MLE principle   Data $x \sim p(x|w)$ → prob. distrib.

$$\hat{w} = \underbrace{\arg\max_{w} p(x|w)}_{\text{the likelihood } L(w)}$$

Coin toss experiment: $x_n \sim B(w)$ $p(x_n = H|w) = w$

$p(x_n = T|w) = 1 - w$

Linear regression $t_n \sim ?$

$N(x|\mu, \sigma^2)$ → univariate Gaussian

$$N(x|\mu,\sigma^2) \Rightarrow p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{if } \mu = 0 \Rightarrow p(x|0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

$$\beta = \frac{1}{\sigma^2} \Rightarrow \beta^{-1} = \sigma^2$$

precision   variance

$$\underline{p(x|0,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}x^2}}$$

$$N(x|\mu,\Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}}$$

$x = [x_1, x_2, \dots, x_N]$

$\Sigma \in R_{N \times N}$ covariance matrix

Σ_{ij}

$$N(w|0,\alpha^{-1}I) = \frac{1}{\sqrt{\det(\frac{2\pi}{\alpha} I)}} \underline{e^{-\frac{w^T \alpha^{+1} I w}{2}}}$$

$$e^{-\frac{w^T w}{2\alpha^{-1}}}$$

Probabilistic account of Linear Regression:

Assume label t is generated as $t = w^T x + \varepsilon$,

where noise $\varepsilon \sim \mathcal{N}(0, \beta^{-1}) \Rightarrow \cancel{p}\, p(\varepsilon) = \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}\varepsilon^2}$

$t \sim \mathcal{N}(w^T x, \beta^{-1}) \Rightarrow p(t) = \frac{1}{\sqrt{2\pi\beta^{-1}}} \cdot e^{-\frac{\beta}{2}(t - w^T x)^2}$

Training data $\mathcal{D} = \{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\}$ i.i.d

$p(t_n | x_n, w, \beta) \equiv \underline{p(t_n | w) = \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}(t_n - w^T x_n)^2}}$

Label vector $\vec{t} = [t_1, t_2, \ldots, t_n]$

The likelihood $p(\vec{t} | w) = \prod_{n=1}^{N} p(t_n | w)$

Use MLE to find $\hat{w} = \underset{w}{\arg\max}\; p(t|w)$

$= \underset{w}{\arg\max}\; \ln p(t|w)$

$\sum_{n=1}^{N} \ln p(t_n | w) = \sum_{n=1}^{N} \left(\underbrace{\ln \sqrt{\frac{\beta}{2\pi}}}_{\text{const. w.r.t. } w} - \frac{\beta}{2}(t_n - w^T x_n)^2 \right)$

$\hat{w} = \underset{w}{\arg\max} \sum_{n=1}^{N} -\frac{\beta}{2}(t_n - w^T x_n)^2 = \underset{w}{\arg\min}\; \frac{1}{2N} \sum_{n=1}^{N} (t_n - w^T x_n)^2$

$$\hat{w} = \underset{w}{\arg\min}\ \frac{1}{2N} \underbrace{\sum_{n=1}^{N} (t_n - w^T x_n)^2}_{J(w)} \qquad x_n, w \in \mathbb{R}_{K \times 1}$$

$$\frac{\partial J}{\partial w} = 0 \Rightarrow \sum_{n=1}^{N} 2(t_n - w^T x_n)(-x_n) = 0 \quad \Big|\ {}^T$$

$$\sum_{n=1}^{N} (t_n - w^T x_n)\, x_n^T = 0$$

$$X = \begin{bmatrix} - - x_1^T - \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$$

$$w^T \underbrace{\sum_{n=1}^{N} x_n x_n^T}_{X^T X} = \underbrace{\sum_{n=1}^{N} t_n x_n^T}_{t^T X}$$

$$t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

$$w^T (X^T X) = t^T X \quad \Big|\ {}^T$$

$$\cancel{w}\,(X^T X)^T w = X^T t$$

$$(X^T X)\, w = X^T t \Rightarrow w = (X^T X)^{-1} X^T t$$

Moore-Penrose inverse of X.

Bayes rule $\quad \underbrace{p(w|t)}_{\text{posterior}} = \dfrac{p(t|w)\cdot p(w)}{p(\cancel{w})}$ prior likelihood

$p(w) = \mathcal{N}(w|\vec{0}, \alpha^{-1}I)$ (MAP)

Maximum A-Posteriori principle : $\hat{w} = \arg\max_{w} p(w|t)$

$$\hat{w} = \arg\max_{w} p(t|w)\,p(w) = \arg\max_{w} \ln p(t|w) + \ln p(w)$$

$$= \arg\max_{w} \sum_{n=1}^{N} -\frac{\beta}{2}(t_n - w^T x_n)^2 + \left(-\frac{w^T w}{2\alpha^{-1}}\right) \qquad /N$$

$$= \arg\min_{w} \left[\frac{1}{2N}\sum_{n=1}^{N} \beta (t_n - w^T x_n)^2\right] + \frac{1}{\frac{1}{\alpha}N}\frac{w^T w}{2} \qquad /\beta$$

$$= \arg\min_{w} \left[\frac{1}{2N}\sum_{n=1}^{N} (t_n - w^T x_n)^2\right] + \left[\frac{\overbrace{\alpha/\beta N}^{\lambda}}{2}\|w\|^2\right]$$