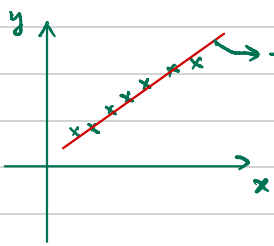
线性回归

$Data = \{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

$$X = (x_1 \ x_2 \ \cdots \ x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N \times p}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1}$$



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

$y = f(x) + \varepsilon$ (ε: noise)

x, y, ε are r.v.

$x \in \mathbb{R}^p$, $y \in \mathbb{R}^1$, $\varepsilon \sim N(0, \sigma^2)$

Bayesian method → 贝叶斯派 (w is r.v.) ⇓ $P(w|Data)$

→ 频率派 (w is an unknown constant) 优化问题、点估计

Lasso Ridge

LSE → Regularized LSE

⇑ MLE (noise is Gaussian)

⇕ MAP (noise is Gaussian)

$$W_{MLE} = \arg\max_w P(Data|w)$$

$$P(w|Data) \propto P(Data|w) \cdot P(w)$$

$$W_{MAP} = \arg\max_w P(w|Data) = \arg\max_w P(Data|w) P(w)$$

$P(w)$: $N(\mu_0, \Sigma_0)$ ↔ Ridge; Laplace ↔ Lasso

Bayesian Linear regression

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

$X = (x_1, x_2, \cdots, x_N)^T$

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1}$

Model:

$$\begin{cases} f(x) = w^T x = x^T w \\ y = f(x) + \varepsilon \end{cases}$$

x, y, ε are r.v.

$\varepsilon \sim N(0, \sigma^2)$

Bayesian Method

$$\begin{cases} \text{Inference}: \text{posterior}(w) \\ \text{Prediction}: x^* \to y^* \end{cases}$$

$$P(w, Y|X) = P(Y|w, X) \cdot P(w|X)$$

$P(w)$

Inference $P(w|\text{Data})$

likelihood, prior

$$P(w|\text{Data}) = P(w|X, Y) = \frac{P(w, Y|X)}{P(Y|X)} = \frac{P(Y|w, X) \cdot P(w)}{\int P(Y|w, X) \cdot P(w)\, dw}$$

$$P(Y|w, X) = \prod_{i=1}^N P(y_i|w, x_i) = \prod_{i=1}^N N(y_i|w^T x_i, \sigma^2)$$

$$P(y|x, w) = N(w^T x, \sigma^2)$$

共轭：Gaussian分布是自共轭的

$P(w) = N(0, \Sigma_p)$

$P(w|\text{Data}) \propto P(Y|w, X) \cdot P(w)$

Gaussian, Gaussian, Gaussian

$$\propto \prod_{i=1}^N N(y_i|w^T x_i, \sigma^2) \cdot N(0, \Sigma_p)$$

$\to N(\mu_w, \Sigma_w)$ μ_w?, Σ_w?

likelihood:

$$P(Y|X,W) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - W^T x_i)^2\right\}$$

$$= \frac{1}{2\pi^{\frac{N}{2}}\sigma^N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - W^T x_i)^2\right\}$$

$$\sum_{i=1}^{N}(y_i - W^T x_i)^2 = (y_1 - W^T x_1 \quad y_2 - W^T x_2 \cdots \quad y_N - W^T x_N)\begin{pmatrix} y_1 - W^T x_1 \\ \vdots \\ y_N - W^T x_N \end{pmatrix}$$

$$(y_1 \; y_2 \cdots y_N) - W^T(x_1 \; x_2 \cdots x_N)$$

$$(Y^T - W^T X^T)(Y - XW)$$

$$\frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left\{-\frac{1}{2}(Y-XW)^T\sigma^{-2}I(Y-XW)\right\}$$

$$= N(XW, \sigma^{-2}I)$$

$$P(W|Data) \propto N(XW, \sigma^{-2}I) \cdot N(0, \Sigma_p)$$

$$\propto \exp\left\{-\frac{1}{2}(Y-XW)^T\sigma^{-2}I(Y-XW)\right\}\exp\left\{-\frac{1}{2}W^T\Sigma_p^{-1}W\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}(Y^T - W^T X^T)(Y-XW) - \frac{1}{2}W^T\Sigma_p^{-1}W\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(Y^TY - 2Y^TXW + W^TX^TXW) - \frac{1}{2}W^T\Sigma_p^{-1}W\right\}$$

$P(x) = N(\mu, \Sigma)$

指数部分: $\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

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

$$= -\frac{1}{2}(x^T\Sigma^{-1}x - 2\mu^T\Sigma^{-1}x + \Delta)$$

二次项: $-\frac{1}{2\sigma^2}W^TX^TXW - \frac{1}{2}W^T\Sigma_p^{-1}W = -\frac{1}{2}(W^T(\sigma^{-2}X^TX + \Sigma_p^{-1})W)$

$\Sigma_W^{-1} = A$

一次项: $+\frac{1}{2\sigma^2}(2)Y^TXW = \sigma^{-2}Y^TXW$

$$\mu_W^T A = \mu_W^T\Sigma_W^{-1} = \sigma^{-2}Y^TX$$

$$A\mu_W = \sigma^{-2}X^TY$$

$$\mu_W = \sigma^{-2}A^{-1}X^TY$$

$P(w|Data) = N(\mu_w, \Sigma_w)$

$\mu_w = \sigma^{-2} A^{-1} X^T Y$

$\Sigma_w = A^{-1}$

$A = \sigma^{-2} X^T X + \Sigma_p^{-1}$

Prediction:

Given x^*, y^* $\quad P(f(x^*)|Data, x^*) = N(x^{*T}\mu_w, x^{*T}\Sigma_w x^*)$

noise-free

① $f(x^*) \Rightarrow f(x^*) = x^{*T} w \rightarrow P(w|Data) = N(\mu_w, \Sigma_w)$

$x^{*T} w \sim N(x^{*T}\mu_w, x^{*T}\Sigma_w x^*)$

② y^* . $y^* = f(x^*) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

$P(y^*|x^*, Data) = N(x^{*T}\mu_w, x^{*T}\Sigma_w x^* + \sigma^2)$

Gaussian Process

distribution　　随机过程

Gaussian distribution ⟶ multivariate Gaussian distribution

一维

$p(x) = N(\mu, \sigma^2)$

⇕

Gaussian Network

$x \in \mathbb{R}^p$

$p(x) = N(\mu, \Sigma)$, $\Sigma_{p \times p} \Rightarrow p < +\infty$

高斯过程是定义在连续域（时间/空间）上的无限多个高斯随机变量所组成的随机过程

↓

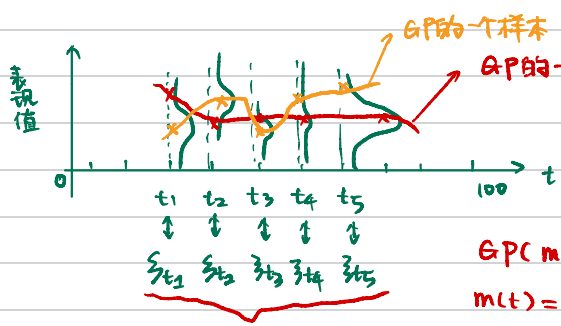
无限维 Gaussian Distribution

⇕

Gaussian Process

$\{\xi_t\}_{t \in T}$, $T \to$ 连续域，if $\forall n \in N^+$, $t_1, t_2, \cdots, t_n \in T$

s.t. $\{\xi_{t_1}, \xi_{t_2}, \cdots, \xi_{t_n}\} \triangleq \xi_{t_1 - t_n} \sim N(\mu_{t_1-t_n}, \Sigma_{t_1-t_n})$. 那么 $\{\xi_t\}_{t \in T}$ 就是一个 Gaussian process



$GP(m(t), k(s,t))$

$m(t) = E[\xi_t]$ → mean function

$k(s,t) = E[(\xi_s - m(s))(\xi_t - m(t))^T]$

↓

kernel & covariance function

人的一生：$[0, 100]$

$t \in [0, 100]$, $\xi_t \in N(\mu_t, \sigma_t^2)$

↳ 这个人在 t 时刻他的表现（0～100分）

$t = 0$，他的一生基本已定，$t > 0$，每一个时刻，μ_t，σ_t^2 是已经确定

Gaussian Process Regression (GPR)

Recall Bayesian Linear Regression

① $P(w|Data) = N(w|\mu_w, \Sigma_w)$

$$\begin{cases} \mu_w = \sigma^{-2} A^{-1} X^T Y \\ \Sigma_w = A^{-1} \end{cases}$$

$$A = \sigma^{-2} X^T X + \Sigma_p^{-1}$$

② Given x^*

$$P(f(x^*)|Data, x^*) = N(x^{*T}\mu_w,\ x^{*T}\Sigma_w x^*)$$

$$P(y^*|Data, x^*) = N(x^{*T}\mu_w,\ x^{*T}\Sigma_w x^* + \sigma^2)$$

$\phi: x \mapsto z$, $x \in \mathbb{R}^p$, $z \in \mathbb{R}^q$, $z = \phi(x)$

Nonlinear → { ① Non-linear Transformation } —kernel→ kernel BLR
{ ② Bayesian 2R }

Noise-free:

$$f(x^*)|X, Y, x^* \sim N(x^{*T}(\sigma^{-2}A^{-1}X^TY),\ x^{*T}A^{-1}x^*)$$

$$A = \sigma^{-2}X^TX + \Sigma_p^{-1}$$

If $\phi: x \mapsto z$, $x \in \mathbb{R}^p$, $z = \phi(x) \in \mathbb{R}^q$, $q > p$

Define: $\Phi = \phi(X) = (\phi(x_1)\ \phi(x_2)\ \cdots\ \phi(x_N))^T$ $N \times q$

$\Rightarrow f(x) = \phi^T(x) w$

$$f(x^*)|X, Y, x^* \sim N(\sigma^{-2}\phi(x^*)^T A^{-1}\Phi^T Y,\ \phi(x^*)^T A^{-1}\phi(x^*))$$

$$A = \sigma^{-2}\Phi^T\Phi + \Sigma_p^{-1}$$

compute A^{-1} → woodbury formula

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

$$A\Sigma_p = \sigma^{-2}\Phi^T\Phi\Sigma_p + I$$

$$A\Sigma_p\Phi^T = \sigma^{-2}\Phi^T\Phi\Sigma_p\Phi^T + \Phi^T$$

$$= \sigma^{-2}\Phi^T(K + \sigma^2 I) \Rightarrow K = \Phi\Sigma_p\Phi^T$$

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

$$\underbrace{\sigma^{-2} \phi(X^*)^T A^{-1} \bar{\Phi}^T Y}_{\text{Expectation}} = \phi(X^*) \Sigma_p \bar{\Phi}^T (K + \sigma^2 I)^{-1} Y$$

$$\Rightarrow f(X^* \mid X, Y, X^*) \sim N\left(\phi(X^*)^T \Sigma_p \bar{\Phi}^T (K + \sigma^2 I)^{-1} Y,\ \phi(X^*)^T \Sigma_p \phi(X^*) - \phi(X^*)^T \Sigma_p \bar{\Phi}^T (K + \sigma^2 I)^{-1} \bar{\Phi} \Sigma_p \phi(X^*)\right)$$

$K = \bar{\Phi} \Sigma_p \bar{\Phi}^T$	$\phi(X^*)^T \Sigma_p \bar{\Phi}^T$
$\phi(X^*)^T \Sigma_p \bar{\Phi}^T$	$\bar{\Phi} \Sigma_p \phi(X^*)$
$\phi(X^*)^T \Sigma_p \phi(X^*)$	$\underset{N \times p}{\bar{\Phi}} = \phi(X) = (\phi(x_1), \cdots, \phi(x_N))^T$

$k(x, x') = \phi^T(x) \Sigma_p \phi(x') \leftrightarrow$ kernel function

Σ_p: positive definite, $\Sigma_p = (\Sigma_p^{\frac{1}{2}})^2$

$k(x, x') = \phi^T(x) \Sigma_p^{\frac{1}{2}} \Sigma_p^{\frac{1}{2}} \phi(x') = (\Sigma_p^{\frac{1}{2}} \phi(x))^T \cdot \Sigma^{\frac{1}{2}} \phi(x') = \langle \psi(x), \psi(x') \rangle$

$\psi(x) = \Sigma_p^{\frac{1}{2}} \phi(x)$

Bayesian linear Regression + kernel trick (Non-linear Transformation Inner product)

GPR → ① weight-space view $\{ f(x) = \phi^T(x) w,\ y = f(x) + \varepsilon$

② function-space view

$f(x)$ is r.v.

$f(x) \sim GP(m(x), k(x, x'))$

$f(x)$