再生核希尔伯特空间

黄政宇
北京大学北京国际数学研究中心
北京大学国际机器学习研究中心

本堂课大纲

- 再生核希尔伯特空间(核函数观点)

- 再生核希尔伯特空间(正则化观点)
  - 核岭回归(Kernel ridge regression)
  - 核岭插值(Kernel interpolation)

- 数值计算
  - Nyström近似
  - 随机特征方法

高斯过程能近似什么样的函数，需要多少数据，
近似出来的是什么函数？

再生核希尔伯特空间

再生核希尔伯特空间(Reproducing Kernel Hilbert Spaces)

f是定义在 R^N 的实值函数，$\mathcal{H}$是由这些函数构成的希尔伯特空间(完备(completeness)，$\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$)，$\mathcal{H}$是一个再生核希尔伯特空间，如果存在一个(对称正定核)函数$\kappa: R^N \times R^N \to R$，满足

- 对任意$x \in R^N$，$\kappa(\cdot, x)$ 属于$\mathcal{H}$

- κ有可再生性，即$\langle f(\cdot), \kappa(\cdot, x)\rangle_{\mathcal{H}} = f(x)$

$f(x)$能逐点定义，有“连续性”

$\lim_n \|f_n - f\|_{\mathcal{H}} \to 0$，那么$\lim_n f_n(x) - f(x) = 0 \quad \forall x$

再生核希尔伯特空间

- 练习

希尔伯特空间$L_2(\Omega)$:

$$\langle f, g\rangle_{\mathcal{H}} = \int f(x)g(x)dx$$

不是再生核希尔伯特空间。

再生核希尔伯特空间

- 例子

假设存在一个半正定的核函数，以及对应于测度μ的特征值分解

$$\kappa(x,x') = \sum_{i=1}^{N} \lambda_i \phi_i(x) \phi_i(x')$$

$$\int \phi_i(x)\phi_j(x)d\mu(x) = \delta_{ij}$$

定义由所有$f(x) = \sum_{i=1}^{N} f_i \phi_i(x)$构成的希尔伯特空间（其中$N$固定），内积为$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \frac{f_i g_i}{\lambda_i}$，这是一个再生核希尔伯特空间。

再生核希尔伯特空间

Moore-Aronszajn 定理

每一个正定的核函数$\kappa: R^N \times R^N \to R$都对应唯一的一个再生希尔伯特空间$\mathcal{H}$。

RKHS 范数$\|\cdot\|_{\mathcal{H}}$实际上是仅与核相关的属性，并且在测度的变化下是不变的。

再生核希尔伯特空间

- 证明

给定一个半正定的核函数，定义函数空间$\mathcal{H}_0$

$$\mathcal{H}_0 := span\{\kappa(\cdot, x): x \in R^N\}$$

$$= \left\{ f = \sum_{i=1}^{n} c_i \kappa(\cdot, x_i): n \in N, c_i \in R, x_i \in R^N \right\}$$

对于$f = \sum_{i=1}^{n} a_i \kappa(\cdot, x_i) \in \mathcal{H}_0$，$g = \sum_{i=1}^{m} b_i \kappa(\cdot, y_i) \in \mathcal{H}_0$,我们定义内积为

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \kappa(x_i, y_j)$$

再生核希尔伯特空间$\mathcal{H}$定义为它的闭包。

再生核希尔伯特空间

Wendland [2005, Corollary 10.48]

$\kappa_{\nu,l}$是定义在带有Lipschitz边界的空间$\mathcal{X} \subset R^d$上的Matern核，$s = \nu + d/2$，那么再生核希尔伯特空间$\mathcal{H}_{\kappa_{\nu,l}}$和阶数为 s 的Sobolev 空间

$$W_2^s(\mathcal{X}) := \{f \in L_2(\mathcal{X}): \|f\|_{W_2^s}^2 := \sum_{\beta \in N_0^d : |\beta| \le s} \left\|D^\beta f\right\|_{L_2}^2 < \infty\}$$

在范数意义上是等价的。即$\mathcal{H}_{\kappa_{\nu,l}} = W_2^s(\mathcal{X})$，且存在常数$c_1$，$c_2 > 0$

$$c_1\|f\|_{W_2^s}^2 \le \|f\|_{\mathcal{H}_{\kappa_{\nu,l}}}^2 \le c_2\|f\|_{W_2^s}^2$$

$\|f\|_{\mathcal{H}}^2$不仅蕴含了大小信息，也蕴含了光滑信息

再生核希尔伯特空间

Wendland [2005, Theorem 10.12]

κ 是定义在 R^N 上的有平移不变性的核 $\kappa(x,y) := \kappa(x-y)$，$\kappa \in C(R^d) \cap L_1(R^d)$，那么再生核希尔伯特空间 $\mathcal{H}_\kappa$ 满足

$$\mathcal{H}_\kappa = \{f \in L_2(R^d) \cap C(R^d): \|f\|_{\mathcal{H}_\kappa}^2 := \frac{1}{(2\pi)^{d/2}} \int \frac{|\mathcal{F}[f](\omega)|^2}{2\mathcal{F}[\kappa](\omega)} d\omega < \infty\}$$

内积定义为

$$\langle f, g\rangle_{\mathcal{H}_\kappa} := \frac{1}{(2\pi)^{\frac{d}{2}}} \int \frac{\mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}}{2\mathcal{F}[\kappa](\omega)} d\omega, \quad f, g \in \mathcal{H}_\kappa$$

傅里叶函数 $\mathcal{F}[\kappa]$ 基本决定了这个再生核希尔伯特空间

再生核希尔伯特空间

- 正则项

Sobolev 核：考虑$f:[0,1]\to R$，$f(0)=f(1)=0$

$$\|f\|_{\mathcal{H}}^2=\int_0^1\left(f(x)\right)^2+\left(f'(x)\right)^2dx$$

Gaussian 核：

$$\|f\|_{\mathcal{H}}^2=\int_{-\infty}^{+\infty}|f(\omega)|^2\exp\frac{\omega^2\sigma^2}{2}d\omega$$

$\|f\|_{\mathcal{H}}^2$ 越小，需要函数尽量光滑

再生核希尔伯特空间

- 再生核希尔伯特空间(核函数观点)

- 再生核希尔伯特空间(正则化观点)
  - 核岭回归(Kernel ridge regression)
  - 核岭插值(Kernel interpolation)

- 数值计算
  - Nyström近似
  - 随机特征方法

核岭回归

- 核岭回归(Kernel ridge regression)

有噪音训练数据 $\{(x_i, y_i)|i = 1,2 \cdots\cdots n\}$

有限维数据 、无限维函数 $\Longrightarrow$不适定

$$\hat{f} = \text{argmin}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{\text{i=1}}^{n} L\big(x_i, y_i, f(x_i)\big) + \lambda \|f\|^2_{\mathcal{H}_k}$$

当 $L(x, y, y') = (y - y')^2$

$$\hat{f} = \text{argmin}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_k}$$

核岭回归

表示定理(REPRESENTER THEOREM)

$$\hat{f} = \text{argmin}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}_k}^2$$

那么$\hat{f}(x)$能表示为$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i)$。

定义 $k_{xX} = [\kappa(x, x_1)\ \kappa(x, x_2) \cdots \kappa(x, x_n)]$，$K_{XX} = [\kappa(x_i, x_j)]_{i,j=1}^{n}$，$y = [y_1\ y_2 \cdots y_n]^T$。当$\lambda > 0$，一个解为

$$\hat{f}(x) = k_{xX}(K_{XX} + n\lambda I_n)^{-1} y = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i)$$

其中

$$(\alpha_1, \cdots, \alpha_n)^T := (K_{XX} + n\lambda I_n)^{-1} y \in R^n$$

当K_{XX}可逆，该解唯一确定。

高斯回归回顾

- 有噪音观测
  - 条件概率

$$\begin{bmatrix} y \\ f_* \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} K(X,X)+\sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

$$f_* \,|X_*, X, y \sim \mathcal{N}(K(X_*,X)(K(X,X)+\sigma^2 I)^{-1} y,$$
$$K(X_*,X_*) - K(X_*,X)(K(X,X)+\sigma^2 I)^{-1} K(X,X_*)$$

  - 方差和观测无关
  - 预处理计算： Cholesky分解 $(K(X,X)+\sigma^2 I)^{-1}$
  - 在线复杂度：$\mathcal{O}(n^2 n_*)$

核岭回归VS高斯回归

等价性

κ是定义在R^d上的正定核，训练数据为$(x_i, y_i)_{i=1}^n \subset R^d \times R$，考虑

- 高斯回归 $f \sim GP(0, \kappa)$，假设数据误差服从$\mathcal{N}(0, \sigma^2)$
- 正则化系数为λ的核岭回归,

那么当$\sigma^2 = n\lambda$时，高斯回归的期望函数和核岭回归的函数一致 $\bar{m} = \hat{f}$。

核岭回归VS高斯回归

误差估计

先验分布 $f \sim GP(0,\kappa)$ 后验分布 $f \sim GP(\bar{m},\bar{\kappa})$，高斯回归或核岭回归的结果：

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i) = k_{xX}(K_{XX} + \sigma^2 I_n)^{-1} y$$

$$:= w^{\sigma}(x)^T y$$

定义核函数

$$\kappa^{\sigma}(x,y) = \kappa(x,y) + \sigma^2 \delta(x,y)$$

以及相应的再生核希尔伯特空间$\mathcal{H}_{\kappa^{\sigma}}$，我们有误差估计

$$\sqrt{\bar{\kappa}(x,x) + \sigma^2} = \sup_{f \in \mathcal{H}_{\kappa^{\sigma}}, \|f\|_{\mathcal{H}_{\kappa^{\sigma}}} \leq 1} \left(f(x) - \sum_{i=1}^{n} w_i^{\sigma}(x) f(x_i)\right)$$

核岭插值

- 核岭插值(Kernel interpolation)

无噪音训练数据 $\{(x_i, y_i, f(x_i))|i = 1,2 \cdots\cdots n\}$

有限维数据 、无限维函数 $\Longrightarrow$不适定

$$\hat{f} = \text{argmin}_{f \in \mathcal{H}_k} \|f\|_{\mathcal{H}_k} \quad \text{使得} \quad f(x_i) = y_i$$

由表示定理，当K_{XX}可逆

$$\hat{f}(x) = k_{xX} {K_{XX}}^{-1} y = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i)$$

唯一确定，其中

$(\alpha_1, \cdots, \alpha_n)^T := {K_{XX}}^{-1} y \in R^n$

核岭插值VS高斯回归

等价性

κ是定义在R^d上的正定核，训练数据为$(x_i, y_i)_{i=1}^n \subset R^d \times R$，记为$X := (x_1, \cdots, x_n)$，假设$K_{XX}$可逆，考虑

- 高斯回归 $f \sim GP(0, \kappa)$，假设数据没有误差
- 核岭插值

高斯回归的期望函数和核岭插值的函数一致 $\bar{m} = \hat{f}$

核岭插值VS高斯回归

误差估计

先验分布 $f \sim GP(0, \kappa)$ 后验分布 $f \sim GP(\bar{m}, \bar{\kappa})$，假设 K_{XX} 可逆，，高斯回归或核岭插值的结果：

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \kappa(x, x_i) = k_{xX} {K_{XX}}^{-1} y$$
$$:= w(x)^T y$$

我们有

$$\sqrt{\bar{\kappa}(x,x)} = \sup_{f \in \mathcal{H}_\kappa, \|f\|_{\mathcal{H}_\kappa} \leq 1} \left(f(x) - \sum_{i=1}^{n} w_i(x) f(x_i) \right)$$

再生核希尔伯特空间

- 再生核希尔伯特空间(核函数观点)

- 再生核希尔伯特空间(正则化观点)
  - 核岭回归(Kernel ridge regression)
  - 核岭插值(Kernel interpolation)

- 数值计算
  - Nyström近似
  - 随机特征方法

数值计算

- 核岭回归和高斯回归

给定正定核函数κ

$$\hat{f} = \mathrm{argmin}_{f\in\mathcal{H}_k} \frac{1}{n}\sum_{\mathrm{i}=1}^{n}(f(x_i)-y_i)^2 + \lambda\|f\|_{\mathcal{H}_k}$$

$$\hat{f}(x) = k_{xX}(K_{XX}+n\lambda I_n)^{-1}Y = \sum_{i=1}^{n}\alpha_i\kappa(x,x_i)$$

对于数据量大$n \gg 0$，计算量是$\mathcal{O}(n^3)$

数值计算

- 低秩近似

$$K_{XX} = QQ^T \qquad Q \in R^{n \times m}$$

那么（Woodbury 矩阵恒等式）

$$(K_{XX} + n\lambda I_n)^{-1} = (QQ^T + n\lambda I_n)^{-1}$$
$$= \frac{I}{n\lambda} - \frac{1}{n\lambda} Q(n\lambda + Q^T Q)^{-1} Q^T$$

对于数据量大$n \gg 0$，计算量是$\mathcal{O}(nm^2)$

数值计算

- Nyström近似

从原始矩阵中抽取一部分行和列形成一个小的子矩阵，并利用这个子矩阵来构建原矩阵的低秩近似，对于

$$K_{XX} = \begin{bmatrix} K_{mm} & K_{m(n-m)} \\ K_{(n-m)m} & K_{(n-m)(n-m)} \end{bmatrix}$$

作前m步(带主元的)-Cholesky分解

$$\begin{aligned} K_{XX} &= \begin{bmatrix} L_m & 0 \\ K_{(n-m)m}L_m^{-T} & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & * \end{bmatrix} \begin{bmatrix} L_m^T & L_m^{-1}K_{m(n-m)} \\ 0 & I \end{bmatrix} \\ &\approx \begin{bmatrix} L_m \\ K_{(n-m)m}L_m^{-T} \end{bmatrix} \begin{bmatrix} L_m^T & L_m^{-1}K_{m(n-m)} \end{bmatrix} \\ &= K_{(n-m)m}K_{mm}^{-1}K_{m(n-m)} \end{aligned}$$

数值计算

- 随机特征方法

$$\kappa(x, x') = \Psi(x)^{\mathrm{T}}\Psi(\mathrm{x}')$$

其中

$$\Psi(x) = [\psi_1(x)\ \psi_2(x) \cdots \psi_m(x)]^T$$

那么：

$$K_{XX} = QQ^T$$

其中：

$$Q = \begin{bmatrix} \Psi(x_1)^T \\ \Psi(x_2)^T \\ \vdots \\ \Psi(x_n)^T \end{bmatrix} \in R^{n \times m}$$

回顾

Bochner定理

一个定义在R^N中的有平移不变性的实值核函数，那么当κ能表示为

$$\kappa(\tau) = \int_{R^N} e^{2\pi i\omega\cdot\tau} p(\omega) d\omega$$

其中$\mu(\omega)$是有限正测度，我们考虑特殊情况$d\mu(\omega) = p(\omega)d\omega$。

定义$\psi(x,\omega) = e^{2\pi i\omega\cdot x}$，那么

$$\kappa(x-x') = \int_{R^N} e^{2\pi i\omega\cdot(x-x')} p(\omega) d\omega$$
$$= \mathbb{E}_{\omega\sim p(\omega)}[\psi(x,\omega)\psi^*(x',\omega)]$$

数值计算

- 随机特征方法

$$\kappa(x, x') = \Psi(x)^{\mathrm{T}}\Psi(\mathrm{x}')$$

其中

$$\Psi(x) = [\psi_1(x)\ \psi_2(x) \cdots \psi_m(x)]^T$$

随机特征：

$$\kappa(x, x') = \mathbb{E}_{\omega \sim p(\omega)}[\psi(x, \omega)\psi(x', \omega)]$$

$$= \sum_{i=1}^{q} \psi(x, \omega_i)\psi(x', \omega_i)$$

可以定义：

$$\Psi(x) = [\psi(x, \omega_1)\ \psi(x, \omega_2)\ \cdots \psi(x, \omega_{\mathrm{m}})]^T$$

参考文献

参考文献

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, Chapter 6

Wendland, Holger. Scattered data approximation. Vol. 17. Cambridge university press, 2004.

Kanagawa, Motonobu, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. "Gaussian processes and kernel methods: A review on connections and equivalences." arXiv preprint arXiv:1807.02582 (2018).

Rahimi, Ali, and Benjamin Recht. "Random features for large-scale kernel machines." Advances in neural information processing systems 20 (2007).