STAT538A: Statistical Learning: the Way of the Kernel	Winter 2020
Lecture 2	
Lecturer: Zaid Harchaoui	*Scribe: Lang Liu*

1 Reproducing Kernel Hilbert Space

Theorem 12 (Aronszajn, 1950)**.** *Let $\mathcal{X}$ be a metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PSD kernel, there exists a unique Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of functions on $\mathcal{X}$ satisfying the following conditions:*

(i) for all $x \in \mathcal{X}, \phi(x) := k(x, \cdot) \in \mathcal{H}$,

(ii) the span of the set $\{\phi(x) | x \in \mathcal{X}\}$ is dense in $\mathcal{H}$, and

(iii) for all $f \in \mathcal{H}$ and $x \in \mathcal{X}, f(x) = \langle \phi(x), f\rangle_{\mathcal{H}}$.

In particular, $\mathcal{H}$ is an RKHS.

Proof The proof is divided into several steps.

Step 1. Construct an inner product space $\mathcal{H}_0 = \operatorname{span}\{\phi(x) : x \in \mathcal{X}\}$ satisfying the conditions (i)-(iii). Recall that

$$\operatorname{span}\{\phi(x) : x \in \mathcal{X}\} = \left\{ f = \sum_{i=1}^{s} \alpha_i \phi(x_i) : x_1 \ldots, x_s \in \mathcal{X}, \alpha_1, \ldots, \alpha_s \in \mathbb{R}, \text{and } s \in \mathbb{N} \right\},$$

it is easy to verify that $\mathcal{H}_0$ is a vector space, and conditions (i) and (ii) holds on $\mathcal{H}_0$. Let $\langle\cdot,\cdot\rangle_{\mathcal{H}_0} : \mathcal{H}_0 \times \mathcal{H}_0 \to \mathbb{R}$ be a bivariate functional defined as

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{\substack{1\le i\le s\\ 1\le j\le r}} \alpha_i\beta_j k(x_i, y_j), \quad \text{for } f = \sum_{i=1}^{s} \alpha_i\phi(x_i) \text{ and } g = \sum_{j=1}^{r} \beta_j\phi(y_j).$$

In particular, we have

$$\langle f, \phi(x)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{s} \alpha_i k(x_i, x) = f(x),$$

that is, the condition (iii) holds on $\mathcal{H}_0$. Next, we will show that $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$ is an inner product on $\mathcal{H}_0$. We only verify the positive definiteness here. Given any $f := \sum_{i=1}^{s} \alpha_i\phi(x_i) \in \mathcal{H}_0$, we have

$$\langle f, f\rangle_{\mathcal{H}_0} = \sum_{i,j=1}^{s} \alpha_i\alpha_j k(x_i, x_j) \ge 0$$

by the positive semi-definiteness of the kernel. Moreover, if $\langle f, f\rangle_{\mathcal{H}_0} = 0$ for some $f \in \mathcal{H}_0$, then for every $x \in \mathcal{X}$ and $\epsilon \in \mathbb{R}$, we obtain

$$\langle f + \epsilon\phi(x), f + \epsilon\phi(x)\rangle_{\mathcal{H}_0} = \langle f, f\rangle_{\mathcal{H}_0} + 2\epsilon\langle f, \phi(x)\rangle_{\mathcal{H}_0} + \epsilon^2 k(x, x) \ge 0.$$

This implies $4\langle f, \phi(x)\rangle^2_{\mathcal{H}_0} - 4\langle f, f\rangle_{\mathcal{H}_0} k(x, x) \le 0$. Notice that $\langle f, f\rangle_{\mathcal{H}_0} = 0$, it holds that $f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}_0} = 0$. Since $x \in \mathcal{X}$ is arbitrary, we have $f \equiv 0$ on $\mathcal{X}$.

Step 2. Prove that the completion of $\mathcal{H}_0$ is a Hilbert space. Let $\{f_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathcal{H}_0$, *i.e.*, $\|f_n - f_m\|_{\mathcal{H}_0} \to 0$ as $n, m \to \infty$. Then for any $x \in \mathcal{X}$, by Cauchy-Schwarz inequality, we have

$$|f_n(x) - f_m(x)| = \left|\langle f_n - f_m, \phi(x)\rangle_{\mathcal{H}_0}\right| \leq \|f_n - f_m\|_{\mathcal{H}_0} \|\phi(x)\|_{\mathcal{H}_0} = \|f_n - f_m\|_{\mathcal{H}_0} \sqrt{k(x,x)} \to 0$$

as $n, m \to \infty$. It follows that $\{f_n(x)\}_{n\in\mathbb{N}}$ is also a Cauchy sequence, and thus the limit $f(x) := \lim_{n\to\infty} f_n(x)$ exists. Therefore, we may view f as the limit of the sequence $\{f_n\}_{n\in\mathbb{N}}$ *w.r.t.* the norm $\|\cdot\|_{\mathcal{H}_0}$. Now let $\mathcal{H}$ be the completion of $\mathcal{H}_0$ by adding all such limits of Cauchy sequences. By the continuity of the inner product, we can define a bivariate function $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ as follows: $\forall f, g \in \mathcal{H}$, there exist two sequences $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_0$ and $\{g_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_0$ such that $\lim_{n\to\infty} f_n = f$ and $\lim_{n\to\infty} g_n = g$, then we define

$$\langle f, g\rangle_{\mathcal{H}} := \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{H}_0}.$$

Let us verify the limit exists. Firstly, note that

$$\left|\|f_n\|_{\mathcal{H}_0} - \|f_m\|_{\mathcal{H}_0}\right| \leq \|f_n - f_m\|_{\mathcal{H}_0} \to 0, \quad \text{as } n, m \to \infty,$$

so there exists a constant $c_f > 0$ such that $\lim_{n\to\infty} \|f_n\|_{\mathcal{H}_0} = c_f$. Similarly, there exists a constant $c_g > 0$ such that $\lim_{n\to\infty} \|g_n\|_{\mathcal{H}_0} = c_g$. This implies $(\langle f_n, g_n\rangle_{\mathcal{H}_0})_{n\geq 1}$ is a Cauchy sequence in $\mathbb{R}$ since

$$\begin{aligned}
\left|\langle f_n, g_n\rangle_{\mathcal{H}_0} - \langle f_m, g_m\rangle_{\mathcal{H}_0}\right| &\leq \left|\langle f_n, g_n - g_m\rangle_{\mathcal{H}_0}\right| + \left|\langle f_n - f_m, g_m\rangle_{\mathcal{H}_0}\right| \\
&\leq \|f_n\|_{\mathcal{H}_0} \|g_n - g_m\|_{\mathcal{H}_0} + \|f_n - f_m\|_{\mathcal{H}_0} \|g_m\|_{\mathcal{H}_0} \\
&\to 0 \quad \text{as } n, m \to \infty.
\end{aligned}$$

Finally, it is not hard to verify that $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is an inner product. In fact, for $f, g, h \in \mathcal{H}$ associated with respective Cauchy sequences $\{f_n\}$, $\{g_n\}$ and $\{h_n\}$, we have

$$\begin{aligned}
\langle f, f\rangle_{\mathcal{H}} &= \lim_{n\to\infty} \langle f_n, f_n\rangle_{\mathcal{H}_0} \geq 0, \\
\langle f, ag\rangle_{\mathcal{H}} &= \lim_{n\to\infty} \langle f_n, ag_n\rangle_{\mathcal{H}_0} = a\langle f, g\rangle_{\mathcal{H}}, \\
\langle f, g + h\rangle_{\mathcal{H}} &= \lim_{n\to\infty} \langle f_n, g_n + h_n\rangle_{\mathcal{H}_0} = \lim_{n\to\infty} (\langle f_n, g_n\rangle_{\mathcal{H}_0} + \langle f_n, h_n\rangle_{\mathcal{H}_0}) = \langle f, g\rangle_{\mathcal{H}} + \langle f, h\rangle_{\mathcal{H}}.
\end{aligned}$$

Therefore, $\mathcal{H}$ is a Hilbert space of functions defined on $\mathcal{X}$.

Step 3. Verify that $\mathcal{H}$ satisfies the conditions (i)-(iii) and is unique. By the definition of completion, conditions (i) and (ii) are automatically fulfilled. As for (iii), it follows from

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \langle \phi(x), f_n\rangle_{\mathcal{H}_0} = \langle \phi(x), f\rangle_{\mathcal{H}}.$$

In order to prove the uniqueness, we assume $\mathcal{G}$ is another Hilbert space satisfying these conditions. We want to show that

$$\mathcal{G} = \mathcal{H} \quad \text{and} \quad \langle\ ,\ \rangle_{\mathcal{G}} = \langle\ ,\ \rangle_{\mathcal{H}}. \tag{1}$$

We first observe that $\mathcal{H}_0 \subset \mathcal{G}$ because of the condition (ii). Also, according to condition (iii), for any $x, y \in \mathcal{X}$, $\langle \phi(x), \phi(y)\rangle_{\mathcal{G}} = k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$. By linearity of the inner product, for every $f, g \in \mathcal{H}_0$, we have $\langle f, g\rangle_{\mathcal{G}} = \langle f, g\rangle_{\mathcal{H}}$. Since both $\mathcal{G}$ and $\mathcal{H}$ are completions of $\mathcal{H}_0$, (1) follows from the uniqueness of the completion. ■

Corollary 13. *A function $k : \mathcal{X} \times \mathcal{X} :\to \mathbb{R}$ is a PSD kernel if and only if there is a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$, such that $\forall x, x' \in \mathcal{X}, k(x, x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{H}}$. We refer to the map ϕ as a* feature map.

Proof "$\Rightarrow$": if k is a PSD kernel, then by Aronszajn's Theorem we know that there exists an RKHS $\mathcal{H}$, and, in particular, a Hilbert space of functions, such that $\phi(x) := k(x, \cdot) \in \mathcal{H}$ and $k(x, x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{H}}$ for every $x, x' \in \mathcal{X}$.

"⇐": it follows directly from the properties of inner product. ■

Remark One PSD kernel may correspond to multiple feature maps. For instance, let $\mathcal{X} = \mathbb{R}$ and $k(x, x') = xx'$. We consider the Hilbert space $\mathcal{H} := \mathbb{R}^2$, then both

$$\phi_1(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \phi_2(x) = \begin{bmatrix} \sqrt{2}x/\sqrt{3} \\ x/\sqrt{3} \end{bmatrix},$$

are feature maps of k.

Given any such feature map, we can define a PSD kernel by $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. With an appropriate choice of ϕ, the kernel k usually admits a closed-form expression. Therefore, we can compute the inner product between two embedded data points $(\phi(x), \phi(x'))$ without actually working in the Hilbert space $\mathcal{H}$. This allows the computations in kernel methods to be tractable, and we usually call it the "kernel trick".

Next, we investigate some properties of an RKHS.

Definition 14 (Evaluation functional)**.** *Let $\mathcal{H}$ be a Hilbert space of functions defined on $\mathcal{X}$. For any $x \in \mathcal{X}$, the* evaluation functional *at x is defined to be a linear operator $k_x : \mathcal{H} \to \mathbb{R}$ such that $k_x(f) = f(x)$ for all $f \in \mathcal{H}$.*

It is straightforward to show that the inner product in a Hilbert space is a continuous linear operator. The next theorem verifies that the converse is also true.

Theorem 15 (Riesz Representation Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space, and L be a continuous linear operator. Then there exists some $g \in \mathcal{H}$ such that*

$$L(f) = \langle f, g \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

And we refer to g as the representer *of the operator L.*

The Riesz representation theorem provides a characterization of RKHS in terms of evaluation functionals.

Proposition 16. *Let $\mathcal{H}$ be a Hilbert space of functions defined on $\mathcal{X}$. Then $\mathcal{H}$ is an RKHS if and only if for all $x \in \mathcal{X}$, the evaluation functional k_x is continuous.*

Proof "⇒": Let k be the reproducing kernel associated with $\mathcal{H}$. Given an arbitrary $x \in \mathcal{X}$, we want to show that for any sequence $(f_n)_{n \in N}$ in $\mathcal{H}$ such that $f_n \to f$ as $n \to \infty$, we have $k_x(f_n) = f_n(x) \to k_x(f) = f(x)$ as $n \to \infty$. In fact,

$$|f_n(x) - f(x)| = |\langle f_n, k(x, \cdot) \rangle_{\mathcal{H}} - \langle f, k(x, \cdot) \rangle_{\mathcal{H}}| = |\langle f_n - f, k(x, \cdot) \rangle_{\mathcal{H}}| \leq \|f_n - f\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}},$$

where the last inequality is due to the Cauchy-Schwarz inequality. Furthermore, we have

$$\|k(x, \cdot)\|_{\mathcal{H}} = \langle k(x, \cdot), k(x, \cdot) \rangle_{\mathcal{H}}^{1/2} = k(x, x)^{1/2} < \infty,$$

and it follows that $f_n(x) \to f(x)$ as $n \to \infty$.

"⇐": Note that k_x is linear, then by Riesz representation theorem, $\exists g_x \in \mathcal{H}$, s.t. , $f(x) = k_x(f) = \langle f, g_x \rangle_{\mathcal{H}}$ for every $f \in \mathcal{H}$. For any $x, y \in \mathcal{X}$, let $k(x, y) = k_y(g_x) = \langle g_x, g_y \rangle_{\mathcal{H}} = g_x(y)$, then $k(x, \cdot) = g_x \in \mathcal{H}$ and k is a reproducing kernel: for all $f \in \mathcal{H}$,

$$\langle k(x, \cdot), f \rangle_{\mathcal{H}} = \langle g_x, f \rangle_{\mathcal{H}} = f(x).$$

It follows that $\mathcal{H}$ is an RKHS. ■

Proposition 17. *Let $\mathcal{H}$ be an RKHS, then it admits a unique reproducing kernel, which is also a PSD kernel. Conversely, if k is a PSD kernel, then there exists an RKHS $\mathcal{H}$ such that the reproducing property holds.*

Proof "⇒": Let $k(\cdot,\cdot)$ and $h(\cdot,\cdot)$ be two reproducing kernels associated with $\mathcal{H}$. Then for any $x \in \mathcal{X}$, we have:

$$\|k(x,\cdot)-h(x,\cdot)\|_{\mathcal{H}}^2 = \langle k(x,\cdot)-h(x,\cdot), k(x,\cdot)-h(x,\cdot)\rangle_{\mathcal{H}} = k(x,x)-h(x,x)-k(x,x)+h(x,x) = 0.$$

It follows that k and h are identical. Moreover, since $k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}}$, we know that k is PSD by the positive definiteness of inner product.

"⇐": If $k(\cdot,\cdot)$ is a PSD kernel, then there exists an RKHS $\mathcal{H}$ with k being its reproducing kernel by Aronszajn's theorem. ■

2 Applications

One of the most attractive feature of reproducing kernel is its reproducing property. It allows us to perform efficient computations while still gaining the advantages of high-dimensional features. In the following, we show how one can apply the kernel trick to compute the distance between two functions.

To begin with, we consider two objects $x_1, x_2 \in \mathcal{X}$. They are mapped to two features $\phi(x_1), \phi(x_2) \in \mathcal{H}$, so it is natural to define a distance between two objects as the distance between their features in $\mathcal{H}$:

$$d_{\mathcal{H}}(x_1,x_2) := \|\phi(x_1)-\phi(x_2)\|_{\mathcal{H}}.$$

At first sight of this definition, it seems necessary to explicitly compute $\phi(x_1), \phi(x_2)$ before computing this distance. However, this distance can actually be expressed in terms of inner products in $\mathcal{H}$, that is,

$$\|\phi(x_1)-\phi(x_2)\|_{\mathcal{H}}^2 = \langle\phi(x_1),\phi(x_1)\rangle_{\mathcal{H}} + \langle\phi(x_2),\phi(x_2)\rangle_{\mathcal{H}} - 2\langle\phi(x_1),\phi(x_2)\rangle_{\mathcal{H}}.$$

Applying the kernel in the above equality, we obtain the distance in terms of the kernel:

$$d_{\mathcal{H}}(x_1,x_2) = \sqrt{k(x_1,x_1)+k(x_2,x_2)-2k(x_1,x_2)}.$$

The effect of kernel is easily understood in this example: it is possible to perform operations implicitly in feature space. This is of utmost importance for kernels that are easy to compute directly, but correspond to a complex feature space. For example, for Gaussian RBF kernel $k(x,y) = \exp(-\|x-y\|^2/\sigma^2)$ in which $x, y \in \mathbb{R}^d$, the square distance $d_{\mathcal{H}}(x,y)^2 = 2[1-\exp(-\|x-y\|^2/\sigma^2)]$.

Then, we consider the following slightly more general problem. Let $S_X = \{x_1,\ldots,x_n\}$ be a fixed finite set of objects in $\mathcal{X}$. Is it possible to access how close a new object x is to the set of objects S_X? Having mapped the set S_X and the object x into the feature space with the feature map ϕ, a natural way to measure the distance between the object x and the set S_X is to define it as the distance between $\phi(x)$ and the centroid of S_X in the feature space, where the centroid is defined as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \phi(x_i).$$

In general, there is no reason to believe $\hat{\mu}$ should be the image of the feature map ϕ at an object $x' \in \mathcal{X}$, but it is well defined as an element of $\mathcal{H}$. Then, we can define the distance between x and S_X as follows:

$$d_{\mathcal{H}}(x,S_X) = \|\phi(x)-\hat{\mu}\|_{\mathcal{H}} = \left\|\phi(x) - \frac{1}{n}\sum_{i=1}^n \phi(x_i)\right\|_{\mathcal{H}}$$

Expanding the above distance in terms of inner products in the feature space and using the kernel trick, we have

$$d_{\mathcal{H}}(x,S_X) = \sqrt{k(x,x) - \frac{2}{n}\sum_{i=1}^n k(x,x_i) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,x_j)}$$

This shows that the distance between x and S_X can be computed entirely from the kernels between pairs of objects in $\{x\} \cup S_X$, even though it is defined as a distance in the feature space between $\phi(x)$ and $\hat{\mu}$ that $\phi^{-1}(\hat{\mu})$ does not necessarily exist in $\mathcal{X}$.

As a further generalization, we can easily define the following function as a distance between two set of objects S_X and S_Y,

$$d_{\mathcal{H}}(S_X, S_Y) = \sqrt{\frac{1}{|S_X|^2} \sum_{x, x' \in S_X} k(x, x') + \frac{1}{|S_Y|^2} \sum_{y, y' \in S_Y} k(y, y') - \frac{2}{|S_X||S_Y|} \sum_{x \in S_X, y \in S_Y} k(x, y)},$$

where $|S_X|$ and $|S_Y|$ are the cardinalities of S_X and S_Y, respectively.