Maximum Likelihood Estimation (MLE)

$\mathcal{X}$: training data set

θ : parameters

$$\theta^*_{MLE} = \arg\max_{\theta} \underline{p(\mathcal{X}|\theta)}$$

$$p(\mathcal{X}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

Maximum a Posteriori Estimation (MAP)

$$\theta^*_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{X})$$

$$= \arg\max_{\theta} \frac{\underline{p(\mathcal{X}|\theta)}\, \overbrace{p(\theta)}}{p(\mathcal{X})}$$

our prior belief about θ

Parametric Regression:

$$\underbrace{y}_{\text{observations}} = \underbrace{f(x)}_{\text{underlying process}} + \underbrace{\epsilon}_{\text{noise}}$$

Learning problem:

approximate $f(x)$ with $g(x|\underbrace{\theta}_{\text{parameters}})$

$$\boxed{\begin{aligned} E[X] &= \gamma \\ E[X+\underset{\text{constant}}{c}] &= \gamma + c \\ VAR[X] &= \kappa^2 \\ VAR[X+c] &= \kappa^2 \end{aligned}}$$

I $\quad p(\epsilon) \sim N(\epsilon; 0, \sigma^2)$

II $\quad p(y|x) \sim N(y; g(x|\theta), \sigma^2)$

$$y = f(x) + \epsilon$$

$$\Big\downarrow \text{approximation}$$

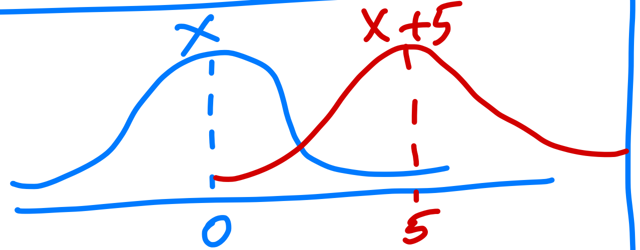
$$y = \underbrace{g(x|\theta)}_{\text{constant}} + \underbrace{\epsilon}_{\text{random variable}}$$

$$\begin{aligned} E[y|x] &= E[g(x|\theta) + \epsilon] \\ &= g(x|\theta) + \underbrace{E[\epsilon]}_{0} \\ &= g(x|\theta) \end{aligned}$$

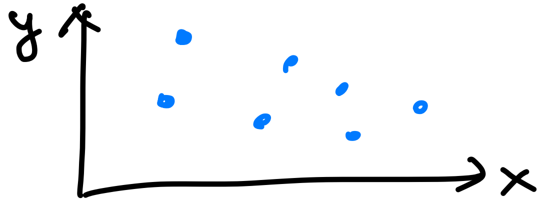
$$\begin{aligned} VAR[y|x] &= VAR[g(x|\theta) + \epsilon] \\ &= 0 + \underbrace{VAR[\epsilon]}_{\sigma^2} \\ &= \sigma^2 \end{aligned}$$

$X \sim N(x; 0, 9)$

$X+5 \sim N(x+5; 5, 9)$



$$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N} \quad x_i \in \mathbb{R}^1 \quad y_i \in \mathbb{R}^1$$



$$\underbrace{(x_i, y_i)}_{\text{i.i.d.}} \sim p(x_i, y_i)$$

$$\boxed{\begin{aligned} p(x,y) &= p(x|y)\cdot p(y) \\ p(x,y) &= p(y|x)\,p(x) \end{aligned}}$$

$$p(x_1, y_1, x_2, y_2, \dots, x_N, y_N) = \prod_{i=1}^{N} p(x_i, y_i)$$

Likelihood $\Rightarrow$

$$L(\theta|\mathcal{X}) = \prod_{i=1}^{N} p(x_i, y_i)$$

$$= \prod_{i=1}^{N} \left[p(y_i|x_i)\, p(x_i) \right]$$

$$\boxed{\log(a.b) = \log(a) + \log(b)}$$

log likelihood $\Rightarrow$

$$= \log\left[\prod_{i=1}^{N} \left[p(y_i|x_i)\, p(x_i) \right] \right]$$

$$= \sum_{i=1}^{N} \left[\log[p(y_i|x_i)] + \cancelto{\text{constant}}{\log[p(x_i)]} \right]$$

$$= + \sum_{i=1}^{N} \log[p(y_i|x_i)]$$

$$\text{maximize} \sum_{i=1}^{N} \log[\underbrace{p(y_i|x_i)}_{N(y_i;\, g(x_i|\theta),\, \sigma^2)}] \Rightarrow \text{maximize} \sum_{i=1}^{N} \log\left[N(\overset{\text{R.V.}}{y_i};\, \overset{\text{mean}}{g(x_i|\theta)},\, \overset{\text{variance}}{\sigma^2}) \right]$$

$$\text{maximize} \sum_{i=1}^{N} \underbrace{\log}_{\times}\left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \underbrace{\exp}_{\times}\left[-\frac{(y_i-g(x_i|\theta))^2}{2\sigma^2}\right]\right]$$

constant

$$\text{maximize} \sum_{i=1}^{N}\left[-\frac{[y_i-g(x_i|\theta)]^2}{2\sigma^2}\right]$$

constant

$$\text{maximize} \sum_{i=1}^{N} -[y_i-g(x_i|\theta)]^2$$

$$\text{minimize} \sum_{i=1}^{N} [y_i-\underbrace{g(x_i|\theta)}_{\hat{y}_i}]^2$$

$$\text{minimize} \sum_{i=1}^{N} (\underbrace{y_i-\hat{y}_i}_{e_i})^2 = \sum_{i=1}^{N} e_i^2$$

$$g(x_i|\theta) = w_0 + w_1 . x_i$$

$$\theta = \{w_0, w_1\}$$

$$g(x_i|\theta) = w_0 + w_1 . x_i + w_2 x_i^2$$

$$\theta = \{w_0, w_1, w_2\}$$

minimize $\sum_{i=1}^{N} \left[y_i - g(x_i|\theta) \right]^2$

$g(x_i|\theta) = w_0 + w_1 x_i$

$\theta^* = \{ w_0^*, w_1^* \} = ?$

$$\text{Error}[\theta|\mathcal{X}] = \sum_{i=1}^{N} \left[y_i - (w_0 + w_1 x_i) \right]^2$$

$$\frac{\partial \text{Error}[\theta|\mathcal{X}]}{\partial w_0} = \sum_{i=1}^{N} \cancel{2} \left[y_i - (\underline{w_0} + w_1 x_i) \right] \cdot (\cancel{-}1) = 0$$

$$\frac{\partial \text{Error}[\theta|\mathcal{X}]}{\partial w_1} = \sum_{i=1}^{N} \cancel{2} \left[y_i - (w_0 + w_1 x_i) \right] \cdot (\cancel{-}\underline{x_i}) = 0$$

$$\sum_{i=1}^{N} y_i = \underbrace{\sum_{i=1}^{N} w_0}_{N w_0} + \sum_{i=1}^{N} w_1 \cdot x_i$$

$$\sum_{i=1}^{N} (y_i \cdot x_i) = \sum_{i=1}^{N} w_0 x_i + \sum_{i=1}^{N} w_1 x_i^2$$

$A^{-1} A \theta = A^{-1} b$

$$\boxed{\theta = A^{-1} \cdot b}$$

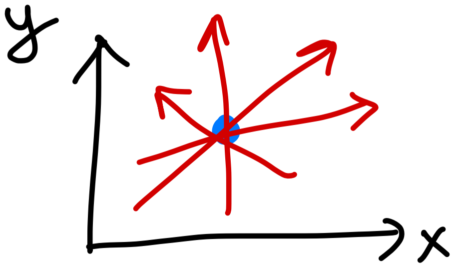
$$\overbrace{\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}}^{A} \overbrace{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}^{\theta} = \overbrace{\begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} (y_i \cdot x_i) \end{bmatrix}}^{b}$$

Exercise: Show this 2x2 matrix is invertible if $N \geq 2$.

$$\theta^* = \begin{bmatrix} w_0^* \\ w_1^* \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} (y_i \cdot x_i) \end{bmatrix}$$

this matrix is invertible if $N \geq 2$.

$\Rightarrow$ If there is a single data point $(N=1)$

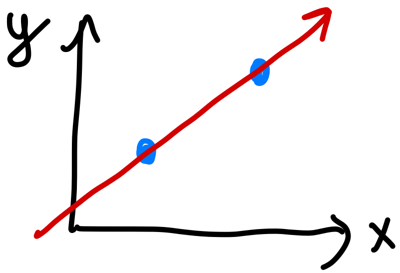


$$A = \begin{bmatrix} 1 & x_1 \\ x_1 & x_1^2 \end{bmatrix} \Rightarrow \det(A) = 1 \cdot x_1^2 - x_1 \cdot x_1 = 0$$

$\Downarrow$

not invertible

$\Rightarrow$ If there are two data points $(N=2)$



$$A = \begin{bmatrix} 2 & (x_1+x_2) \\ (x_1+x_2) & x_1^2+x_2^2 \end{bmatrix} \Rightarrow \det(A) = 2x_1^2 + 2x_2^2 - x_1^2 - x_2^2 - 2x_1x_2 = (x_1-x_2)^2 > 0$$

$\Downarrow$

invertible

Polynomial Regression:

$\boxed{w_0 = w_0 . x_i^0}$

$$g(x_i \mid w_0, w_1, w_2, \ldots w_k) = \underbrace{w_0 + w_1 . x_i^1 + w_2 x_i^2 + \ldots + w_k x_i^k}_{k^{th} \text{ order polynomiel}}$$

$k=1 \Rightarrow$ lineer regression

$$\underbrace{\begin{bmatrix} N & \sum_{i=1}^{N} x_i & \cdots & \sum_{i=1}^{N} x_i^k \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 & \cdots & \sum_{i=1}^{N} x_i^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_i^k & \sum_{i=1}^{N} x_i^{k+1} & \cdots & \sum_{i=1}^{N} x_i^{2k} \end{bmatrix}}_{A \,=\, D^T . D} \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{bmatrix}}_{\theta} = \underbrace{\begin{bmatrix} \sum_{i=1}^{N} (y_i . x_i^0) \\ \sum_{i=1}^{N} (y_i . x_i^1) \\ \vdots \\ \sum_{i=1}^{N} (y_i . x_i^k) \end{bmatrix}}_{b}$$

$$\theta = A^{-1} . b$$

A is invertible if $N \geq k+1$.

$K=0 \Rightarrow$ constant fit

$$N \cdot w_0 = \sum_{i=1}^{N} y_i \Rightarrow \boxed{w_0^* = \frac{\sum_{i=1}^{N} y_i}{N}} \ \bar{y}$$

$\hookrightarrow$ sample mean

$$D = \begin{bmatrix} 1 & x_1 & \cdots & x_1^K \\ 1 & x_2 & \cdots & x_2^K \\ \vdots & \vdots & \ddots & \\ 1 & x_N & \cdots & x_N^K \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ \vdots & & \ddots & \vdots \\ x_1^K & x_2^K & \cdots & x_N^K \end{bmatrix}}_{D^T} \cdot \underbrace{\begin{bmatrix} 1 & x_1 & \cdots & x_1^K \\ 1 & x_2 & \cdots & x_2^K \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^K \end{bmatrix}}_{D} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i & \cdots & \sum_{i=1}^{N} x_i^K \\ & & & \\ & & & \end{bmatrix}$$

$$D^T \cdot D = A$$

if $N < K+1$, $D^T D$ is not invertible.

if $N \geq K+1$, $D^T \cdot D$ is invertible.