Gaussian Process
高斯过程

zhoujie
2018.11.13

Gaussian Process

GPR: weight-space

Recall Bayesian Linear Regression

① $P(w|Data) = N(w|\mu_w, \Sigma_w)$

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

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

$\begin{cases} f(x) = w^Tx = x^Tw \\ y = f(x) + \varepsilon, \ \varepsilon \sim N(0, \sigma^2) \end{cases}$

② Given x^*.

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

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

Noise-free: $X = (x_1 \ x_2 \cdots x_N)^T, \ Y = (y_1 \ y_2 \cdots y_N)^T$

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

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

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

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

Then: $f(x) = \phi(x)^Tw$

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

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

How to compute A^{-1}? $\longrightarrow$ woodbury formula.

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

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

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

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

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

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

$y = f(x) + \varepsilon$, $f(x)$ is not linear function

Non-linear $\longrightarrow$ $\begin{cases} \text{① Non-linear Transformation} \\ \text{② Bayesian LR} \end{cases}$ $\xrightarrow[\text{Kernel trick}]{\text{Kernel}}$ Kernel BLR

$\phi(x)$

$\phi: x \mapsto z$

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

$K(x, x') = $ kernel function

$$\Longleftrightarrow \underbrace{\sigma^{-2}\phi(x^*)^T A^{-1}\Phi^T Y}_{f(x^*)|X,Y,x^*\text{'s expectation}} = \phi(x^*)\Sigma_p \Phi^T(K+\sigma^2 I)^{-1}Y$$

$$\boxed{\begin{array}{l}(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1} \\ \text{Woodbury Formula}\end{array}}$$

Likewise : $f(x^*)|X,Y,x^*$'s covarience : $\phi(x^*)^T\Sigma_p\phi(x^*) - \phi(x^*)^T\Sigma_p\Phi^T(K+\sigma^2 I)^{-1}\Phi\Sigma_p\phi(x^*)$

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

$K = \Phi\Sigma_p\Phi^T$ $\quad$ $\phi(x^*)^T\Sigma_p\Phi^T$

$\phi(x^*)^T\Sigma_p\Phi^T$ $\quad$ $\Phi\Sigma_p\phi(x^*)$

$\phi(x^*)^T\Sigma_p\phi(x^*)$

$K(x,x') = \phi^T(x)\Sigma_p\phi(x') \overset{?}{\longleftrightarrow}$ Kernel function

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

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

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

GPR: weight-space → function-space

Recall Gaussian Process: function-space 和 weight-space 结果一样，但更加简单.

$\{\xi_t\}_{t\in T}$, T: continuous time/space.

$\forall n \in N^+ (n\geq 1)$, $\underbrace{t_1, t_2, \cdots, t_n}_{\text{index}} \to \underbrace{\xi_1, \xi_2, \cdots, \xi_n}_{\text{r.v}}$

令 $\xi_{1:n} = (\xi_1\ \xi_2 \cdots \xi_n)^T$,

If $\xi_{1:n} \sim N(\mu_{1:n}, \Sigma_{1:n})$,

$t \to \xi_t$
t: index
ξ_t: r.v

Then $\{\xi_t\}_{t\in T}$ is Gaussian Process.

$$\boxed{\xi_t \sim GP(\underbrace{m(t)}_{\text{mean function}}, \underbrace{K(t,s)}_{\text{covariance function}})}$$

为什么 GP 由 $m(t)$, $K(t,s)$ 来表达. (高斯过程存在性定理)

回到 weight-space view. (关注对象为 w)

$$\begin{cases} f(x) = \phi(x)^T w \\ y = f(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \end{cases}$$

Bayesian Method:

给定 prior: $w \sim N(0, \Sigma_p)$

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

$\therefore E[f(x)] = E[\phi(x)^T w] = \phi(x)^T E[w] = 0$

$\forall x, x' \in \mathbb{R}^p$,

$$\begin{aligned} cov(f(x), f(x')) &= E[(f(x) - E[f(x)])(f(x') - E[f(x')])] \\ &= E[f(x) \cdot f(x')] \\ &= E[\phi(x)^T w \cdot \phi(x')^T w] \\ &= E[\phi(x)^T w \cdot w^T \phi(x')] \end{aligned}$$

$$\mathrm{cov}(f(x), f(x')) = E[\phi(x)^T \underline{w \cdot w^T} \phi(x')]$$

$$= \phi(x)^T \underbrace{E[w \cdot w^T]} \phi(x')$$

$$= \phi(x)^T \cdot \Sigma_p \phi(x') = \langle \psi(x), \psi(x') \rangle = K(x, x')$$

启发：$f(x)$ 是否看作是一个高斯过程.

$\{f(x)\}_{x \in \mathbb{R}^p}$

GPR:

① weight-space view：关注的是 w

② function-space view：关注的是 $f(x)$

$x^*, \longrightarrow y^*$

$$P(y^* | Data, x^*) = \int_w P(y^* | w, x^*) \cdot P(w)\, dw$$

$$P(y^* | Data, x^*) = \int P(y^* | f, x^*) \cdot P(f)\, df$$

Gaussian Process $\rightarrow$ for GPR

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

① $f(x)$ is function

② $f(x)$ is r.v. $\rightarrow$ 高斯分布

$m(x)$ $\rightarrow$ mean function

$K(x, x')$ $\rightarrow$ covariance function

$t \rightarrow \xi_t$, $\{\xi_t\}_{t \in T} \sim GP$

$x \rightarrow f(x)$, $\{f(x)\}_{x \in \mathbb{R}^p} \sim GP$

$T = \mathbb{R}^p \rightarrow$ 相容的

$t \leftrightarrow x$

$\xi_t \leftrightarrow f(x)$

GPR: function space view

$\{f(x)\}_{x\in\mathbb{R}^p} \sim GP(m(x), K(x,x'))$

$m(x) = E[f(x)]$

$K(x,x') = E[(f(x)-m(x))(f(x')-m(x'))^T]$

Regression:

Data: $\{(x_i, y_i)\}_{i=1}^N$, $y = f(x) + \varepsilon$

$X = (x_1\ x_2 \cdots x_N)^T_{N\times p}$

$Y = (y_1\ y_2 \cdots y_N)^T_{N\times 1}$

$f(X) \sim N(\mu(X), K(X,X))$

$Y = f(X) + \varepsilon \sim N(\mu(X), K(X,X) + \sigma^2 I)$

$f(x)$ is Normal Dist.

$\boxed{\text{Prediction:}}$ Given $\boxed{X^* = (x_1^*\ x_2^* \cdots x_M^*)}$ $\to x^* \to y^* = f(x^*) + \varepsilon$

$Y^* = f(X^*) + \varepsilon$

$$\begin{array}{l} x_a \\ x_b \end{array}\begin{pmatrix} Y \\ f(X^*) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(X) \\ \mu(X^*) \end{pmatrix}_{\mu_a,\ \mu_b}, \begin{pmatrix} K(X,X)+\sigma^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{pmatrix}_{\Sigma_{aa},\ \Sigma_{ab},\ \Sigma_{ba},\ \Sigma_{bb}} \right)$$

$\boxed{p(f(X^*) \mid Y, X, X^*)}$ → 条件概率 → $N(\mu^*, \Sigma^*)$

$\phi(x^*)^T w$

已知联合高维分布，求条件概率.

$p(x_b \mid x_a)$

$\mu^* = K(X^*, X)\cdot(K(X,X)+\sigma^2 I)^{-1}(Y - \mu(X)) + \mu(X^*)$

$\Sigma^* = K(X^*, X^*) - K(X^*, X)(K(X,X)+\sigma^2 I)^{-1} K(X, X^*)$

$y^* = f(x^*) + \varepsilon$

公式: $x \sim N(\mu, \Sigma)$

$$x=\begin{pmatrix} x_a \\ x_b \end{pmatrix}\begin{matrix} p \\ q \end{matrix} \quad \mu=\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma=\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$$x_b | x_a \sim N(\mu_{b|a}, \Sigma_{b|a})$$

$$\boxed{\begin{cases} \mu_{b|a} = \Sigma_{ba}\Sigma_{aa}^{-1}(x_a-\mu_a)+\mu_b \\ \Sigma_{b|a} = \Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \end{cases}}$$

$$P(f(x^*) | Y; X, x^*) = N(\mu^*, \Sigma^*)$$

noise-free

$$P(y^* | Y, X, x^*) = N(\mu_y^*, \Sigma_y^*)$$

noise

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

$$\begin{cases} \mu_y^* = \mu^* \\ \Sigma_y^* = \Sigma^* + \sigma^2 I \end{cases}$$

Thank You !