* 建模基本流程

1. $\hat{y} = f(w, x)$ → θ

2. $C = C(\hat{y}, y)$ → MLE view : cross entropy、MSE

3. $w^* = \underset{w}{\arg\min}\, C$ → Difficult when it's deep

⇓ optimization method

⇓ Gradient descent

⇓ $w_{t+1} = w_t - LR\, \nabla_w C$

⇓

$$x \longrightarrow f(w^*, x) \xrightarrow{\text{predict}} \hat{y}$$

* 梯度下降法

<補充>
2D 示意圖

國中： 斜率 $= \dfrac{\Delta y}{\Delta x}$

高中： 一階導數 $= \lim\limits_{\Delta x \to 0} \dfrac{f(x+\Delta x) - f(x)}{\Delta x}$

現在： $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$ $\quad x \in R^n$

$w_{t+1} = w_t - LR\, \nabla_w C$?

$$\phi = \lim_{h \to 0} \frac{C(w_t + h \cdot u) - C(w_t)}{h}$$

$u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = u_1 e_1 + u_2 e_2 + \dots$

$\|u\| = 1$

by Taylor ^{1st} at w_t

$$C(w_t + hu) \approx C(w_t) + \left.\frac{\partial C}{\partial w_{t1}}\right|_{w=w_t} \cdot (w_{t1} + hu_1 - w_{t1}) + \dots$$

$$+ \left.\frac{\partial C}{\partial w_{tn}}\right|_{w=w_t} \cdot (w_{tn} + hu_n - w_{tn})$$

$$= C(w_t) + h \cdot \nabla C(w_t)^T \cdot u$$

代入 ϕ

$$\lim_{h\to 0}\frac{C(w_t+h\cdot u)-C(w_t)}{h}=\lim_{h\to 0}\frac{h\cdot\nabla C(w_t)^T\cdot u}{h}=\nabla C(w_t)^T u$$

$$\Rightarrow \quad \phi=\|\nabla C(w_t)\|\cdot\cos\theta$$

$$\Rightarrow \quad u^*=\operatorname*{argmin}_u \phi=\phi(\theta=\pi)$$

ex : Regression

1. $\hat{y}=Xw+b$
2. $C=\|\hat{y}-y\|^2$
3. $\nabla_w C$ 、 $\nabla_b C = ?$

$$w=\begin{bmatrix} w_1\\ \vdots\\ w_d\end{bmatrix}_{d\times 1} \qquad X=\begin{array}{c} 1\ 2\ \cdots\ d \\ \begin{bmatrix} \text{———} \\ \text{———} \\ \text{———} \\ \text{———} \\ \vdots \end{bmatrix}_{N\cdot d}\end{array}$$

$$b=\begin{bmatrix} b_1\\ \vdots\\ b_N\end{bmatrix}_{N\times 1} \qquad y=\begin{bmatrix} y_1\\ \vdots\\ y_N\end{bmatrix}_{N\times 1}$$

$$\begin{aligned}
C&=(\hat{y}-y)^T(\hat{y}-y)\\
&=(Xw+b-y)^T\cdot(Xw+b-y)\\
&=((Xw)^TXw+(Xw)^Tb-(Xw)^Ty+b^TXw+b^Tb-b^Ty\\
&\quad -y^TXw-y^Tb+y^Ty)
\end{aligned}$$

$$\begin{aligned}
\nabla_w C&=(2X^TXw+X^Tb-X^Ty+X^Tb-X^Ty)\\
&=2X^T(Xw+b-y)=2X^T(\hat{y}-y)
\end{aligned}$$

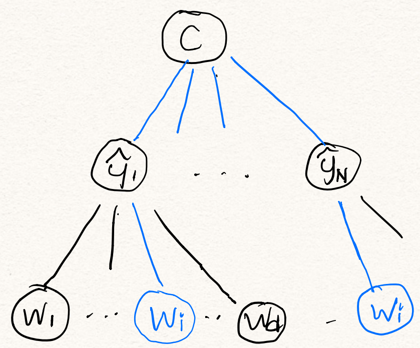
<補充>

1. Normal Equation?
2. 幾何意義?

$$\nabla_b C = 2(\hat{y} - y)$$

方法 2：

Chain Rule



$$\frac{\partial C}{\partial w_i} = \sum_{j}^{N} \frac{\partial C}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial w_i}$$

$$= \sum_{j}^{N} 2(\hat{y}_j - y_j) \cdot x_{ji}$$

$$= 2 \cdot (\hat{y} - y) \; ? \; X$$

$$\Downarrow$$

$$\nabla_W C = 2X^T(\hat{y} - y)$$

$$\nabla_b C = 2(\hat{y} - y)$$

$$\hat{y}_j = \sum_{k}^{d} w_k x_{jk} + b_j$$

* 常見最佳化方法

(1) SGD：

$$W_{t+1} = W_t - LR \cdot \nabla_W C$$

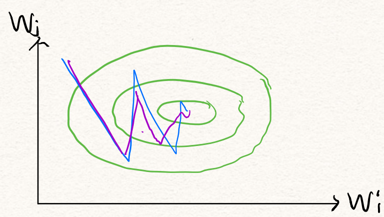
<補充>
weight decay

(2) Momentum：

$$V_{t+1} = \rho V_t + G_{t+1}$$

$$W_{t+1} = W_t - LR \cdot (\rho V_t + G_{t+1})$$

$$V_0 = 0$$



(3) AdaGrad :

$$W_{t+1} = W_t - LR \cdot G_{t+1} \odot \frac{1}{\sqrt{r_{t+1}} + \varepsilon}$$

$$\left(r_{t+1} = \sum_{\tau}^{t+1} G_\tau \odot G_\tau\right)$$

elementwise operation

- LR 逐漸調小，某方向太快讓它緩緩
  有時是缺點，收斂太快

(4) RMSprop :

- AdaGrad + Momentum Concept on r

$$W_{t+1} = W_t - LR \cdot G_{t+1} \odot \frac{1}{\sqrt{r_{t+1}} + \varepsilon}$$

$$\left(r_{t+1} = \rho r_t + (1-\rho) G_{t+1} \odot G_{t+1}\right)$$

0.99 or 0.9

(5) Adam :

- First moment

$$\hat{v}_{t+1} = \frac{\rho_1 v_t + (1-\rho_1) G_{t+1}}{(1-\rho_1^{t+1})}$$

- Second moment

$$\hat{r}_{t+1} = \frac{\rho_2 r_t + (1-\rho_2) G_{t+1} \odot G_{t+1}}{(1-\rho_2^{t+1})}$$

$$\Rightarrow W_{t+1} = W_t - LR \cdot \frac{\hat{v}_{t+1}}{\sqrt{\hat{r}_{t+1}} + \varepsilon}$$

<補充>

為什麼除以 $(1-\rho^{t+1})$?

$$v_{t+1} = \rho_1(\rho_1 v_{t-1} + \cdots) + (1-\rho_1) G_{t+1}$$

$$= (1-\rho_1) \sum_{\tau} G_\tau \rho^{t-\tau}$$

$$E(v_{t+1}) = (1-\rho_1^{t+1}) E(G_t)$$

$$E(G_t) = \frac{E(v_{t+1})}{1-\rho_1^{t+1}}$$