NDD 1. → rough work.

Single feature → $h_\theta(x) = \theta_0 + \theta_1 x_1$

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$. $\quad x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1}$

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$

$\{\theta_0 \theta_1 \cdots \theta_n\}$ $n+1 \times 1$

$h_\theta x = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_n x_n = \theta^T x$ $\quad 1 \times n+1$

Cost function.

$$J(\theta_0, \theta_1 \cdots \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$J(\theta)$

Concept of Gradient Decent. (one feature)

$\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$ $\quad$ say $\theta_0 = 0 \quad \theta_1 = 1$

Changin θ_0 / θ_1 to minimise $J(\theta_0, \theta_1)$

repeat until convergence.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ \& } j = 1)$$

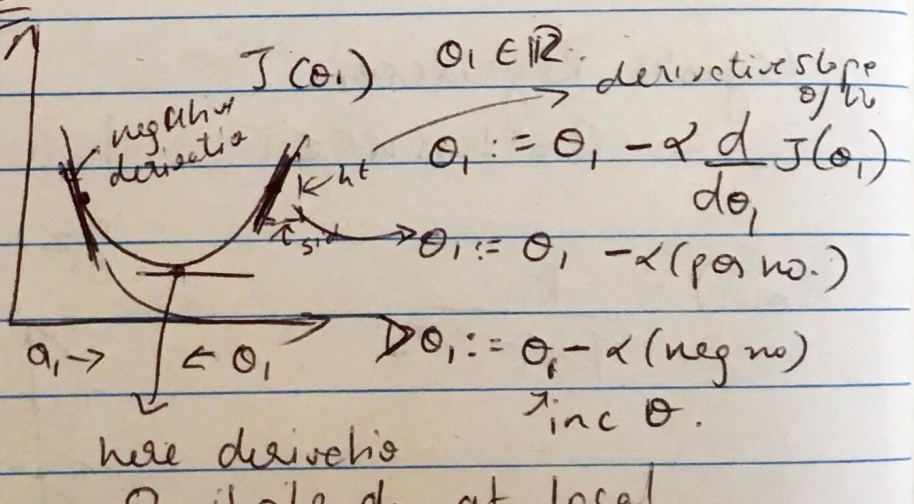
↗ Simultaneously ✱ update θ_0 & θ_1

α → learning rate

$\min_{\theta_1} J(\theta_1)$

α small → slow

α large → overshoot fail to convergl diverge.



Gradient decent for n variables. $n \geq 1$.

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

simultaneously update θ_j for all j.

Stochastic gradient decent.

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

$$J_{train}(\theta) = \frac{1}{m} \sum_{i=1}^{m} cost(\theta, (x^{(i)}, y^{(i)}))$$

1) Randomly shuffle dataset.

2) Repeat {

for $i = 1, \ldots, m$ { $\theta_j = \theta_j - \alpha \underbrace{(h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}}$

for $j = 0, \ldots n$.

$\frac{\partial}{\partial \theta_j} cost(\theta(x^{(i)}, y^{(i)}))$

1-10 times depends on size of dataset

} }

Checking for convergence. If diverging use smaller α.

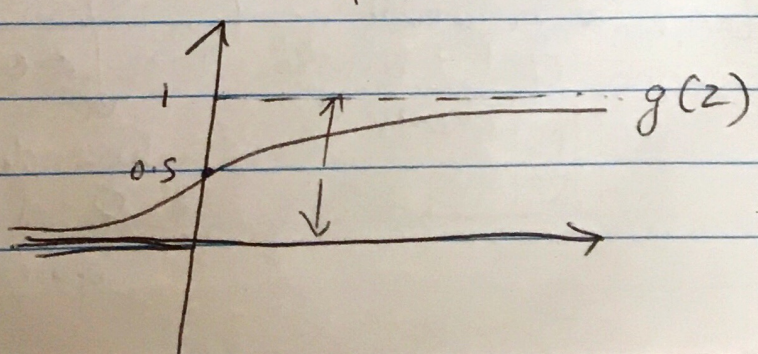
Can slowly decrease α.

Logistic Regression.

$0 \leq h_\theta(x) \leq 1$ $\quad h_\theta(x) = g(\theta^T x)$

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

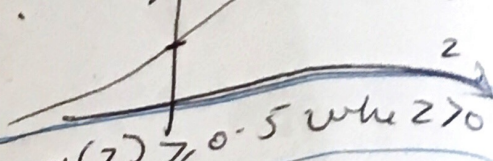
$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$



$h_\theta(x) = P(y=1 \mid x;\theta)$ 6y θ

Suppose "y=1" if $h_\theta(x) \geq 0.5 \rightarrow \theta^T x \geq 0$

"y=0" if $h_\theta(x) < 0.5 \rightarrow \theta^T x < 0$



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

$h_\theta(x) \geq 0.5 \rightarrow \theta^T x \geq 0$

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

if $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$ predict $y=1$ if $-3 + x_1 + x_2 \geq 0$

$\theta^T x$ ↓ $x_1 + x_2 \geq 3$

decision boundary where $h_\theta(x) = 0.5$

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

$y=1$ if $-1 + x_1^2 + x_2^2 \geq 0$. $\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$

$x_1^2 + x_2^2 \geq 1$ ← circle.

Cost func. ⇒ training set $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \cdots (x^{(m)}, y^{(m)})\}$

m eg. $x \in \begin{bmatrix} x_0 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1}$ $x=1$, $y \in \{0, 1\}$.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \qquad J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{cost}(h_\theta(x^{(i)}), y)$$

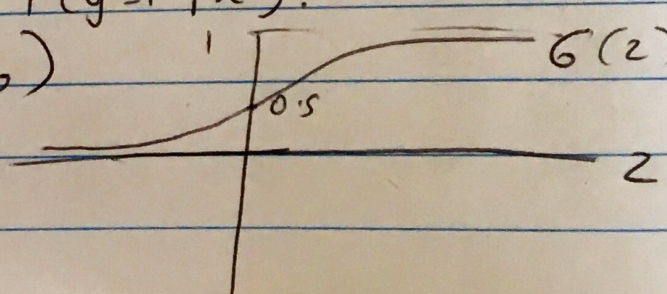
$\text{cost}(h_\theta(x), y) = \frac{1}{2}(h_\theta(x) - y)^2$ ← not convex

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

Given x, want $\hat{y} = P(y=1 \mid x)$. $x \in \mathbb{R}^{n+1}$.

op $\hat{y} = g(w^T x + b)$

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



Decision Trees.

w/ y

$H(S) = -p(+) \log_2 p(+) - p(-) \log_2 p(-)$ 6+4

$p+$ $p-$ → in the subset not whole dataset.

Information gain → avg of entropies of subset

$$-\sum_{i}^{c} p_i \log_2 (p_i)$$

$$Gain(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} H(S_v)$$

Gain ratio:

Split Entropy$(S, A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \log \frac{|S_v|}{|S|}$

or continuous.

$$Gain\ ratio = \frac{Gain(S, A)}{Split\ Entropy(S, A)}$$

↳ $-\int_{y \in Y} p(y) \log(p_y) \cdot dy$

Multiclass & regression.

$$H(S) = -\sum_{c} p(c) \log p(c)$$

Random Forest.

$$p(c|v) = \frac{1}{T} \sum_{t=1}^{T} p_t(c|v)$$

Randomization

$T_j \in T$

$$\theta_j^* = \arg\max_{\theta_j \in T_j} F_j$$