

Linear Algebra

X : vector of random variable X_1, X_2, \dots, X_n

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

C_X : covariance matrix of X

$$C_X = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_2, X_1) & \cdots & \text{Cov}(X_n, X_1) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) & \cdots & \text{Cov}(X_n, X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_1, X_n) & \text{Cov}(X_2, X_n) & \cdots & \text{Var}(X_n) \end{bmatrix}$$
$$= E[(X - E(X))(X - E(X))^T]$$

$$Y = AX + b$$

$$\begin{aligned} C_Y &= E[(Y - E(Y))(Y - E(Y))^T] \\ &= E[(AX + b - E(AX + b))(AX + b - E(AX + b))^T] \\ &= E[(AX - E(AX))(AX - E(AX))^T] \\ &= E[A(X - E(X))(X - E(X))^T A^T] \quad \} \quad (AB)^T = B^T A^T \\ &= A E[(X - E(X))(X - E(X))^T] A^T \\ &= AC_X A^T \end{aligned}$$

$f_X(y)$: probability density function

$$Y = G(X), \quad X = G^{-1}(Y) = H(Y)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} H_1(Y) \\ H_2(Y) \\ \vdots \\ H_n(Y) \end{bmatrix}$$

$$J = \det \begin{pmatrix} \partial H_1 / \partial y_1 & \partial H_1 / \partial y_2 & \cdots & \partial H_1 / \partial y_n \\ \partial H_2 / \partial y_1 & \partial H_2 / \partial y_2 & \cdots & \partial H_2 / \partial y_n \\ \vdots & \vdots & & \vdots \\ \partial H_n / \partial y_1 & \partial H_n / \partial y_2 & \cdots & \partial H_n / \partial y_n \end{pmatrix}$$

$$f_Y(y) = f_X(H(y)) |J|$$

$$Y = AX + b = G(X)$$

$$X = A^{-1}(Y - b) = H(Y)$$

$$J = \det(A^{-1}) = \frac{1}{\det(A)}$$

$$\begin{aligned} f_Y(y) &= f_X(H(y)) |J| \\ &= f_X(A^{-1}(y - b)) \left| \frac{1}{\det(A)} \right| \end{aligned}$$

Probability density function of multivariate Gaussian distribution

$$Z = [Z_1, Z_2, \dots, Z_n]^T, \quad Z_i \sim N(0, 1), \quad Z_i \text{ i.i.d.}$$

$$\begin{aligned} f_Z(z) &= \prod_{i=1}^n f_{Z_i}(z_i) \\ &= \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2} \sum_{i=1}^n z_i^2\right) \\ &= \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2} z^T z\right) \end{aligned}$$

Probability density function of multivariate normal distribution

$$X = AZ + m, \quad Z \sim (0, I)$$

$$E(X) = m$$

$$C_X = AC_ZA^T = AA^T$$

$$\det(C_X) = \det(AA^T) = \det(A) \det(A^T) = (\det(A))^2$$

$$\sqrt{\det(C_X)} = |\det(A)|$$

$$f_X(x) = f_Z(A^{-1}(x-m)) \left| \frac{1}{\det(A)} \right|$$

$$= \frac{1}{\sqrt{2\pi}^n} \left| \frac{1}{\det(A)} \right| \exp\left(-\frac{1}{2}(A^{-1}(x-m))^T A^{-1}(x-m)\right)$$

$$= \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{\det(C_X)}} \exp\left(-\frac{1}{2}(x-m)^T (A^{-1})^T A^{-1}(x-m)\right)$$

$$= \frac{1}{\sqrt{2\pi}^n \sqrt{\det(C_X)}} \exp\left(-\frac{1}{2}(x-m)^T (A^T A)^{-1}(x-m)\right)$$

$$= \frac{1}{\sqrt{2\pi}^n \sqrt{\det(C_X)}} \exp\left(-\frac{1}{2}(x-m)^T C_X^{-1}(x-m)\right)$$

NLL to MSE

$$-\log p_\theta(y|x) = -\frac{1}{2}(f_\theta(x) - y)^T \Sigma_\theta(x)^{-1}(f_\theta(x) - y) - \frac{1}{2} \log |\Sigma_\theta(x)| - \frac{n}{2} \log 2\pi$$

$$\begin{aligned} \left(\begin{array}{l} N(f_\theta(x), \Sigma_\theta(x)) \\ \Sigma_\theta(x) = I \end{array} \right) &= -\frac{1}{2}(f_\theta(x) - y)^T (f_\theta(x) - y) \\ &= -\frac{1}{2} \|f_\theta(x) - y\|^2 \end{aligned}$$

Analyze error

$$L(\theta, \pi, y) = -\frac{1}{2} \|f_\theta(\pi) - y\|^2$$

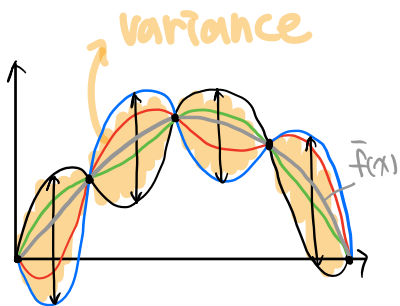
$$D = \{(\pi_1, y_1), (\pi_2, y_2), \dots, (\pi_n, y_n)\}$$

$$p(D) = \prod_{i=1}^n p(\pi_i) p(y_i | \pi_i)$$

$$\begin{aligned} E_{D \sim p(D)} [\|f_\theta(\pi) - y\|^2] &= E_{D \sim p(D)} [\|f_D(\pi) - f(\pi)\|^2] \\ &= \sum p(D) \|f_D(\pi) - f(\pi)\|^2 \end{aligned}$$

$$\text{let } \bar{f}(\pi) = E_{D \sim p(D)} [f_D(\pi)]$$

$$\begin{aligned} E_{D \sim p(D)} [\|f_D(\pi) - f(\pi)\|^2] &= E_{D \sim p(D)} [\|f_D(\pi) - \bar{f}(\pi) + \bar{f}(\pi) - f(\pi)\|^2] \\ &= E_{D \sim p(D)} [\|(f_D(\pi) - \bar{f}(\pi)) + (\bar{f}(\pi) - f(\pi))\|^2] \\ &= E_{D \sim p(D)} [\|f_D(\pi) - \bar{f}(\pi)\|^2] + E_{D \sim p(D)} [\|\bar{f}(\pi) - f(\pi)\|^2] \\ &\quad + \cancel{E_{D \sim p(D)} [2(f_D(\pi) - \bar{f}(\pi))^T (\bar{f}(\pi) - f(\pi))]} \\ &= \underbrace{E_{D \sim p(D)} [\|f_D(\pi) - \bar{f}(\pi)\|^2]}_{\text{Variance}} + \underbrace{E_{D \sim p(D)} [\|\bar{f}(\pi) - f(\pi)\|^2]}_{\text{Bias}^2} \end{aligned}$$



Variance \uparrow : overfitting
 bias \uparrow : underfitting

Bias²
 related with
 model capacity
 not data

Regularization

Bayesian perspective : add prior on parameters

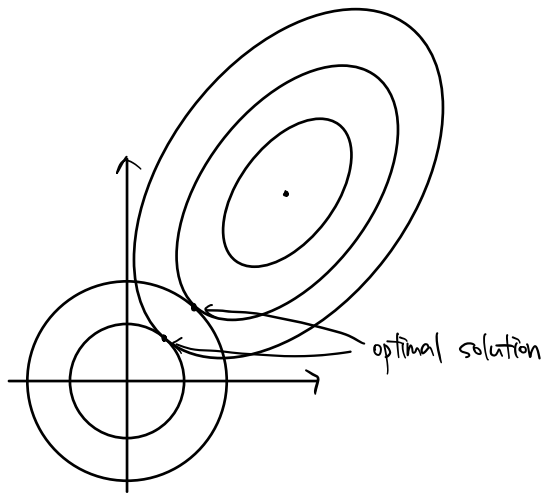
Given D , what is most likely θ ?

$$p(\theta|D) = \frac{p(\theta, D)}{p(D)} \propto p(\theta, D) = \underbrace{p(D|\theta)} \underbrace{p(\theta)}_{\text{prior}}$$

$$p(D|\theta) = \prod_{i=1}^N p_{\theta}(y_i | x_i) p(x_i)$$

chapter 2 - given θ , maximize probability of D

\Rightarrow New loss function : $-\sum \log p_{\theta}(y_i | x_i) - \log p(\theta)$



small number \rightarrow higher probability

if $\theta \sim N(0, \sigma^2)$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\theta^2}{2\sigma^2}\right)$$

$$\log p(\theta) = \sum_i -\frac{\theta_i^2}{2\sigma^2} - \frac{1}{2} \log 2\pi\sigma^2$$

$$= -\lambda \|\theta\|^2 \quad (\lambda = \frac{1}{2\sigma^2}, \text{hyperparameter})$$

\rightarrow L2 loss

L2 regularization

weight decay

if $\theta \sim \text{Laplace}(0, b)$

$$\frac{1}{2b} \exp\left(-\frac{|\theta|}{b}\right)$$

$$\log p(\theta) = \sum_i -\frac{|\theta_i|}{b} - \log 2b$$

$$= -\lambda \|\theta\|_1$$

\rightarrow L1 loss

L1 regularization

MLE (Maximum Likelihood Estimation) : $p(D|\theta)$

MAP (Maximum A Posterior) : $p(\theta|D)$