STAT 598Y STATISTICAL LEARNING THEORY INSTRUCTOR: JIAN ZHANG

LECTURE 12: REPRODUCING KERNEL HILBERT SPACES AND KERNEL METHODS

【数学】RKHS

张楚珩
清华大学 交叉信息院博士在读

RKHS 即 reproducing kernel Hilbert space。

原文传送门

Lever, Guy, and Ronnie Stafford. "Modelling policies in mdps in reproducing kernel hilbert space." Artificial Intelligence and Statistics. 2015.

Reproducing kernel Hilbert space (Lecture notes from Purdue)

特色

本来是想记录一下前一篇文章的，但是由于 RKHS 之前没学过，因此先来看一下 RKHS。由于内容较多，因此单独列出来。

过程

1. Hilbert space

Hilbert space 就是可以定义内积的空间，定义如下。

Definition. A Hilbert space is an inner product space which is also complete and separable [1] with respect to the norm/distance function induced by the inner product. For any $f, g \in \mathcal{H}$ and $\alpha \in \mathbb{R}$, $\langle ., . \rangle$ is an inner product if and only if it satisfies the following conditions:

1. $\langle f, g \rangle = \langle g, f \rangle$;
2. $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$ and $\langle \alpha f, g \rangle = \alpha \langle f, g \rangle$;
3. $\langle f, f \rangle \geq 0$ and $\langle f, h \rangle = 0$ if and only if $f = 0$.

The norm/distance induced by the inner product is defined as $\|f\| = \sqrt{\langle f, f \rangle}$ and $\|f - g\| = \sqrt{\langle f - g, f - g \rangle}$. $\langle ., . \rangle$ is called a semi-inner product if the third condition only says $\langle f, f \rangle \geq 0$. In this case, the induced norm is actually a semi-norm.

[1] A vector space $\mathcal{H}$ is complete if every Cauchy sequence in $\mathcal{H}$ converges to an element in $\mathcal{H}$. A sequence satisfying $\lim_{m,n \to \infty} \|f_n - f_m\| = 0$ is called a Cauchy sequence.

这个定义其实就是讲了内积和范数的定义。由此，Hilbert space 也叫完备内积空间。

Hilbert space 的分解：

A closed linear subspace $\mathcal{G}$ of a Hilbert space $\mathcal{H}$ is also a Hilbert space. The distance between an element $f \in \mathcal{H}$ and $\mathcal{G}$ is defined as $\inf_{g\in\mathcal{G}} \|f-g\|$. Since $\mathcal{G}$ is closed, the infimum can be attained and we have $f_\mathcal{G} \in \mathcal{G}$ such that $\|f - f_\mathcal{G}\| = \inf_{g\in\mathcal{G}} \|f-g\|$. Such $f_\mathcal{G}$ is called the *projection* of f onto $\mathcal{G}$. It can be shown that such $f_\mathcal{G}$ is unique, and $\langle f - f_\mathcal{G}, g\rangle = 0$ for all $g \in \mathcal{G}$. The linear subspace $\mathcal{G}^c = \{f : \langle f, g\rangle = 0, \forall g \in \mathcal{G}\}$ is called the *orthogonal complement* of $\mathcal{G}$. It can be shown that $\mathcal{G}^c$ is also closed and $f = f_\mathcal{G} + f_{\mathcal{G}^c}$ for any $f \in \mathcal{H}$, where $f_\mathcal{G}$ and $f_{\mathcal{G}^c}$ are projections of f onto $\mathcal{G}$ and $\mathcal{G}^c$. The decomposition $f = f_\mathcal{G} + f_{\mathcal{G}^c}$ is called a tensor sum decomposition and is denoted by $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^c$, $\mathcal{G}^c = \mathcal{H} \ominus \mathcal{G}$ or $\mathcal{G} = \mathcal{H} \ominus \mathcal{G}^c$.

A simple example of decomposition would be $\mathcal{H} = \mathbb{R}^2$ and $\mathcal{G} = \{(x, 0) : x \in \mathbb{R}\}$ and $\mathcal{G}^c = \{(0, y) : y \in \mathbb{R}\}$. Any element (x, y) in $\mathcal{H}$ can be decomposed as $(x, y) = (x, 0) + (0, y)$ and this decomposition is unique.

Riesz 表示定理讲的是定义域在 Hilbert 空间上的实值函数都能够表示为内积（kernel）的形式。

Theorem 12-1 (Riesz). *For every contniuous linear functional L in a Hilbert space $\mathcal{H}$, there exists a unique $g_L \in \mathcal{H}$ such that $L(f) = \langle g_L, f\rangle$ for $\forall f \in \mathcal{H}$.*

PROOF.

Define $\mathcal{N}_L = \{f : L(f) = 0\}$ to be the null space of L. Since L is continuous we have $\mathcal{N}_L$ a closed linear subspace. Assume $\mathcal{N}_L \subset \mathcal{H}$ then there exists a nonzero element $g_0 \in \mathcal{H} \ominus \mathcal{N}_L$. We have

$$(L(f))g_0 - (L(g_0))f \in \mathcal{N}_L,$$

and thus

$$\langle (L(f))g_0, (L(g_0))f, g_0\rangle = 0.$$

Thus we get

$$L(f) = \left\langle \frac{L(g_0)}{\langle g_0, g_0\rangle} g_0, f \right\rangle.$$

Hence we take $g_L = (L(g_0))g_0/\langle g_0, g_0\rangle$. If $\mathcal{N}_L = \mathcal{H}$ we simply take $g_L = 0$. If there are g_L and $\tilde{g}_L$ both representers for L then we have $\langle g_L - \tilde{g}_L, f\rangle = 0$ for any $f \in \mathcal{H}$ and thus $\|g_L - \tilde{g}_L\| = 0$ and then $g_L = \tilde{g}_L$.

证明过程如上所示，注意到 $L:\mathcal{H}\to\mathbb{R}$ 。由于函数连续，如果它能取值到零，那么一定有一整个子空间都能取值到零，因此 $\mathcal{N}_f$ 是一个封闭线性子空间。第一个式子：考虑如果 $f\in\mathcal{N}_f$ ，那么第一项为零，第二项在 $\mathcal{N}_f$ 中；如果 $f\notin\mathcal{N}_f$ ，考虑到 $g_0\notin\mathcal{N}_f$ ，因此可以选择 g_0 把不在 $\mathcal{N}_f$ 的部分全部抵消。第二个式子：第一个逗号是减号，写错了。

2. Reproducing kernel Hilbert space

核（kernel）的定义如下，即如果把核看做是一个矩阵的话，它是对称正定矩阵。

Definition. A kernel $k : \mathcal{X} \times \mathcal{X} \mapsto R$ if (1) it is symmetric; (2) it is positive semi definite. I.e. any $x_1, \ldots, x_n$ the gram matrix K is positive semi definite.

Reproducing kernel 和 RKHS 的定义如下

Definition. $k(.,.)$ is a *reproducing kernel* of a Hilbert space $\mathcal{H}$ if for $\forall f \in \mathcal{H}$, we have $f(x) = \langle k(x,.), f(.)\rangle$.

Definition. A RKHS is a Hilbert space $\mathcal{H}$ with a reproducing kernel whose span is dense in $\mathcal{H}$.

An equivalent definition of RKHS woud be "a Hilbert space of functions with all evaluation functinos bounded and linear" or "all evaluation functionals are continuous".

注意到这里的 $f:\mathcal{X}\to\mathbb{R}$ ，等价定义之间能够通过前面的 Riesz 表示定理联系起来。

Notes 里面给出了如下定理

Theorem 12-2 (Mercer's). *Let $(\mathcal{X},\mu)$ be a finite measure space and $k\in L_\infty(\mathcal{X}\times\mathcal{X},\mu\times\mu)$ be a kernel such that $T_k:L_2(\mathcal{X},\mu)\mapsto L_2(\mathcal{X},\mu)$ is positive definite, i.e. $\int k(x,z)f(x)f(z)d\mu(x)d\mu(z)\geq 0$ for all $f\in L_2(\mathcal{X},\mu)$. Let $\phi_i\in L_2(\mathcal{X},\mu)$ be the normalized eigenfunctions of T_k associated with the eigenvalues $\lambda_i\geq 0$. Then*
(1) The eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ are absolutely summable;
(2) $k(x,z)=\sum_{i=1}^{\infty}\lambda_i\phi_i(x)\phi_i(z)$ holds the series converges absolutely and uniformly.

看起来比较绕，可以这样来理解：把 kernel 看做是一个无限维的矩阵（因为定义域 $\mathcal{X}$ 中的元素可能是无限多的），如果该矩阵正定对称，就可以对这个矩阵做特征值分解，这样就能得到无限多个特征值和特征向量，即上述定理中 (2) 有了的形式。

一个正定对称的 kernel 可以确定一个 Hilbert 空间

$$\mathcal{H}=\{f:f(x)=\sum_i\alpha_i\sqrt{\lambda_i}\phi_i(x),\|f\|_{\mathcal{H}}<\infty\}$$

可以选取正交归一的 $[\sqrt{\lambda_1}\phi_1,\sqrt{\lambda_2}\phi_2,\cdots]^T$ 作为这个 Hilbert 空间的基底，这样 Hilbert 空间上的任意实值函数都可以看做是这组基底的线性组合，即

$$f(x)=\sum_i f_i\sqrt{\lambda_i}\phi_i(x),\quad f\in\mathcal{H}$$

首先，我们可以看到这个 Hilbert 空间是一个 RKHS，因为它满足

$$\langle k(x,\cdot),f(\cdot)\rangle_{\mathcal{H}}=\langle\sum_i\lambda_i\phi_i(x)\phi_i(\cdot),\sum_i f_i\sqrt{\lambda_i}\phi_i(\cdot)\rangle_{\mathcal{H}}=\sum_i f_i\sqrt{\lambda_i}\phi_i(\cdot)=f(\cdot)$$

当我们把它们表示为如下形式时： $\Phi(f)=[f_1,f_2,\cdots]^T\in\mathbb{R}^\infty$ ， $\Phi(k(x,\cdot))=[\sqrt{\lambda_1}\phi_1(x),\sqrt{\lambda_1}\phi_1(x),\cdots]$ ，函数就能表示为一个简单的形式： $f(x)=\Phi(f)^T\Phi(k(x,\cdot))$ 。

其次，该空间上的范数和内积都可以直接使用使用表示 Φ 来得到，即

$$\langle f,g\rangle_{\mathcal{H}}=\sum_i f_i g_i=\Phi(f)^T\Phi(g),\quad \|f\|_{\mathcal{H}}^2=\sum_i f_i^2=\|\Phi(f)\|_2^2$$

最后，由于 $k(x,y)=\langle k(x,\cdot),k(y,\cdot)\rangle_{\mathcal{H}}=\Phi(k(x,\cdot))^T\Phi(k(y,\cdot))$ ，即我们只需要知道 kernel function ，而不需要知道具体的特征向量 Φ 映射。这也称作 kernel trick。

3. Representer theorem

下面这个定理说明如果解一个定义在若干个点上的泛函优化问题，那么最优解一定在这些点的 kernel function 张成的空间上。可以把下面的优化问题想象成机器学习里面的 empirical risk minimization。

Theorem 12-3 (Representer). *Given a reproducing kernel k and let $\mathcal{H}$ be the corresponding RKHS. Then for a function $L:\mathbb{R}^n\mapsto\mathbb{R}$ and non-decreasing function $\Omega:\mathbb{R}\mapsto\mathbb{R}$, the solution of the optimization problem*

$$\min_{f\in\mathcal{H}}J(f)=\min_{f\in\mathcal{H}}\left\{L(f(x_1),\ldots,f(x_n))+\Omega(\|f\|_{\mathcal{H}}^2)\right\}$$

can be expressed as

$$f^*=\sum_{i=1}^n\alpha_i k(x_i,.).$$

Furthermore, if $\Omega(.)$ is strictly increasing, then all solutions have this form.

证明思路主要就是要利用 RKHS 的再生性（reproducing property）

PROOF.

Define the subspace $\mathcal{G}$ to be the span of

$$\mathsf{span}\{k(x_i,.), 1 \le i \le n\}.$$

Decompose f as $f = f_{\mathcal{G}} + f_{\mathcal{G}^c}$. We have

$$\|f\|_{\mathcal{H}} = \|f_{\mathcal{G}}\|_{\mathcal{G}} + \|f_{\mathcal{G}^c}\|_{\mathcal{H}}$$

by orthogonality of $\mathcal{G}$ with $\mathcal{G}^c$. Since Ω is non-decreasing, we have

$$\Omega(\|f\|_{\mathcal{H}}^2) \ge \Omega(\|f_{\mathcal{G}}\|_{\mathcal{H}}^2).$$

On the other hand, since the kernel k has the reproducing property, we have

$$\begin{aligned} f(x_i) &= \langle f, k(x_i,.)\rangle \\ &= \langle f_{\mathcal{G}}, k(x_i,.)\rangle + \langle f_{\mathcal{G}^c}, k(x_i,.)\rangle \\ &= \langle f_{\mathcal{G}}, k(x_i,.)\rangle \\ &= f_{\mathcal{G}}(x_i). \end{aligned}$$

So this implies that $L(f(x_1),\ldots,f(x_n)) = L(f_{\mathcal{G}}(x_1),\ldots,f_{\mathcal{G}}(x_n))$, i.e. the first component of the optimization objective only depends on the projection of f onto $\mathcal{G}$ which is the span of $k(x_i,.)'s$. Since $\Omega(\|f\|_{\mathcal{H}}^2) \ge \Omega(\|f_{\mathcal{G}}\|_{\mathcal{H}}^2)$, we have the minimizer can be expressed as $f^*(.) = \sum_{i=1}^n \alpha_i k(x_i,.)$. If $\Omega(.)$ is strictly increasing, then $f_{\mathcal{G}^c}$ must be zero and all minimizers must take the above form.

4. Kernel 的例子

这里前面列出来的几种 kernel 都是比较常见的 kernel，尤其是 RBF kernel，在深度学习和强化学习里面都用的比较多。

Some simple examples of kernel:

- Linear kernel: $k(x,z) = x^T z$ or more generally, $k(x,z) = x^T B z$ for $B \succcurlyeq 0$.
- Polynomial kernel: $k(x,z) = (x^T z + c)^d$ where $c \ge 0$ and $d \in \mathbb{N}_+$.
- RBF kernel: $k(x,z) = \exp(-\gamma\|x-z\|^2)$.

We could also construct kernels based on simple ones. For instance, we have kernels (it can be shown that $k(.,.)$ satisfies the conditions of a kernel):

- $k(x,z) = \sum_i \alpha_i k_i(x,z)$ where $\alpha_i \ge 0$ and $k_i(.,.)$ are kernels;
- $k(x,z) = k_1(x,z)k_2(x,z)$;
- $k(x,z) = \exp(k_1(x,z))$;
- $k(x,z) = P(k_1(x,z))$ where $P(t)$ is a polynomial of t with nonnegative coefficients.

5. Rademacher Average

考虑一族比较平滑的函数 $\mathcal{F}_t = \{f : f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq t\}$ ，给定一些样本点，定义这些样本点在这一族函数上的（某种）平均值为 Rademacher average。以下定理给出了它的上界。

Theorem. *Let $\mathcal{H}$ be a RKHS with kernel k, and let $K \in \mathbb{R}^{n\times n}$ so that $K_{ij} = k(x_i, x_j)$. Define $\mathcal{F}_t = \{f : f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq t\}$. Then we have*

$$\hat{\mathcal{R}}_n(\mathcal{F}_t) := \mathbb{E}\left[\sup_{f\in\mathcal{F}_t} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i) | X_1, \ldots, X_n\right] \leq \frac{t}{n}\sqrt{\mathrm{trace}(K)}$$

and

$$\mathcal{R}_n(\mathcal{F}_t) \leq \frac{t}{\sqrt{n}}\sqrt{\sum_{i=1}^{\infty}\lambda_i}$$

where λ_i's are the eigenvalues of the operator $T_k : f \mapsto \int k(.,x)f(x)dP(x)$.

其中 ϵ_i 是 n 个独立采样的 noise，它们满足均值为 0 ，方差为 1 。证明如下

By the reproducing property we have

$$\begin{aligned}
\sup_{f\in\mathcal{F}_t} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i) &= \sup_{f\in\mathcal{F}_t} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \langle k(x_i,.), f\rangle \\
&= \sup_{\|f\|_{\mathcal{H}}\leq t} \left\langle \frac{1}{n}\sum_{i=1}^{n}\epsilon_i k(x_i,.), f\right\rangle \\
&\leq t\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i k(x_i,.)\right\| \\
&= t\sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n}\epsilon_i\epsilon_j k(x_i,x_j)}.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{F}_t) &= \mathbb{E}\left[\frac{t}{n}\sqrt{\sum_{i,j=1}^{n}\epsilon_i\epsilon_j k(x_i,x_j)} | X_1,\ldots,X_n\right] \\
&\leq \frac{t}{n}\sqrt{\mathbb{E}\left[\sum_{i,j=1}^{n}\epsilon_i\epsilon_j k(x_i,x_j)|X_1,\ldots,X_n\right]} \\
&= \frac{t}{n}\sqrt{\sum_{i=1}^{n}k(x_i,x_i)} \\
&= \frac{t}{n}\sqrt{\mathrm{trace}(K)},
\end{aligned}$$

where we used the property that $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{V}[\epsilon_i] = 1$ and Jensen's inequality. Since $k(x,x) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(x)$, where ϕ_i's are an orthonomral basis, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_t) &= \mathbb{E}[\hat{\mathcal{R}}_n(\mathcal{F}_t)] \\
&\leq \frac{t}{\sqrt{n}}\mathbb{E}\sqrt{\frac{1}{n}\sum_{i=1}^{n} k(x_i,x_i)} \\
&\leq \frac{t}{\sqrt{n}}\sqrt{\mathbb{E}[k(X,X)]} \\
&\leq \frac{t}{\sqrt{n}}\sqrt{\sum_{i=1}^{\infty}\lambda_i}.
\end{aligned}$$

6. Frechet derivative

普通的导数可以写成

$$\begin{aligned}
f'(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} \quad &\Leftrightarrow \quad \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h} - f'(x)\right] = 0 \\
&\Leftrightarrow \quad \lim_{h\to 0}\frac{f(x+h)-f(x)-f'(x)h}{h} = 0 \qquad (2)
\end{aligned}$$

对于一个定义在 Hilbert 空间上的映射 $f:\mathcal{H}\to\mathbb{R}$ （可以认为 Hilbert 空间中的一点代表一个函数），也可以求导，其导数也是一个 Hilbert 空间上的导数。因此可以定义下面的这个 $Df|_a$ 算子来作为导数，即 Frechet derivative。

The Fréchet derivative is the derivative for functions on a Banach space. Let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces, and $\mathcal{U} \subset \mathcal{V}$ be an open subset of $\mathcal{V}$. A function $f : \mathcal{U} \to \mathcal{W}$ is called Fréchet differentiable at $x \in \mathcal{U}$ if there exists a bounded linear operator $Df|_x : \mathcal{V} \to \mathcal{W}$ such that,

$$\lim_{r \to 0} \frac{\|f(x+r) - f(x) - Df|_x(r)\|_{\mathcal{W}}}{\|r\|_{\mathcal{V}}} = 0.$$

发布于 2019-06-29

数学分析

文章被以下专栏收录

强化学习前沿
读呀读paper

进入专栏