3 梯度下降法

* 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. $w^* = \underset{w}{\operatorname{argmin}} C$ → 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

國中：$斜率 = \frac{\Delta y}{\Delta x}$ <補充> 2D 未老圖

高中：一階導數 $= \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$

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

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

$$\phi = \lim_{h \to 0} \frac{C(w_t + h \cdot u) - C(w_t)}{h} \qquad u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = u_1 e_1 + u_2 e_2 + \cdots$$

by Taylor 1st at w_t $\qquad \|u\| = 1$

$$C(w_t + hu) \approx C(w_t) + \frac{\partial C}{\partial w_{t1}}\Big|_{w=w_t} \cdot (w_{t1} + hu_1 - w_{t1}) + \cdots$$

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

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

代入 ϕ

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

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

$$\Rightarrow \quad u^* = \arg\min_u \phi = \phi(\theta = \pi)$$

§ Regression 為例

1. $\hat{y} = Xw + b$

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

2. $C = \|\hat{y} - y\|^2$

3. $\nabla_w C$ 、 $\nabla_b C = ?$

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

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

$$= (Xw + b - y)^T \cdot (Xw + b - y)$$

$$= ((Xw)^T Xw + (Xw)^T b - (Xw)^T y + b^T Xw + b^T b - b^T y$$

$$- y^T Xw - y^T b + y^T y)$$

$$\nabla_w C = (2X^T Xw + X^T b - X^T y + X^T b - X^T y)$$

$$= 2X^T(Xw + b - y) = 2X^T(\hat{y} - y)$$

<補充>

1. Normal Equation ?

2. 幾何意義 ?

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

方法 2：

Chain Rule

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

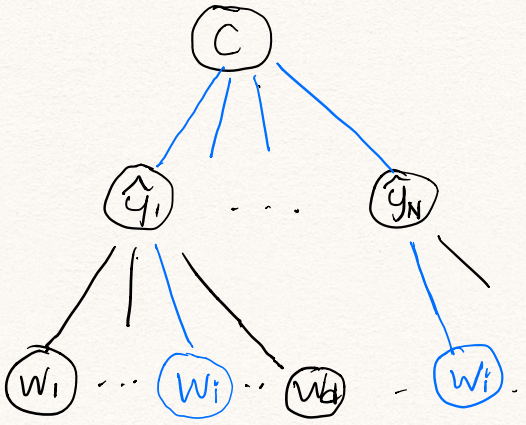
$$= \sum_j^N 2(\hat{y}_j - y_j) \cdot x_{ji}$$

$$= 2 \cdot (\hat{y} - y) \textcircled{?} x$$

⇓

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

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



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

* $C = \|\hat{y} - y\|^2$ 怎麼來的？

$$y^{(i)} = X^{(i)} W + \varepsilon^{(i)}$$ → 常態

$$P(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right)$$

$$P(y^{(i)} | x^{(i)}; W) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - X^{(i)} W)^2}{2\sigma^2}\right)$$

$$L(w) = \prod_{i=1}^{N} P(y^{(i)} \mid x^{(i)}; w)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(y^{(i)} - x^{(i)}w)^2}{2\sigma^2}\right)$$

$$\ell(w) = \log L(w)$$

$$= \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(y^{(i)} - x^{(i)}w)^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(y^{(i)} - x^{(i)}w)^2}{2\sigma^2}\right)$$

$$= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{N} (y^{(i)} - x^{(i)}w)^2$$

$$C := \frac{1}{2} \sum_{i=1}^{N} (y^{(i)} - x^{(i)}w)^2$$

$$= \frac{1}{2} \|\hat{y} - y\|^2$$

* 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:

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

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

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

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

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

ρ: 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 $(1-\rho^{t+1})$?

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

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

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

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

(6) Second-Order Optimization

$$J(\theta) \approx J(\theta_0) + (\theta-\theta_0)^T \nabla_\theta J(\theta_0) + \frac{1}{2}(\theta-\theta_0)^T H (\theta-\theta_0)$$

$$\nabla_\theta J(\theta_0) + H(\theta-\theta_0) = 0$$

$$\Rightarrow \theta^* = \theta_0 - H^{-1} \nabla_\theta J(\theta_0)$$