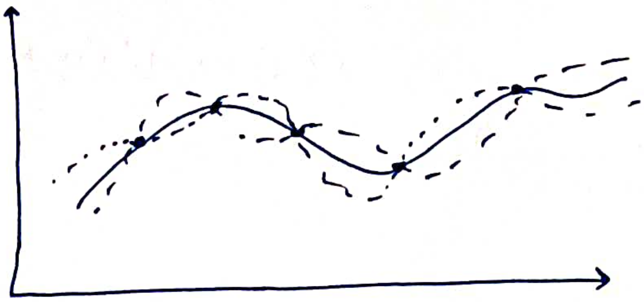
Gaussian Process Regression

Objective



There can be many functions which pass through these points.

Distribution over functions

$$\mu = K^{*} \cdot K^{-1} \cdot y \qquad V = K^{**} - K^{*} \cdot K^{-1} K^{*T}$$

① Kernel First

$y = f(x)$ ← Noiseless.

$f(x) \sim \mathcal{GP}(m(x), K(x, x'))$, $m(x) = 0$

$K(x, x') = e^{-\frac{1}{2}\|x - x'\|^2}$ If $x = x'$ $K(x, x') = 1$
$x - x' \to \pm\infty$ $K(x, x') \to 0$

$$K(X, X) = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \cdots & K(x_N, x_N) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \cdots & 1 \end{bmatrix}$$

Prior to seeing any data = GP "prior"

$$\begin{pmatrix} f(x) \\ f(x^*) \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K[X, X] & K[X, X^*] \\ K[X^*, X] & K[X^*, X^*] \end{bmatrix}\right)$$

If $\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$ then

$$y \mid x \sim \mathcal{N}\left(\mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mu_x), \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}\right)$$

$$\therefore f(x^*) \mid f(x) \sim \mathcal{N}\left(K(X^*, X)\, K(X,X)^{-1} f(x),\; K(X^*, X^*) - K(X^*, X)\, K(X,X)^{-1} K(X, X^*) \right)$$

If we introduce noise,

$y = f(x) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \qquad f(x) \sim \mathcal{N}(0, K)$

$\therefore y \sim \mathcal{N}(0, K + \sigma^2 I)$

$$\begin{pmatrix} y \\ f(x^*) \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix}\right)$$

$f(x^*) \mid y \sim \mathcal{N}(\bar{f}, \bar{V})$

$\bar{f} = K(X^*, X)\left(K(X,X) - \sigma^2 I\right)^{-1} y$

$\bar{V} = K(X^*, X^*) - K(X^*, X)\left(K(X,X) + \sigma^2 I\right)^{-1} K(X, X^*)$

② Prior

$y = f(x) + \epsilon \qquad f(x) = x^T w \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$

$L(y \mid X, w) = \prod P(Y_i = y_i) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\|y - Xw\|^2} = \mathcal{N}(Xw, \sigma^2 I)$

Bayesian Prior on w : $w \sim \mathcal{N}(0, \Sigma_p)$

Posterior : $w \mid y \propto L(y \mid X, w) \cdot \pi(w) = \ldots\ldots = \mathcal{N}(m, A^{-1})$

where $A = \frac{1}{\sigma^2} X^T X + \Sigma_p^{-1}$, $m = \frac{1}{\sigma^2} A^{-1} X^T y$

$f(x^*) = x^{*T} w \sim \mathcal{N}(x^{*T} m, x^{*T} A^{-1} x^*)$

$f(X^*) = X^{*T} w \sim \mathcal{N}(X^{*T} m, X^{*T} A^{-1} X^*)$

③ MSE

We want the best "linear" estimator - linear in terms of Y

$$\hat{y}(x) = \sum \lambda_i y_i = y^T \lambda \qquad y^* = \Lambda y$$

$$\min_{\lambda} \mathbb{E}[\hat{y}(x) - y(x)]^2 = \mathbb{E}[\lambda^T y y^T \lambda - 2\lambda^T y \cdot y^* - y^{*2}]$$

$$\mathbb{E}[\hat{y}(x) - y(x^*)]^2$$

$$= \lambda^T K \lambda - 2\lambda^T K^* - \sigma^2$$

$$\frac{\partial}{\partial \lambda} = 2K\lambda - 2K^* = 0 \Rightarrow \lambda = K^{-1}K^*$$

$$\hat{y}(x) = y^T K^{-1} K^* = K^{*T} K^{-1} y$$

$$\Lambda y \Rightarrow (y^* - \Lambda y)^T (y^* - \Lambda y) = y^{*T} y^* - 2y^{*T} \Lambda y + y^T \Lambda^T \Lambda y$$

$$\frac{\partial}{\partial \Lambda} = 2(y^* - \Lambda y) y^T = 0 \Rightarrow \Lambda (y y^T) = y^* y^T \Rightarrow \Lambda K = K^*$$

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

$$\hat{y} = K^* K^{-1} y$$

SVD:

In linear Regression, we know that $\hat{\beta} = (X^T X)^{-1} X^T y$

Let's decompose X via SVD: $X = USV^T$

$$\hat{\beta} = (VS^T U^T U S V^T)^{-1} V S^T U^T y = (V S^2 V^T)^{-1} V S^T U^T y$$

$$= (V^T)^{-1} (S^2)^{-1} V^{-1} V S^T U^T y$$

$$= V (S^2)^{-1} V^T V S^T U^T y$$

$$= \underbrace{V S^T U^T . U S (S^2)^{-1}}\, (S^2)^{-1} S^T U^T y$$

$$= X^T B$$

$$\hat{y}(X) = X \hat{\beta} = X X^T B$$

What if linear function is too limited?

Project to a higher dimensional feature space $\phi(x)$

$$f(X^*) = \phi(X^*) \cdot w \sim \mathcal{N}\left(\phi(X^*)\, m,\ \phi(X^*)\, A^{-1} \phi(x^*)^T\right)$$

Matrix A is $p \times p$ matrix, it will be difficult to invert if $p \to \infty$.

Let's define:

$$K(X, X) = X \Sigma_p X^T = K$$

$$\frac{1}{\sigma^2} X^T (K + \sigma^2 I) = \frac{1}{\sigma^2} X^T (X \Sigma_p X^T + \sigma^2 I) = A \Sigma_p X^T$$

$$A^{-1} \frac{1}{\sigma^2} X^T (K + \sigma^2 I)(K + \sigma^2 I)^{-1} = A^{-1} A \Sigma_p X^T (K + \sigma^2 I)^{-1}$$

$$\boxed{\frac{1}{\sigma^2} A^{-1} X^T = \Sigma_p X^T (K + \sigma^2 I)^{-1}} \quad -$$

$$\mathbb{E}(f(x^*)) = X^* m = X^* \frac{1}{\sigma^2} A^{-1} X^T y = X^* \Sigma_p X^T (K + \sigma^2 I)^{-1} y$$

$$x^* \Sigma_p X^T = K[X^*, X)$$

$$\therefore \mathbb{E}(f(X^*)) = K[X^*, X] \left(K(X, X) + \sigma^2 I\right)^{-1} y$$

$$V(f(X^*)) = K(X^*, X^*) - K[X^*, X)\left(K(X, X) + \sigma^2 I\right)^{-1} K(X, X^*)$$

$$(Z + U W V^T)^{-1} = Z^{-1} - Z^{-1} U (W^{-1} + V^T Z^{-1} U)^{-1} V^T Z^{-1}$$

<u>Matrix Inversion Lemma</u>

Now, if we replace $Z = \Sigma_p$ $\quad W^{-1} = \sigma^2 I \quad U = V = X$

"Minimal MSE

$$(y - XX^T B)^T (y - XX^T B) = y^T y - 2y^T XX^T B + B^T XX^T XX^T B.$$

$$\frac{\partial}{\partial B} = -2y^T XX^T + 2XX^T XX^T B = 0 \qquad K := XX^T$$

$$\Rightarrow K \cdot K \cdot B = Ky \Rightarrow B = K^{-1} y$$

$$\hat{y}(x^*) = X^* \hat{B} = X^* X^T K^{-1} y \qquad K^* = X^* X^T \qquad \hat{y}(x^*) = K^* K^{-1} y$$