11

Review: Bayesian Linear Regression

$D = \{(x_i, y_i)\}_{i=1}^{N}$, $X = [x_1, \dots, x_N] \in \mathbb{R}^{D \times N}$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$

$P(y|x,D) = \int P(y|x,w)P(w|D)\,dw \propto \exp\left(-\frac{1}{2}(y-\mu_{y|x,D})^T \Sigma^{-1}_{y|x,D}(y-\mu_{y|x,D})\right)$

- $\mu_{y|x,D} = \dfrac{x^T(xx^T + XX^T + \frac{\sigma^2}{\sigma_0^2}I)^{-1}Xy}{1 - x^T(xx^T+XX^T+\frac{\sigma^2}{\sigma_0^2}I)^{-1}x}$

- $\Sigma_{y|x,D} = \dfrac{\sigma^2}{1 - x^T(xx^T+XX^T+\frac{\sigma^2}{\sigma_0^2}I)^{-1}x}$

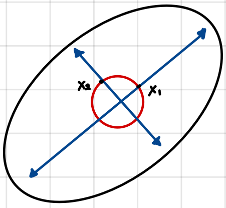
Woodbury Matrix Identity — using

$\Rightarrow (A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$

$\Rightarrow (\underset{A}{1} - \underset{U}{x}\,\underset{C}{(xx^T+XX^T+\alpha I)^{-1}}\,\underset{V}{x})^{-1} = 1 + x^T(XX^T+\alpha I)^{-1}x$

$Var[y|x,D] = \sigma^2(1 + \underbrace{x^T(\boxed{XX^T} + \alpha I)^{-1}x}_{\in \mathbb{R}^{D\times D}})$ $\quad (\alpha = \frac{\sigma^2}{\sigma_0^2})$

↳ Like a Covariance matrix



- x_1 방향이 x_2의 방향보다 크다.

$\Rightarrow Var(x_1) > Var(x_2)$

$(XX^T + \alpha I)^{-1}$

$\Rightarrow (\underset{A}{X}\underset{C}{I}\underset{V}{X^T} + \underset{A}{\alpha I})^{-1}$ ← Using Woodbury Matrix Identity

$= \frac{1}{\alpha}I - \frac{1}{\alpha^2}X(I + \frac{1}{\alpha}X^TX)^{-1}X^T$

$Var[y|x,D] = \sigma^2(1 + x^T(\frac{1}{\alpha}I - \frac{1}{\alpha^2}X(I+\frac{1}{\alpha}X^TX)^{-1}X^T)x)$

$= \sigma^2 + \frac{\sigma^2}{\alpha}x^T(I - X(\alpha I + X^TX)^{-1}X^T)x$ —: Small x

$= \underbrace{\sigma^2 + \sigma_0^2 x^Tx - \sigma_0^2 x^TX(XX^T+\alpha I)^{-1}X^Tx}$ $\in \mathbb{R}^{N\times N}$, Data의 수 / Dimension에 따라 더 개선된 식을 사용하는 것이 가능하다.

↑ Inner product

$XX^T = [x_1 \cdots x_N]^T[x_1 \cdots x_N] = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}[x_1 \cdots x_N] = \begin{bmatrix} x_1^Tx_1 & x_1^Tx_2 & \cdots & x_1^Tx_N \\ \vdots & & & \\ x_N^Tx_1 & \cdots & & x_N^Tx_N \end{bmatrix}$ ← Training dataset

$E[y|x,D] = \dfrac{x^T(xx^T + XX^T + \frac{\sigma^2}{\sigma_0^2}I)^{-1}Xy}{1 - x^T(xx^T+XX^T+\frac{\sigma^2}{\sigma_0^2}I)^{-1}x}$

↳ 아래식 전개 과정에서 분모가 여전히 생겨난다.
→ 잠깐 분자로 미루자.

$(1 + x^T(XX^T+\alpha I)^{-1}x)\,x^T(xx^T+XX^T+\alpha I)^{-1}Xy$ ← E[Y|X,D] 분자에 $\underbrace{(1+x^T(XX^T+\alpha I)^{-1}x)\text{를 곱한 식}}$

분모를 미리 제거하기 위함

$I \Leftarrow C$

$(\underset{U}{x}\underset{V}{x^T} + \underbrace{XX^T+\alpha I}_{A})^{-1} = (XX^T+\alpha I)^{-1} - (XX^T+\alpha I)^{-1}x(1+x^T(XX^T+\alpha I)^{-1}x)^{-1}\cdot x^T\cdot(XX^T+\alpha I)^{-1}$

$(1+x^T(XX^T+\alpha I)^{-1}x)\,x^T(xx^T+XX^T+\alpha I)^{-1}Xy = (1+x^T(XX^T+\alpha I)^{-1}x)\cdot x^T(XX^T+\alpha I)^{-1}Xy - \cancel{(1+x^T(XX^T+\alpha I)^{-1}x)}\,x^T\cdot\cancel{(XX^T+\alpha I)}\,x\cdot\cancel{(1+x^T(XX^T+\alpha I)^{-1}x)^{-1}}\,x^T(XX^T+\alpha I)^{-1}Xy$

$C \in \mathbb{R}$ $\qquad C^{-1}$

$= x^T\underbrace{(XX^T+\alpha I)^{-1}}_{\in \mathbb{R}^{D\times D}}Xy = x^TX\underbrace{(XX^T+\alpha I)^{-1}}_{\in \mathbb{R}^{N\times N}}y$

이 항등식 이용

$\underbrace{(XX^T+\alpha I)}_{N\times N}\underset{N\times D}{X} \overset{\Rightarrow N\times D}{} = XX^TX + \alpha X = \overset{N\times D}{X}\,\overset{D\times D}{\underbrace{(X^TX+\alpha I)}} \overset{\Rightarrow N\times D}{}$ ← 항등식에 의해 차원은 달라진다.

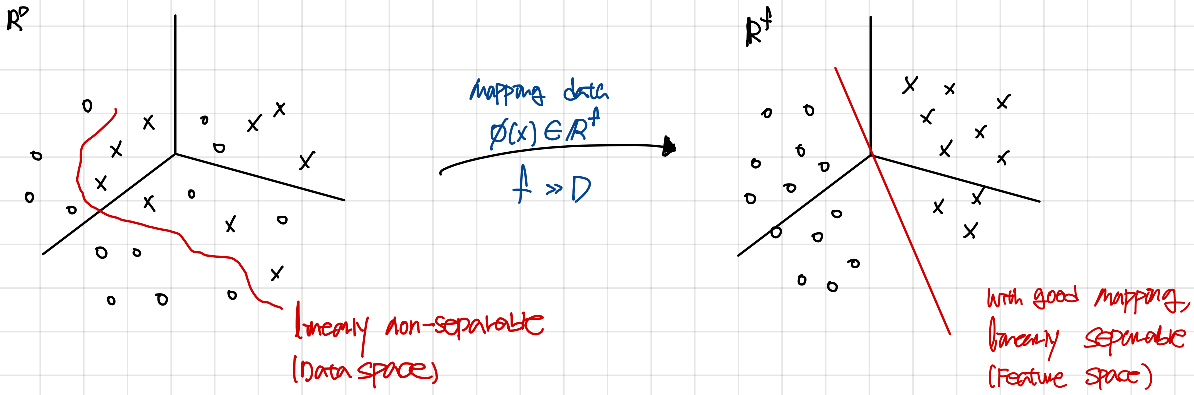
$X(X^TX+\alpha I)^{-1} = (XX^T+\alpha I)^{-1}X$, $X = (XX^T+\alpha I)^{-1}X(X^TX+\alpha I)$

$x^T(XX^T+\alpha I)^{-1}Xy = X^T\mu_w \Rightarrow E[y|x,D] = X^T\mu_w$

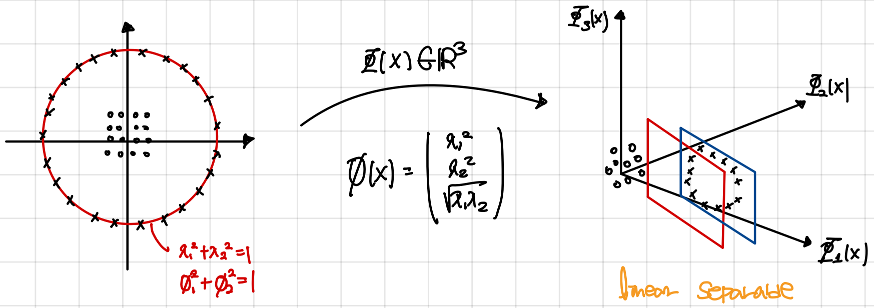
$Var[y|x,D] = \sigma^2 + \sigma^2x^T(XX^T+\alpha I)x = \sigma^2 + x^T\Sigma_w x$

kernel

$y = W^T x + b$



Ex)



$\phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2x_1x_2})$

$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$

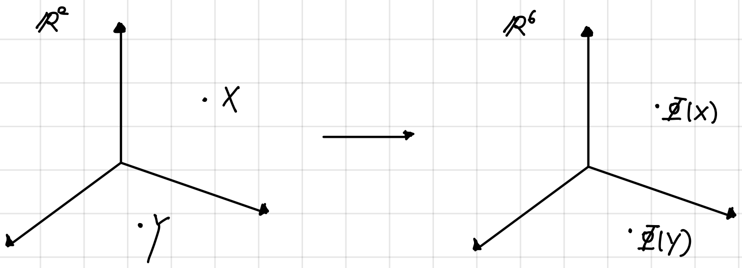
$$\phi(X)^T \phi(Y) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2x_1x_2} \end{pmatrix}^T \begin{pmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2y_1y_2} \end{pmatrix} = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2 = (x_1y_1 + x_2y_2)^2 = \underbrace{(X^TY)^2}_{\in \mathbb{R}} \equiv K(X, Y)$$

Ex) polynomial kernel

kernel func : $K(x, y) = (x^T y + 1)^d \in \mathbb{R}$

let $d = 2$ $(x, y \in \mathbb{R}^2)$

$$\Rightarrow K(x, y) = (x_1y_1 + x_2y_2 + 1)^2 = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2 + 2x_1y_1 + 2x_2y_2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ 1 \\ \sqrt{2x_1x_2} \\ \sqrt{2x_1} \\ \sqrt{2x_2} \end{pmatrix} \begin{pmatrix} y_1^2 \\ y_2^2 \\ 1 \\ \sqrt{2y_1y_2} \\ \sqrt{2y_1} \\ \sqrt{2y_2} \end{pmatrix} = \Phi^T(x)\Phi(y)$$



Review

Predictive distribution

$$P(y|x,D) \sim \mathcal{N}(y;\, \mu_{y|x,D}, \Sigma_{y|x,D})$$

$$-\ \mu_{y|x,D} = x^T(XX^T + \tfrac{\sigma^2}{\sigma_0^2}I)^{-1}Xy = x^T\mu_w$$

Posterior mean

$$-\ \Sigma_{y|x,D} = \sigma^2(1 + x^T(XX^T + \tfrac{\sigma^2}{\sigma_0^2}I)^{-1}x) = \sigma^2 + x^T\Sigma_w x$$

Posterior Covariance

$$-\ y = w^Tx + \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Noise, Randomness (Model과 상관없음)

Epistemic Uncertainty
: Uncertainty in the model itself
This can be reduced to 0 as $N \to \infty$

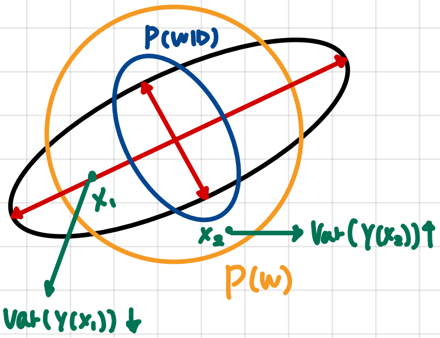
$$\Sigma_w = \frac{\sigma^2}{N}\left(\frac{1}{N}XX^T + \frac{\sigma^2}{N\sigma_0^2}I\right)^{-1}$$

$= \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$, if $N \to \infty$ $\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$ becomes constant matrix

$\Rightarrow \lim_{N\to\infty} \frac{\sigma^2}{N}(\text{constant}) = 0$

Aleatoric Uncertainty
: Uncertainty in data that cannot be reduced even if $N \to \infty$

Data Covariance



Prior $P(w) = \mathcal{N}(0, \sigma_0^2 I)$

- If Data Covariance ↑ → Posterior Covariance ↓
- If Data Covariance ↓ → Posterior Covariance ↑

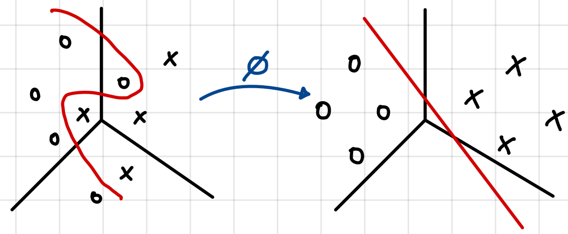
⇐ Posterior Covariance 식에서 σ_0가 분모에 있음

Posterior distribution 모양의 이유

—: predictive model

Expansion to model nonlinear w.r.t x

$$x \in \mathbb{R}^D \longmapsto \Phi(x) \in \mathbb{R}^l$$



$$y(x) = w^Tx \longmapsto y(x) = \alpha^T\phi(x)$$

Kernel function

$$k(x,y) = \phi(x)^T\phi(y)$$

Example of Kernel function

1. Polynomial Kernel

$$k(x,y) = (x^Ty + 1)^d,\quad d \in \mathbb{N}$$
$$= \phi(x)^T\phi(y)$$

2. Gaussian Kernel

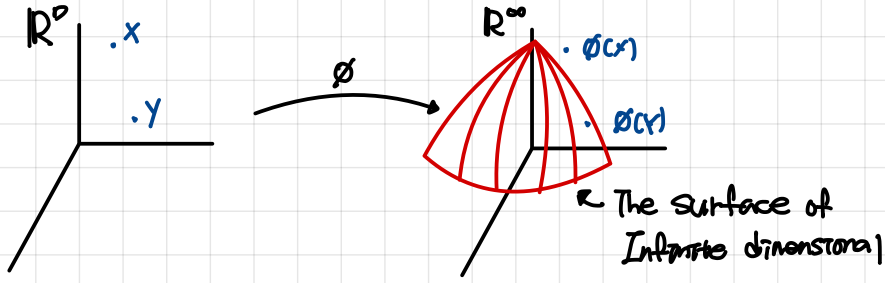
$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2}\right) \in \mathbb{R}$$
$$= \exp\left(-\frac{x^Tx + y^Ty - 2x^Ty}{2}\right)$$
$$= \exp\left(-\frac{x^Tx}{2}\right)\cdot\exp\left(-\frac{y^Ty}{2}\right)\cdot\exp(x^Ty)$$

$$\exp(t) = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots$$

$x^Ty = x_1y_1 + x_2y_2 + \cdots + x_Dy_D$

$$\exp(x^Ty) = \frac{1 + x^Ty + \frac{1}{2!}(x^Ty)^2 + \frac{1}{3!}(x^Ty)^3 + \cdots}{\sqrt{1 + x^Tx + \frac{1}{2!}(x^Tx)^2 + \cdots}\sqrt{1 + y^Ty + \frac{1}{2!}(y^Ty)^2 + \cdots}} = \phi(x)^T\phi(y)$$

$$\phi(x) = \frac{1}{\sqrt{1 + x^Tx + \frac{1}{2!}(x^Tx)^2 + \cdots}}\cdot\begin{pmatrix} x_1 \\ \vdots \\ x_D \\ \frac{1}{\sqrt{2}}x_1^2 \\ \vdots \\ \frac{1}{\sqrt{2}}x_D^2 \\ x_1x_2 \\ x_1x_3 \\ \vdots \end{pmatrix}$$



$k(x,x) = 1$
$= \phi(x)^T\phi(x)$
$\Rightarrow \|\phi(x)\| = 1$

Back to Bayesian linear Regression, Another Version for $\mu_{Y|X,D}$, $\Sigma_{Y|X,D}$

- $\mu_{Y|X,D} = x^TX(X^TX + \frac{\sigma^2}{\sigma_0^2}I)^{-1}y = k^T(K + \frac{\sigma^2}{\sigma_0^2}I)^{-1}y$
- $\Sigma_{Y|X,D} = \sigma^2 + \sigma_0^2 x^Tx - \sigma_0^2 x^TX(X^TX + \frac{\sigma^2}{\sigma_0^2}I)^{-1}X^Tx = \sigma^2 + \sigma_0^2 k^T(K + \frac{\sigma^2}{\sigma_0^2}I)k$

kernelize

$x \longmapsto \phi(x)$

$X = [x_1 \cdots x_N] \longmapsto \Phi = [\phi(x_1) \cdots \phi(x_N)]$

$x^Tx \longmapsto \phi(x)^T\phi(x) = k(x,x)$

$x^TX \longmapsto [\phi(x)^T\phi(x_1) \quad \phi(x)\phi(x_2) \quad \cdots \quad \phi(x)\phi(x_N)]$

$= [k(x,x_1) \quad k(x,x_2) \quad \cdots \quad k(x,x_N)]$

$= k^T \in R^{1\times N}$ and $k_i = k(x,x_i)$

Small k

$X^TX \longmapsto \Phi^T\Phi$

$$= \begin{bmatrix} \phi(x_1)^T\phi(x_1) & \cdots & \phi(x_1)^T\phi(x_N) \\ \vdots & & \vdots \\ \phi(x_N)^T\phi(x_1) & & \phi(x_N)^T\phi(x_N) \end{bmatrix} = \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_N) \\ \vdots & & \vdots \\ k(x_N,x_1) & & k(x_N,x_N) \end{bmatrix} \equiv K \in R^{N\times N}$$

Large K

$K_{ij} = k(x_i,x_j)$

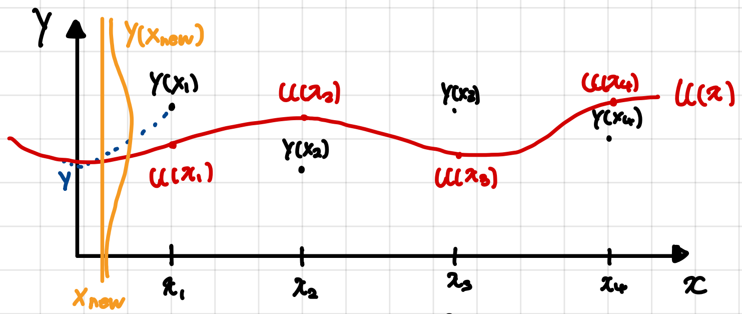
In fact, the above process is an example of the Gaussian processes

Gaussian Processes (= Infinite dimensional Gaussian)

Recall:

$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$

$\Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$



$Y(x_1), Y(x_2), Y(x_3), Y(x_4)$ are jointly Gaussian

$\mu(x)$: mean function

$\Sigma(x_1,x_2)$: Covariance function
$= \mathrm{cov}(Y(x_1), Y(x_2))$

$GP(\mu,\Sigma)$: Gaussian Process of Mean function $\mu(x)$ and Covariance Matrix $\Sigma(x_1,x_2)$

Assume that $Y(x_1) \cdots Y(x_N)$ are known for one realization.

We'd like to know the prediction for $Y(x)$ for a new x