

LECTURE 12: REPRODUCING KERNEL HILBERT SPACES AND KERNEL METHODS

We first define Hilbert space and then introduce the concept of Reproducing Kernel Hilbert Space (RKHS) which plays an important role in machine learning.

Definition. A Hilbert space is an inner product space which is also complete and separable¹ with respect to the norm/distance function induced by the inner product. For any $f, g \in \mathcal{H}$ and $\alpha \in \mathbb{R}$, $\langle \cdot, \cdot \rangle$ 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 $\|f - g\| = \sqrt{\langle f - g, f - g \rangle}$. $\langle \cdot, \cdot \rangle$ 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.

Examples of Hilbert space includes:

1. \mathbb{R}^n with $\langle a, b \rangle = a^T b$;
2. ℓ_2 space of square summable sequence with inner product $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$;
3. The space of L_2 square integrable functions with inner product $\langle f, g \rangle = \int f(x)g(x)dx$.

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.

Theorem 12-1 (Riesz). For every continuous 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 two 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 $g_L = \tilde{g}_L$.

□.

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

Reproducing Kernel Hilbert Space

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, \dots, x_n the gram matrix K is positive semi definite.

Properties: (1) $k(x, x) \geq 0$; (2) $k(x, z) \leq \sqrt{K(x, x)K(z, z)}$.

There are a couple of ways to define RKHS which are equivalent.

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

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 would be “a Hilbert space of functions with all evaluation functionals bounded and linear” or “all evaluation functionals are continuous”.

Theorem 12-2 (Mercer’s). Let (\mathcal{X}, μ) 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.

We can construct a RKHS as the completed space of the span of eigenfunctions defined by the kernel:

$$\mathcal{H} = \overline{\left\{ f : f(x) = \sum_i \alpha_i \phi_i(x) \text{ s.t. } \|f\|_{\mathcal{H}} < \infty \right\}}$$

Given $f = \sum_i \alpha_i \phi_i$ and $g = \sum_i \beta_i \phi_i$, the inner product and the norm induced by the inner product are defined as

$$\langle f, g \rangle_{\mathcal{H}} = \left\langle \sum_i \alpha_i \phi_i(x), \sum_i \beta_i \phi_i(x) \right\rangle_{\mathcal{H}} = \sum_i \frac{\alpha_i \beta_i}{\lambda_i}$$

and

$$\|f\|_{\mathcal{H}}^2 = \left\langle \sum_i \alpha_i \phi_i(x), \sum_i \alpha_i \phi_i(x) \right\rangle_{\mathcal{H}} = \sum_i \frac{\alpha_i^2}{\lambda_i}.$$

It is easy to see that the representer property holds:

$$\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = \left\langle \sum_i \alpha_i \phi_i(\cdot), \sum_i \lambda_i \phi_i(x) \phi_i(\cdot) \right\rangle_{\mathcal{H}} = \sum_i \frac{\alpha_i \lambda_i \phi_i(x)}{\lambda_i} = f(x).$$

The RKHS concept can be utilized in SVM and other kernel machines which is known as the kernel trick. Given the eigenvalues λ_i ’s and eigenfunctions ϕ_i ’s of a reproducing kernel $k(\cdot, \cdot)$, we can map the $x \in \mathbb{R}^p$ into a higher dimensional feature space:

$$x \mapsto \Phi(x) = \left[\sqrt{\lambda_1} \phi_1(x), \dots, \sqrt{\lambda_i} \phi_i(x), \dots \right].$$

The dimensionality of the feature vector $\Phi(x)$ is the same as the number of nonzero eigenvalues of $k(\cdot, \cdot)$, which could be of infinite dimensional. By Mercer’s theorem, the standard ℓ_2 inner product between any two feature vectors $\Phi(x)$ and $\Phi(z)$ can now be computed by the reproducing kernel since

$$k(x, z) = \langle \Phi(x), \Phi(z) \rangle_{\ell_2}.$$

Representer Theorem

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*

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

can be expressed as

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

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

PROOF.

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

$$\text{span}\{k(x_i, \cdot), 1 \leq i \leq n\}.$$

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

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

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

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

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

□

Examples of Kernel

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 $c \geq 0$ 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 $k(\cdot, \cdot)$ satisfies the conditions of a kernel):

- $k(x, z) = \sum_i \alpha_i k_i(x, z)$ where $\alpha_i \geq 0$ and $k_i(\cdot, \cdot)$ 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.

Rademacher Average

We next present a result which computes the upper bound of the Rademacher average of a function class which is a ball in the RKHS. Consider the following learning problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}},$$

where \mathcal{H} is a RKHS with kernel k . The optimization problem is equivalent to

$$\min_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq t} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$$

for some properly chosen $t > 0$. So we would like to invest the function class $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq t\}$'s Rademacher average.

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*

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

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 $T_k : f \mapsto \int k(\cdot, x) f(x) dP(x)$.

PROOF.

By the reproducing property we have

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

Therefore we have

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

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

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

□