Logistic Regression :

- Borrow two big ideas from Guassion Naive Bayes:

① It is nice to model $P(y|x)$ using just the sigmoid function

$$P(y|x) = \frac{1}{1+e^{-y(w^Tx)}}, \quad y \in \{1, -1\}$$

② it is nice to view w as a parameter of $P(y|x)$, that we can fit it into data

---

- Discrimative algorithm (Naires Bayes counter part)

- Assumption

$$P(y|x) = \frac{1}{1+e^{-y(w^Tx)}}$$

w is the parameter that we need to fit/estimate

- Estimating w:

MLE estimate : $w_{MLE} = \arg\max_w P(y|x;w)$; เลือก w ที่ทำให้ $P(y|x)$ สูงสุด

$$= \arg\max_w \prod_{i=1}^{n} P(y_i|x_i;w)$$

$$w_{MLE} = \arg\max_w \prod_{i=1}^{n} \frac{1}{1+e^{-y_i(w^Tx_i)}}$$

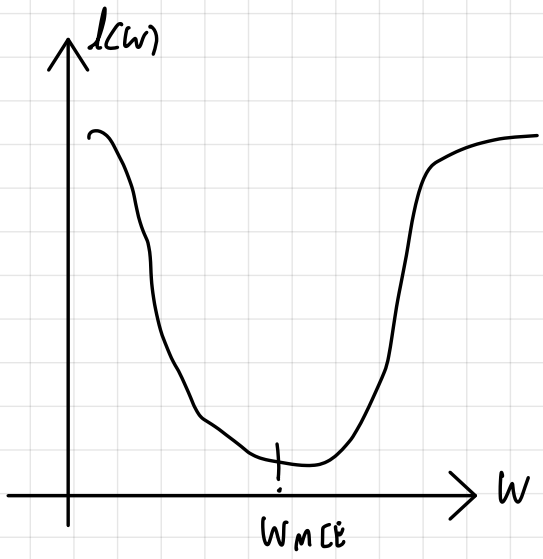
$$= \arg\max_w \sum_{i=1}^{n} \log_e\left(\frac{1}{1+e^{-y_i(w^Tx_i)}}\right)$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} \log_e 1 - \log_e(1+e^{-y_i(w^T x_i)})$$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \log_e(1+e^{-y_i(w^T x_i)}) \rightarrow \text{negative log liklihood } l(w)$$

Note - There is no closed-form solution to W_{MLE}
- We will use "Gradient Descent" to find W_{MLE}, that is, find w to minimize

Negative log likelihood $l(w) = \sum_{i=1}^{n} \log_e(1+e^{-y_i(w^T x_i)})$

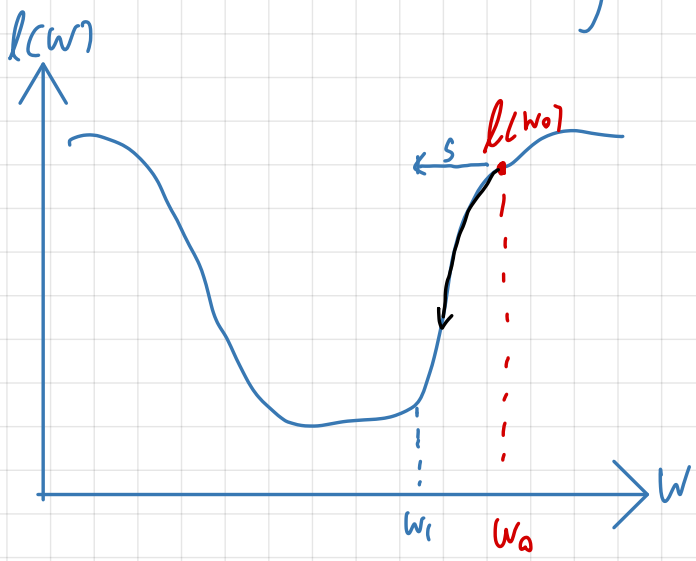


fact: l(w) is a continous, convex, and differentiable function
แปลว่าต่อเนื่องไม่มีจุดหัก

การหาจุดต่ำสุดของ function ที่มีสภาวะ 3 สภาวะ: convex, differentiable, continous ***

Gradient Descent:

- Based on "Hill-climbling" scheme



Hill-climbing
1) start off by some point w_0
2) Repeat until convergence:

$$w_{t+1} = w_t + s$$

if $\| w_{t+1} - w_t \|_2 < \epsilon (0.1 \ldots)$, convergence!

The problem is how to define s?

- In Gradient Descent, we set

$$S = -\alpha \nabla l(w)$$ ***

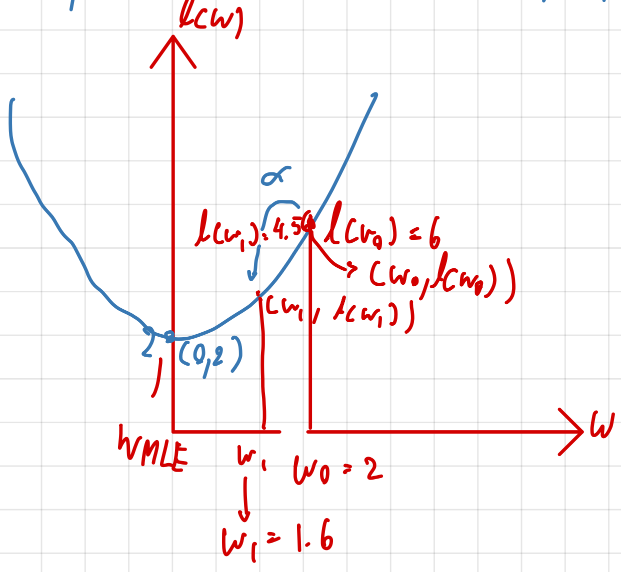
(α → learning rate)

$\nabla l(w) \approx \frac{dl}{dw}$ ← dimension เดียว (ตัวแปรเดียว)

↑ gradient (หลาย dimension)

Setting

let say $l(w) = w^2 + 2$; $y = x^2 + 2$



$w_0 = 2 \Rightarrow l(2) = 6$

$S = -\alpha \frac{dl(w)}{dw} = -\alpha(2w)$

$= -2\alpha w$

$= -4\alpha$

set $\alpha = 0.1 \Rightarrow S = -0.4$

$\alpha > 0$ is a scalar called "learning rate"

update

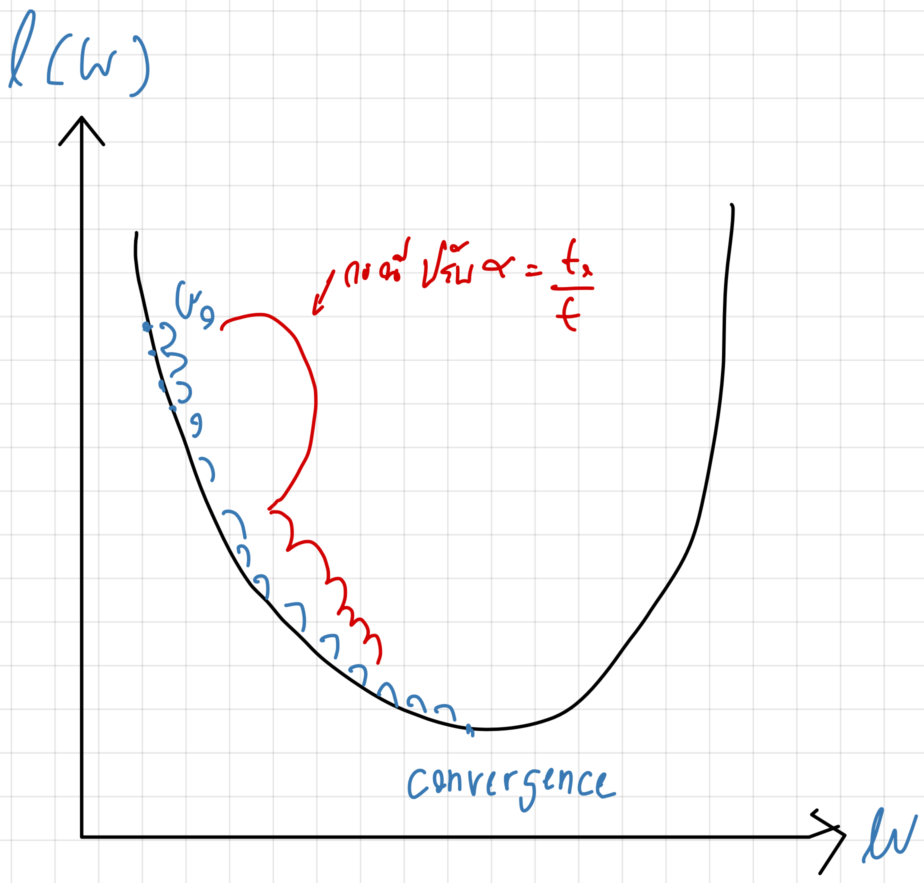
$w_1 = w_0 + S = 2 + (-0.4) = 1.6$

$l(w_1) = (1.6)^2 + 2 = 4.56$

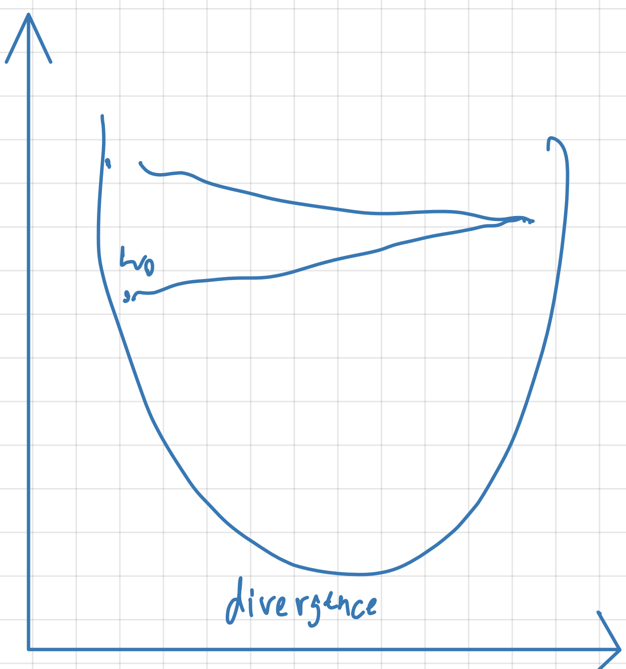
$\nabla l(w) = \frac{dl}{dw}$ → เป็นค่าที่บอกตำแหน่งที่เลย WMLE (ความชันต่ำสุด) หรือเข้าใกล้ WMLE

→ พยายามทำให้ $l(w_1) \approx$ WMLE

α : เป็น step หรือความยาวในการข้ามไปในแต่ละ step ถ้าให้ค่ามากสามารถข้าม WMLE ได้ ถ้าน้อยเกินจะใช้เวลานานในการหา WMLE



- If α is too small, the algorithm will find W_{MLE} very slowly.



- If α is too large, the algorithm will never find W_{MLE}

Best pratice: A safte choice is to set $\alpha = \frac{t_0}{t}$, which guarantees that it will eventually become small enough to converge (for any $t_0 > 0$)

(t → จน.ครั้งที่ update w)

Prove guarantess update S

Talyor's Expansion:

- if $\|\vec{s}\|_2$ is small (meaning w+s is close to w) then the following holds:

$$\ell(w+s) \approx \ell(w) + \nabla \ell(w)^T s$$

vector x vector = scalar

$$\nabla \ell(w) = \begin{bmatrix} \frac{\partial \ell}{\partial w_1} \\ \vdots \\ \partial \ell / \partial w_d \end{bmatrix} \quad w = \mathbb{R}^d$$

- Gradient Descent's update $s = -\alpha \nabla \ell(w)$, where $\alpha > 0$

- Therefore, we can derive

$$\ell(w+s) \approx \ell(w) + \nabla \ell(w)^T(-\alpha \nabla \ell(w))$$

$$\approx \ell(w) + (-\alpha \nabla \ell(w)^T \nabla \ell(w))$$

$$\approx \ell(w) - \underbrace{\alpha}_{>0} \underbrace{\|\nabla \ell(w)\|_2}_{>0} < \ell(w)$$

∴ เมื่อ $\ell(w) - s$ ย่อมจะน้อยกว่า $\ell(w)$ เนื่องจาก ค่า s มากกว่า 0 ดังนั้นเมื่อมีการ update โดยใช้ s จะเป็นการลดค่า $\ell(w)$ แน่นอน

What's to pick up?

- logistic loss function $\ell(w) = \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i})$ also measures how well the Classifier perfroms on the data (ideally, we wish for $\ell(w)$ to be smallest possible)

- The gradient Descent is an algorithm that finds an optimal point of serval loss functions.

P(y|x) P(x|y)

- logistic regression is a discrimative counterpart is Naive Bayes

Naive Bayes with logistic Regression:

- with little data and if the modeling distribution is right, Naive Bayes tends to beat logistic Regression.

- As data becomes larger, logistic Regression will outperfrom Naive Bayes which suffers from the fact that the modeling assumption may not be the right one

- ข้อได้เปรียบของ Naive Bayes คือการจำลอง distribution ขึ้นดังนั้น data น้อย generate ได้ ถ้า data มาก มัน อธิบาย distribution ในตัว ดังนั้น modeling distribution นั้นไม่จำเป็น