Maximum Likelihood Estimation (MLE)

$\mathcal{X}$: training data set

θ : parameters

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

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

Maximum a Posteriori Estimation (MAP)

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

$$= \arg\max_{\theta} \frac{\boxed{p(\mathcal{X} \mid \theta)\, \overbrace{p(\theta)}^{\text{our prior belief about } \theta}}}{p(\mathcal{X})}$$

Parametric Regression:

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

Learning problem

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

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

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

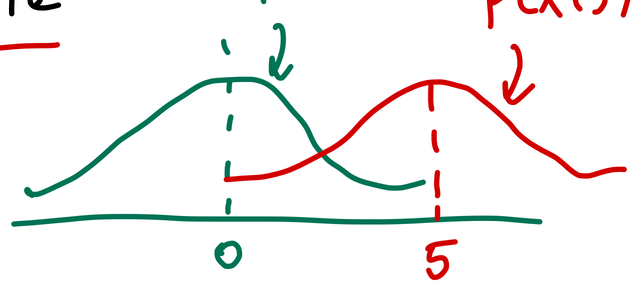
II $\quad p(y|x) \sim N(y;\ g(x|\theta), \sigma^2)$

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

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

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

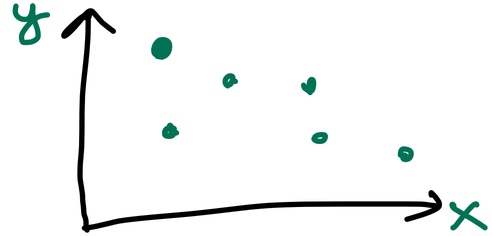
$$X+5 \sim N(x+5;\ 5, 9)$$



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

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

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



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

$$p(x,y) = p(x|y)\,p(y)$$

$$p(x,y) = \underline{p(y|x).p(x)}$$

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

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

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

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

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

maximize $\sum_{i=1}^{N} \underbrace{\log p(y_i|x_i)}_{\log\left[\mathcal{N}(y_i;\, g(x_i|\theta),\, \sigma^2)\right]}$

$\log(a.b) = \log(a) + \log(b)$

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

constant

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

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

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

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

$$g(x_i|\theta) = w_0 + w_1 x_i \Rightarrow \theta = \{w_0, w_1\}$$

$$g(x_i|\theta) = w_0 + w_1 x_i + w_2 x_i^2 \Rightarrow \theta = \{w_0, w_1, w_2\}$$

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

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

$\Rightarrow$ Find w_0^*, w_1^*

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

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

$$\sum_{i=1}^{N} y_i = N \boxed{w_0} + \left(\sum_{i=1}^{N} x_i\right) \boxed{w_1}$$

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

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

EXERCISE #2: show that this matrix is invertible.

$$\underbrace{\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}}_{A_{2\times2}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}_{\theta} = \underbrace{\begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} (y_i x_i) \end{bmatrix}}_{b}$$

$$A \cdot \theta = b$$

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

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

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

Polynomial Regression!

$$g(x_i \mid w_0, w_1, \ldots, w_k) = \underbrace{w_0}_{w_0 \cdot x_i^0} + \underbrace{w_1 x_i}_{w_1 \cdot x_i^1} + \cdots + w_k \cdot x_i^k$$

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

zeroth order polynomial $\Rightarrow$ $N \cdot w_0 = \sum_{i=1}^{N} y_i$

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

$$D = \begin{bmatrix} 1 & x_1 & \cdots & x_1^K \\ 1 & x_2 & \cdots & x_2^K \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^K \end{bmatrix}_{N \times (K+1)}$$

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

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

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

\# of data points

\# of parameters.

$$A\theta = b$$

$$(D^T.D).\theta = D^T.y$$

$$\boxed{\theta^* = (D^T.D)^{-1}.D^T.y}$$