nabla q(x) = Hx + (g - Ha)$

$\nabla^2 q(x) = H + 0 = H$ → "Hessian Matrix" → "Semi-Positive definite" (PSD)

$\Downarrow$

$z^T H z \geq 0$ Ans

Algorithm:

$$x_{t+1} = x_t - \frac{\nabla f(x)}{\nabla^2 f(x)}$$

$$x_{t+1} = x_t - \frac{\nabla J(\theta)}{\nabla^2 J(\theta)}$$

Newton: $\theta^{(t+1)} = \theta^{(t)} - (H^{-1}) \nabla J(\theta)$ → Newton's Update rule.

H^{-1}: $\nabla^2_\theta J(\theta)$, $\nabla J(\theta)$

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

Probabilistic Interpretation of Linear Regression

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

For Linear Regression,

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

I am "expecting" that my data is distributed around (along) my model $(\theta_0 + \theta_1 x)$ here.

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

ϵ → "noise"

(1) $J(\theta)$ — We have data pts first

(2) We have line first

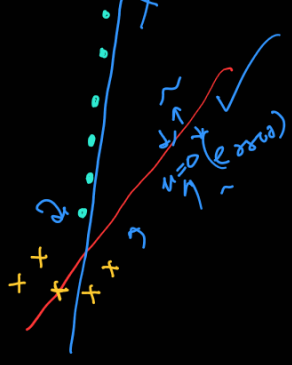
$y^{(i)} = \theta^T x^{(i)} + \epsilon \qquad \epsilon \sim N(0,1)$

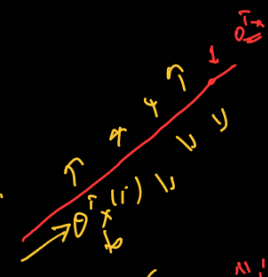
Now this becomes the eqn of how data in general given some $x^{(i)}$

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

$\mu = 0$ (mean = 0)

($\epsilon = 0$) This is because we want on an average, the pt should fall on the line





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

This can be interpreted as: —

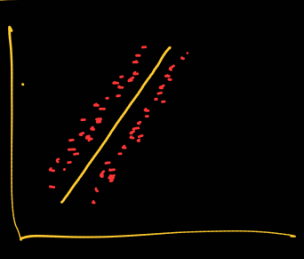
$$P(y^{(i)} | x^{(i)}; \theta) \sim N(\theta^T x^{(i)}, \sigma^2)$$

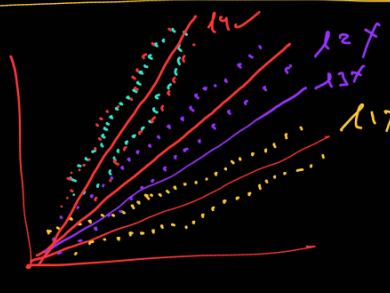
I am assuming that distribution of labels $y^{(i)}$ is Normal, with avg $\theta^T x^{(i)}$ and some standard deviation. This also means that on an average, my $y^{(i)}$'s should fall on the line $\theta^T x^{(i)}$

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

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

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





"We can perform "MLE" to find the model that best generated the data.

(max" likelihood estimation)

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\ p(y^{(1)}, y^{(2)}, \ldots, y^{(m)} \mid x^{(1)}, x^{(2)}, \ldots, x^{(m)}; \theta)$$

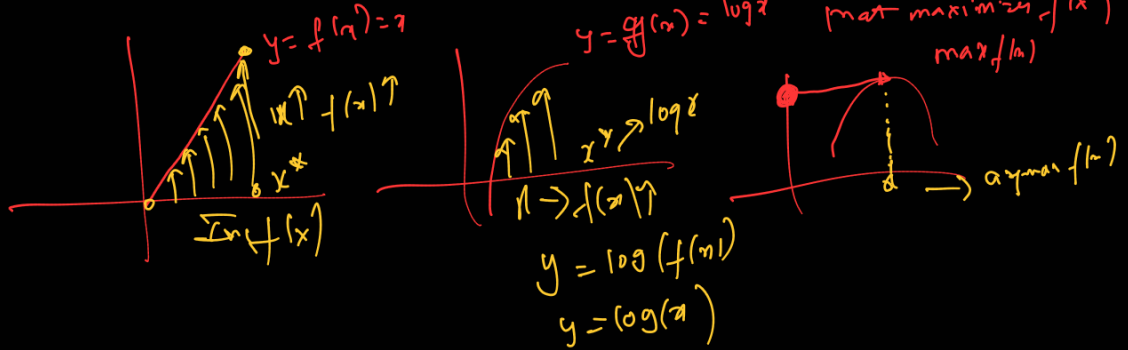
Joint dist Joint distribution

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

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

$\max f(x)$: I want max value of $f(x)$

$\operatorname{argmax} f(x)$: I want that "x" that maximizes $f(x)$



$\theta_{MLE} \Rightarrow LL_{MLE} = LL(\theta)$

"Log likelihood"

"Easy" "Take"

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

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

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

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

$$LL(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \sum_{i=1}^{m} \log e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

$f(x) \rightarrow C + f(x)$

$$LL(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

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

$$LL(\theta) = \underset{\theta}{\operatorname{argmax}} - \sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})}{2\sigma^2}$$

$2\sigma^2$ constant

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

$$\frac{1}{2\sigma^2 \cdot m}$$

$$Lh(\theta) = \frac{m}{\sigma^2} \underset{\theta}{\operatorname{argmax}} \frac{1}{2m} \sum_{i=1}^{m} -\left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

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

$$\min f(x) \Big|_{x} = -f(x)\Big|_{\max}$$

$$\max f(x) \Big|_{x} = -f(x)\Big|_{\min}$$

$$Lh(\theta) = \frac{m}{\sigma^2} \underset{\theta}{\operatorname{argmax}} - J(\theta)$$

$$Lh(\theta)\Big|_{\max} \quad -J(\theta) \uparrow \max \qquad J(\theta) \downarrow \text{minimize}$$

$$Lh(\theta)\Big|_{\max} = J(\theta)_{\text{minimization}}$$

"We were right about Probabilistic Int of Linear Regression ✓ these both are interchangable"

Logistic Regression (Classification) → (Cat vs Dog)

Linear Regression → continous domain:
100.28, 75.67, 10,821.274 ...

70% chance that is a cat

30% chance that it is a dog

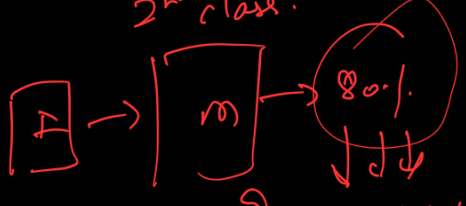
Logistic Regression → Labels depending on how many categories I have.

50% ↑ → 1 class

Binary Classification (2 Categories)

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

0.78 → 1
(1 − 0.78) → 0
0.22

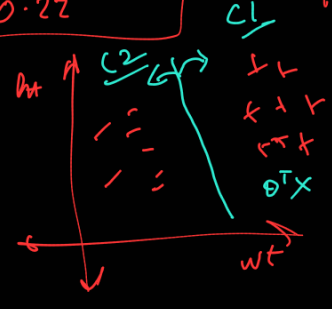


2nd is obviously.

$y^{(i)} = \{0, 1\}$

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

$\rightarrow p^y (1-p)$



Bernoulli $(\phi) = 1$ [Total]

Multinoulli $(\phi_m) = 1$

Binomial Distribution

↳ Repeated bernoulli trials

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

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

Multinomial Dist

↳ Repeated multinoulli trials

$1 - (p_1 + p_2 + p_3 + p_4)$

8 Teams

T1 T2 · T3 T6

P(D2 ··· $1-(p_1+p_4)$

Φ_m (100) chocolates

40	10	50
0.4	0.1	0.5

$$\left(y^{(i)} \mid x^{(i)}; \theta\right) \sim \text{Bernoulli}(\phi)$$

Let $d^{(i)}$ be the distance b/w the line and the point $(x^{(i)})$ then ideally it have a high probability for a point that is farther from the line & vice-versa.

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

Actually I am mapping my distances from the line into probabilities.

differentiable ✓

$(0,1)$ ✓

$z \to \infty \Rightarrow g(z) = 1$ ✓

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

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

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

$$d^{(i)} = w^T x^{(i)} - w^T \vec{x} \quad (w^T\vec{x} \to \text{Constant})$$

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

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

Similarity

UFE: ⇒

time

x_0 is always = 1

$$d^{(i)} = w^T\left(x^{(i)} - \vec{x}\right)$$

$$\begin{bmatrix}\theta_0 & \theta_1 & \theta_2 \cdots \theta_n\end{bmatrix} \begin{bmatrix}x_0 \\ x_1 \\ \vdots \\ x_n\end{bmatrix} = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n + \theta_0 (x_0 = 1)$$

$$\theta^T x^{(i)} + \theta_0 = \theta^T x^{(i)} + b$$

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

$$p\left(y^{(i)} = 1 \mid x^{(i)}; \theta\right) = g\left(\theta^T x^{(i)}\right)$$

↳ $g(z) = \frac{1}{1+e^{-z}}$

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

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

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

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

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

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

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

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

$\sigma'(z) = \frac{d}{dx}(x^{-1}) = -1x^{-1-1} = \frac{-1}{x^2}$

$-1(1+e^{-z})^{-2} \cdot (0+e^{-z}) \cdot (-1)$

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

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

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

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

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

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

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

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

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

$= \frac{1}{e^{z}+1}$

$= \frac{1}{1+e^{+z}}$

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

$= \sigma(-z)$

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

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

$p(y^{(i)}=0 \mid x^{(i)}; \theta) = 1 - p(y^{(i)}=1 \mid x^{(i)}; \theta)$

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

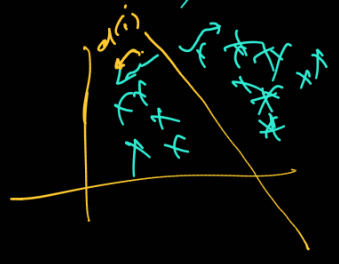
Bernoulli: $p(y=\phi)^{(y)} \cdot p(y=\phi)^{(1-y)} = \phi^{(y)}(1-\phi)^{(1-y)}$ → Fact!

Log likelihood

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

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

$p(y^{(i)}=1 \mid x^{(i)};\theta)$ $\quad$ $p(y^{(i)}=0 \mid x^{(i)};\theta)$



$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)}=1 \mid x^{(i)};\theta)^{y^{(i)}}\, p(y^{(i)}=0 \mid x^{(i)};\theta)^{(1-y^{(i)})}$$

$$L(\theta) = \prod_{i=1}^{m} g(\theta^T x^{(i)})^{y^{(i)}} \cdot \left(1-g(\theta^T x^{(i)})\right)^{(1-y^{(i)})}$$

$\Rightarrow$ HYPOTHESIS

$$L(\theta) = \prod_{i=1}^{m} h_\theta(\theta^T x^{(i)})^{y^{(i)}} \left(1-h_\theta(\theta^T x^{(i)})\right)^{(1-y^{(i)})}$$

$$Lh(\theta) = \log \prod_{i=1}^{m} h_\theta(\theta^T x^{(i)})^{y^{(i)}} \left(1-h_\theta(\theta^T x^{(i)})\right)^{(1-y^{(i)})}$$

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

(h_θ = sigmoid)

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

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

$g(f(x)) = g'(f(x)) \cdot f'(x)$

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} \frac{y^{(i)}}{\sigma(\theta^T x^{(i)})} \cdot \sigma(\theta^T x^{(i)})\left(1-\sigma(\theta^T x^{(i)})\right) x^{(i)} + \frac{(1-y^{(i)})}{\left(1-\sigma(\theta^T x^{(i)})\right)} \cdot \left(0-\sigma(\theta^T x^{(i)})\left(1-\sigma(\theta^T x^{(i)})\right)\right) x^{(i)}$$

$$\nabla_x \log f(x) = \frac{1}{f(x)} \cdot \nabla f(x)$$

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

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

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

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

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

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

Logistic Regression and Linear Regression have similar properties.

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

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

→ Generalized Class of Linear Models (GLM)

$$L(\theta) = \prod_{i=1}^{m} P(y^{(i)}=1 \mid x^{(i)};\theta)^{y^{(i)}} P(y^{(i)}=0 \mid x^{(i)};\theta)^{(1-y^{(i)})} \quad \log\left(A^y (1-\theta)^{(1-y)}\right)$$

$$LL(\theta) = \log \prod_{i=1}^{m} P(y^{(i)}=1 \mid x^{(i)};\theta)^{y^{(i)}} P(y^{(i)}=0 \mid x^{(i)};\theta)^{(1-y^{(i)})}$$

$$LL(\theta) = \sum_{i=1}^{m} \log P(y^{(i)}=1 \mid x^{(i)};\theta)^{y^{(i)}} + \sum_{i=1}^{m} \log P(y^{(i)}=0 \mid x^{(i)};\theta)^{(1-y^{(i)})}$$

$$-\sum_{i=1}^{m} y^{(i)} \log P(y^{(i)}=1 \mid x^{(i)};\theta) + (1-y^{(i)}) \log P(y^{(i)}=0 \mid x^{(i)};\theta)$$

HYPOTHESIS: $g(h_\theta(x^{(i)})) = \sigma(\theta^T x^{(i)})$

$$h_\theta(x^{(i)}) = \theta^T x^{(i)}$$
$$P(y^{(i)}=1 \mid x^{(i)};\theta) = g\left(h_\theta(x^{(i)})\right) = g(\theta^T x^{(i)})$$

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

$$\sum_{i=1}^{m} \log \sigma(\theta^T x^{(i)}) \quad \text{for } (y=1) \qquad \rightarrow 1-(y=1)$$

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

$$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} \frac{y^{(i)}}{\sigma(\theta^T x^{(i)})} \cdot \sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)})) \cdot x^{(i)} + \frac{(1-y^{(i)})}{(1-\sigma(\theta^T x^{(i)}))} \cdot \left(0 - \sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)})) \cdot x^{(i)}\right)$$

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

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

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

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

$$= \sum_{i=1}^{m} x^{(i)} \left(y^{(i)} - g(h_\theta(x^{(i)})) \right) \rightarrow \text{Logistic case}$$

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

Identity $f(x) = x$

$g(f(x)) \quad g \neq f(x)$

$g = x$