So far

$\mathcal{D} = \{(x^{(i)}, y_i) \mid x^{(i)} \in \mathbb{R}^d, y_i \in \mathbb{R}, i \in [n]\}$

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x^{(i)})^2 + \lambda \|w\|^2$$

① $\lambda \to 0$ least squares

② Ridge regression

③ $\lambda \|w\|^2 \to \lambda \|w\|_1 = \lambda \sum_{i=1}^{d} |w_i|$

LASSO

$$\min_{\substack{w \\ \|w\| \leq B}} \sum_{i=1}^{n} \ell(\hat{y}_i, y_i) \qquad \hat{y}_i = w^T x^{(i)}$$

SVR

$$\ell(\hat{y}_i, y_i) = \begin{cases} 0 & |\hat{y}_i - y_i| \leq \epsilon \\ |\hat{y}_i - y_i| - \epsilon & \text{otherwise} \end{cases}$$

$$\ell(\hat{y}_i, y_i) = \max(0, |\hat{y}_i - y_i| - \epsilon)$$

$$\min_{\{w,b\}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$y_i - w^T x^{(i)} - b \leq \epsilon + \xi_i$$

$$w^T x^{(i)} + b - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i \geq 0, \; \xi_i^* \geq 0$$

$$w^* = \sum_{i=1}^{n}(\lambda_i - \lambda_i^*)\, x^{(i)}$$

$$\hat{y} = w^{*T} x + b^* = \sum_{i=1}^{n}(\lambda_i - \lambda_i^*)\left(x^{(i)}\right)^T x + b$$

<u>Develop</u> Non-linear predictors

<u>Kernel trick</u>

$$w^{*T} x + b = \sum_{i=1}^{n}(\lambda_i - \lambda_i^*)\, K(x^{(i)}, x) + b$$

$K(x, x^{(i)})$ is a Kernel function

$E(\epsilon) = 0, \quad E(\epsilon^2) = \sigma^2$

$$Y = f^*(x) + \epsilon$$

$$f^*(x) = x^T w^*$$

$$f^*(x) = \sum_{j=1}^{d} w_j \phi_j(x) = \Phi(x)^T w \qquad j = 1, \ldots, d$$

$$w \sim N(0, \tfrac{1}{\alpha} I) \qquad \epsilon \sim N(0, \tfrac{1}{\beta})$$

Let $x^{(1)}, \ldots, x^{(n)}$ be data points

$$Z^{(n)} = \begin{bmatrix} \Phi(x^{(1)})^T \\ \\ \Phi(x^{(n)})^T \end{bmatrix} w = \Phi^{(n)}(x)^T w$$

$$E(Z^{(n)}) = \Phi^{(n)}(x)^T E(w) = 0$$

$$E(Z^{(n)} Z^{(n)T}) = \Phi^{(n)}(x)^T E(ww^T) \Phi^{(n)}(x)$$

$$= \frac{1}{\alpha} \Phi^{(n)}(x)^T \Phi^{(n)}(x)$$

$$\left(\Phi^{(n)}(x)^T \Phi^{(n)}(x)\right)_{ij} = \frac{1}{\alpha} \Phi(x^{(i)})^T \Phi(x^{(j)})$$

$$= \frac{1}{\alpha} K(x^{(i)}, x^{(j)})$$

$$C^{(n)} = E\left(z^{(n)} z^{(n)T}\right)$$

$$\left(C^{(n)}\right)_{ij} = \frac{1}{\alpha} K(x^{(i)}, x^{(j)})$$

$$Y = Z + \epsilon \qquad \epsilon \sim N\left(0, \frac{1}{\beta}\right)$$

$$E\left(z(x^{(i)})\, z(x^{(j)})\right) = K(x^{(i)}, x^{(j)})$$

$$P(Y = y \mid Z = z) = N\left(y \mid z, \frac{1}{\beta}\right)$$

$$P\left(Y^{(n)} = y^{(n)} \mid Z^{(n)} = z^{(n)}\right) = N\left(y^{(n)} \,\middle|\, z^{(n)}, \frac{I}{\beta}\right)$$

$$P\left(Z^{(n)} = z^{(n)}\right) = N\left(z^{(n)} \mid 0, K\right)$$

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix} \qquad X \sim N(\mu, C)$$

$$C = \begin{bmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{bmatrix} \qquad \begin{aligned} E(x) &= \mu \\ E(x-\mu)(x-\mu)^T &= C \end{aligned}$$

$$X_a | X_b \sim \ ?$$

Precision matrix. $\quad \Lambda = C^{-1}$

$$\Lambda = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$$

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M & -MBD^{-1} \\ - & = \end{bmatrix}$$

$$M = \left(A - BD^{-1}C\right)^{-1}$$

$$\Lambda_{aa} = \left(C_{aa} - C_{ab}C_{bb}^{-1}C_{ba}\right)^{-1}$$

$$\Lambda_{ab} = -\left(C_{aa} - C_{ab}C_{bb}^{-1}C_{ba}\right)^{-1}C_{ab}C_{bb}^{-1}$$

$$\frac{1}{2}(x-\mu)^T\Lambda(x-\mu)$$

$$= (x_a - \mu_a)^T\Lambda_{aa}(x_a - \mu_a) + \frac{1}{2}(x_a - \mu_a)^T\Lambda_{ab}(x_b - \mu_b)$$

$$+ \frac{1}{2}(x_b - \mu_b)^T\Lambda_{ba}(x_a - \mu_a)$$

$$+ \quad - \quad — \quad -$$

$$= \frac{1}{2}(x_a - \tilde{\mu})^T C_{a|b}^{-1}(x_a - \tilde{\mu}) + f(x_b)$$

$$C_{a|b}^{-1} = \Lambda_{aa} \quad \Big| \quad C_{a|b} = \Lambda_{aa}^{-1}$$

$\min g(x_a)$

$$= \frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa}(x_a - \mu_a) + \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}(x_b - \mu_b)$$
$$+ \frac{1}{2}(x_b - \mu_b)^T \Lambda_{ba}(x_a - \mu_a)$$
$$+ \cdots$$

$$\nabla g(x_a) = \Lambda_{aa}(x_a - \mu_a) + \Lambda_{ab}(x_b - \mu_b)$$

$$\tilde{\mu} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab}(x_b - \mu_b)$$

$$\Lambda_{aa}^{-1} \Lambda_{ab} = - C_{ab} C_{bb}^{-1}$$

$$\tilde{\mu}_{a|b} = \mu_a + C_{ab} C_{bb}^{-1} (x_b - \mu_b)$$

$$C_{a|b} = C_{aa} - C_{ab} C_{bb}^{-1} C_{ba}$$

$$x \sim N(\mu, \Lambda^{-1})$$

$$y|x \sim N(Ax+b, L^{-1})$$

$$y \sim N(A\mu + b, L^{-1} + A\Lambda^{-1}A^{T})$$

$$z^{(n)} \sim N(0, K^{(n)})$$

$$y^{(n)} | z^{(n)} \sim N\left(z^{(n)}, \frac{1}{\beta} I\right)$$

$$A = I, \; b = 0, \; L^{-1} = \frac{1}{\beta} I, \; \mu = 0$$

$$\Lambda^{-1} = K$$

$$y^{(n)} \sim N\left(0, \frac{1}{\beta} I^{(n)} + K^{(n)}\right)$$

$$y^{(n+1)} \sim N\Big(0, \underbrace{\frac{1}{\beta} I^{(n+1)} + K^{(n+1)}}_{C^{(n+1)}}\Big)$$

$$y^{(n+1)} = \begin{bmatrix} y_1 \\ \\ y_n \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} y^{(n)} \\ \\ y_{n+1} \end{bmatrix}$$

$$y_{n+1} \mid y^{(n)} \sim N\left(k^T \left(C^{(n)}\right)^{-1} y^{(n)}, \sigma^2\right)$$

$$C^{(n+1)} = \begin{bmatrix} C^{(n)} & k \\ \\ k^T & c \end{bmatrix}$$

$C_{aa} = c$

$\mu_a = 0$ $\quad C_{ab} = k$ $\quad C_{bb} = K^{(n)}$

$\mu_b = 0$ $\quad C_{ba} = k^T$ $\quad x_b = y^{(n)}$

$$C_{a|b} = c - k^T (K^{(n)})^{-1} k = \sigma^2$$

Summary

Gaussian Process: A stochastic process is called Gaussian process if for every finite subcollection in the index set T

$X = [X_{t_1}, \ldots, X_{t_k}]$ is a Multivariate Gaussian.

Model

$$Y = z(x) + \epsilon \qquad \epsilon \sim N(0, \tfrac{1}{\beta})$$

For any $x^{(1)} \ldots x^{(n)}$.

$$z^{(n)} = [z(x^{(1)}), \ldots, z(x^{(n)})]^T \qquad E(z^{(n)}) = 0$$

$$E(z^{(n)} z^{(n)T}) = K^{(n)} \qquad (K^{(n)})_{ij} = K(x^{(i)}, x^{(j)})$$

For any n

$$y^{(n)} \mid z^{(n)} \sim N\left(z^{(n)}, \frac{1}{\beta} I\right)$$

$$z^{(n)} \sim N\left(0, K^{(n)}\right)$$

$$y^{(n)} \sim N\left(0, C^{(n)}\right)$$

$$C^{(n)} = \frac{1}{\beta} I_n + K^{(n)}$$

If Z follows a Gaussian process then Y also follows a Gaussian Process

$$Y^{(n+1)} \sim N\left(0, C^{(n+1)}\right)$$

$$C^{(n+1)} = \begin{bmatrix} K^{(n)} & k \\ k^T & c \end{bmatrix}$$

$k \in \mathbb{R}^n$

$(k)_i = K\left(x^{(i)}, x^{(n+1)}\right)$

$c = K\left(x^{(n+1)}, x^{(n+1)}\right)$

$$Y_{n+1} \mid Y^{(n)} \sim N(\mu, \sigma^2)$$

$$\mu = k^T \left(C^{(n)}\right)^{-1} Y^{(n)}$$

$$\sigma^2 = c - k^T \left(C^{(n)}\right)^{-1} k$$