Lecture V.1 – Build your own RKHS in 4 easy steps

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

STAT/BIOST 527
Spring 2023

Outline

1. RKHS – why bother?
2. From kernel $K()$ to Reproducing Kernel Hilbert Space (RKHS)
3. Properties of RKHS's

Reading AoNPS Ch.: , HTF Ch.:

RKHS – why bother?

- **Practical goal** = learning a predictor $f : \mathsf{X} \to \mathbb{R}$
- If f depends on a kernel $K()$ and we ensure $K() \succ 0$, then f will be guaranteed to have some **nice properties** and be (statistically) safe to use

- RKHS $\Leftrightarrow$ $K() \succ 0$
- So what does RKHS give us?

$$
\begin{aligned}
x \in \mathsf{X} \quad &\leftrightarrow \quad \phi(x) \in \mathcal{H}, \text{ with } \phi(x) \equiv K_x() \text{ the feature map} && (1)\\
f : \mathsf{X} \to \mathbb{R} \text{ non-linear} \quad &\leftrightarrow \quad \text{linear functional } f \text{ on } \mathcal{H} \equiv f \in \mathcal{H} && (2)\\
f(x) \quad &\leftrightarrow \quad f^T \phi(x) && (3)\\
\text{any } f \in L^2(\mathsf{X}) \quad &\leftrightarrow \quad \text{representable in basis } [\psi_{1:\infty}] \text{ induced by } K() && (4)\\
&\qquad\quad \text{and approximation by } \psi_{1:m} \text{ converges uniformly} && (5)
\end{aligned}
$$

- How do you obtain such predictor f?
  - SVM / Kernel machines (frequentist + regularization)
  - Gaussian Processes (Bayesian)
  - Neural Net (NTK)
  - ...

Ingredients

- a **base space** X, which in ML is the input space.
  For example $\mathbb{R}^d$, or $\{x \in \mathbb{R}^d, \|x\| \leq R\}$.
- a **kernel** $K()$ over X, that defines a scalar product.

- $L^2(\mathsf{X})$, the space of functions that have finite 2-norm on X.

$$L^2(\mathsf{X}) = \{f : \mathsf{X} \to \mathbb{R}, \int_{\mathsf{X}} f^2(x)dx < \infty\} \tag{6}$$

- A kernel defines a scalar product on X iff it is positive definite in the following sense

$$\int_{\mathsf{X}} f(x)f(x')K(x,x')dxdx' > 0, \quad \text{for all } f \not\equiv 0, f \in L^2(\mathsf{X}). \tag{7}$$

  In particular, from (7) it follows that for any set $x^{1:n}$, the Gram matrix

$$G = \left[K(x^i, x^j)\right]_{i,j=1}^n \geq 0. \tag{8}$$

  Exercise Prove this

Ingredients

- a base space X, which in ML is the input space.
  For example $\mathbb{R}^d$, or $\{x \in \mathbb{R}^d, \|x\| \leq R\}$.
- a kernel $K()$ over X, that defines a scalar product.

- $L^2(\mathsf{X})$, the space of functions that have finite 2-norm on X.

$$L^2(\mathsf{X}) = \{f : \mathsf{X} \to \mathbb{R}, \int_{\mathsf{X}} f^2(x)dx < \infty\} \quad (6)$$

- A kernel defines a scalar product on X iff it is positive definite in the following sense

$$\int_{\mathsf{X}} f(x)f(x')K(x,x')dxdx' > 0, \quad \text{for all } f \neq 0, f \in L^2(\mathsf{X}). \quad (7)$$

In particular, from (7) it follows that for any set $x^{1:n}$, the Gram matrix

$$G = \left[K(x^i, x^j)\right]_{i,j=1}^n \geq 0. \quad (8)$$

Exercise Prove this

Remark 1

Scalar Product

- *A* **scalar product** $\langle\ \rangle$ *on* X *(also called* **inner product***).*

  $\langle\ \rangle : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ *is a scalar product iff it is*

  1. *Symmetric* $\langle x, x' \rangle = \langle x', x \rangle$.
  2. *Positive definite* $\langle x, x \rangle > 0$ *for all* $x \neq 0$.
  3. *Bilinear (i.e. linear in each argument)* $\langle \alpha x_1 + \beta x_2, x' \rangle = \alpha \langle x_1, x' \rangle + \beta \langle x_2, x' \rangle$ *(and similarly for second argument). Note that it suffices to be symmetric and linear in first argument.*

The recipe

Given X, kernel K over X

1. The **feature map** $x \mapsto K_x() = K(x,)$
   - Every $x \in \mathsf{X}$ maps to the function $K_x : \mathsf{X} \to \mathbb{R}$, defined as $K_x(u) = K(x, u)$ for all $u \in \mathsf{X}$.
   - Hence, each x is also a function in $L^2(\mathsf{X})$; we write this $\mathsf{X} \hookrightarrow L^2(\mathsf{X})$. But the set $\{K_x, x \in \mathsf{X}\}$ has a lot of "holes", it's not useful! Must be "filled in".
2. Start by expanding it into a linear space, the space of all finite sums of K_x's.

$$\mathcal{H}_0 = \text{span}\{K_x, x \in \mathsf{X}\} = \{\sum_{i=1}^{n} \alpha_i K_{x^i}, \text{ for } n = 1, 2, \ldots, \alpha_{1:n} \in \mathbb{R}, x^{1:n} \in \mathsf{X}\} \quad (9)$$

This is still not enough, we would like to include limits of sequences in $\mathcal{H}_0$, e.g. infinite sums. For limits we need a distance.

The recipe

Given $\mathcal{X}$, kernel K over $\mathcal{X}$

1. The **feature map** $x \mapsto K_x() = K(x,)$
   - Every $x \in \mathcal{X}$ maps to the function $K_x : \mathcal{X} \to \mathbb{R}$, defined as $K_x(u) = K(x, u)$ for all $u \in \mathcal{X}$.
   - Hence, each x is also a function in $L^2(\mathcal{X})$; we write this $\mathcal{X} \hookrightarrow L^2(\mathcal{X})$. But the set $\{K_x, x \in \mathcal{X}\}$ has a lot of "holes", it's not useful! Must be "filled in".
2. Start by expanding it into a linear space, the space of all finite sums of K_x's.

$$\mathcal{H}_0 = \text{span}\{K_x, x \in \mathcal{X}\} = \{\sum_{i=1}^{n} \alpha_i K_{x^i}, \text{ for } n = 1, 2, \ldots, \alpha_{1:n} \in \mathbb{R}, x^{1:n} \in \mathcal{X}\} \quad (9)$$

This is still not enough, we would like to include limits of sequences in $\mathcal{H}_0$, e.g. infinite sums. For limits we need a distance.

Remark 2

Complete metric space *In a complete space $\mathcal{H}$, if a sequence $\{f_n\}_{n=1}^{\infty}$ has a limit f, then f is also in $\mathcal{H}$; moreover (and this is the actual definition), if a sequence is Cauchy, meaning that distance$(f_n, f_m) \to 0$ for $m, n \to \infty$, then the limit f exists and is in $\mathcal{H}$.*

Remark 3

Hilbert space *A* **Hilbert space** *is an infinite dimensional vector space that has a scalar product and is complete.*

The recipe (2)

3. Define a scalar product $\langle\ \rangle_{\mathcal{H}}$ on $\mathcal{H}_0$, by means of the kernel K. Let

$$\langle K_x, K_{x'} \rangle_{\mathcal{H}} = K(x, x'). \tag{10}$$

Hence, the scalar product defined by K on X, is transported to $\mathcal{H}_0$. This is sufficient to define the scalar product on all of $\mathcal{H}_0$ because for any $f, g \in \mathcal{H}_0$,

$$\langle f, g \rangle_{\mathcal{H}} = \langle \sum_{i=1}^{n} \alpha_i K_{u^i}, \sum_{j=1}^{m} \beta_j K_{v^j} \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \langle K_{u^i}, K_{v^j} \rangle_{\mathcal{H}} \tag{11}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j K(u^i, v^j) \tag{12}$$

Exercise Prove that $\langle\ \rangle_{\mathcal{H}}$ is a scalar product.

4. The scalar product $\langle\ \rangle_{\mathcal{H}}$ allows us to define a norm

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}. \tag{13}$$

Now we can complete $\mathcal{H}_0$ to $\mathcal{H}$.

Voila! $\mathcal{H}$ is your **Reproducing Kernel Hilbert Space (RKHS)**.

Recipe, summarized

Input X, kernel K

1. Map $\mathsf{X} \hookrightarrow L^2(\mathsf{X})$ by the feature map $x \mapsto K_x() = K(x,)$
2. Make it a linear space
   $\mathcal{H}_0 = \text{span}\{K_x, x \in \mathsf{X}\} = \{\sum_{i=1}^n \alpha_i K_{x^i}, \text{ for } n = 1, 2, \ldots, \alpha_{1:n} \in \mathbb{R}, x^{1:n} \in \mathsf{X}\}$
3. Define scalar product $\langle\ \rangle_{\mathcal{H}}$ on $\mathcal{H}_0$, by $\langle K_x, K_{x'} \rangle_{\mathcal{H}} = K(x, x')$.
4. Complete $\mathcal{H}_0$ to $\mathcal{H}$ using $\|\ \|_{\mathcal{H}}$.

The name RKHS explained

- Reproducing Kernel **Hilbert Space**
  – means the space of functions has a scalar product and is complete
- Reproducing **Kernel** Hilbert Space
  – the scalar product comes from a kernel
- **Reproducing** Kernel Hilbert Space
  – in addition, this space has the Reproducing property (coming next!)

The Reproducing Property $\langle f, K_x \rangle_{\mathcal{H}} = f(x)$

- Let's prove it. Remember $f(x) = \sum_{i=1}^{n} a_i K_{u_i}(x)$ for $f \in \mathcal{H}_0$, $x \in \mathsf{X}$.

$$\langle f, K_x \rangle_{\mathcal{H}} = \sum_{i=1}^{n} a_i \langle K_{u_i}, K_x \rangle_{\mathcal{H}} \tag{16}$$

$$= \sum_{i=1}^{n} a_i K(u_i, x) = f(x) \tag{17}$$

- In other words, if we map x into $\mathcal{H}$ by $x \mapsto K_x$ and calculate the scalar product with some $f \in \mathcal{H}$, the result is the same as applying f to $x \in \mathsf{X}$.
- One can say that K_x reproduces x
- Or alternatively that $f \in \mathcal{H}$, by Riesz's Theorem, defines the linear functional $\langle f, \rangle_{\mathcal{H}}$. This functional on $\mathcal{H}$ reproduces the effect of f on X.

Mercer's Theorem

- Define the transport operator $T : L^2(\mathsf{X}) \to \mathcal{H}$

$$Tf = \int_{\mathsf{X}} f(u)K(\,,u)du \quad \Leftrightarrow \quad Tf(x) = \int_{\mathsf{X}} f(u)K(u,x)du \tag{18}$$

- Let $\{(\lambda_i, \psi_i)\}_i$ be the eigenvalue, eigenfunction pairs of T
- **The Mercer Theorem** says that, under certain conditions on X and K, the operator T
  1. has a discrete spectrum,
  2. is positive semidefinite $\lambda_i \geq 0$ for $i = 1, 2, \ldots$
  3. the eigenfunctions $\{\psi_i\}_{i=1}^{\infty}$ form an orthogonal basis for $L^2(\mathsf{X})$

Mercer's Theorem

- Define the transport operator $T : L^2(\mathcal{X}) \to \mathcal{H}$

$$Tf = \int_{\mathcal{X}} f(u)K(\cdot, u)du \quad \Leftrightarrow \quad Tf(x) = \int_{\mathcal{X}} f(u)K(u, x)du \tag{18}$$

- Let $\{(\lambda_i, \psi_i)\}_i$ be the eigenvalue, eigenfunction pairs of T
- The Mercer Theorem says that, under certain conditions on $\mathcal{X}$ and K, the operator T
  1. has a discrete spectrum,
  2. is positive semidefinite $\lambda_i \geq 0$ for $i = 1, 2, \ldots$
  3. the eigenfunctions $\{\psi_i\}_{i=1}^{\infty}$ form an orthogonal basis for $L^2(\mathcal{X})$

Remark 4

(Linear) Operator

- *An (linear) operator T is a (linear) function from a space of functions to another.*
- *For example the derivative maps a function $f : \mathbb{R} \to \mathbb{R}$ to its derivative f'; we can write that derivative :$\mathcal{C}^1(\mathbb{R}) \to \mathcal{C}^0(\mathbb{R})$ is a linear operator.*
- *For an operator T and function f, we denote by $g = Tf \equiv T(f)$, the function resulting from applying T to f.*
- *Furthermore, if we calculate this function g at point x, we write $g(x) = Tf(x)$*
- *Operators have eigenfunctions and eigenvalues defined as $T\psi = \lambda\psi$ for some $\lambda \in \mathbb{R}$*
- *The set of eigenvalues $\{\lambda$, such that $T\psi = \lambda\psi$ for some $\psi\}$ is the* **spectrum** *of T.*
- *The spectrum of an operator is usually more complicated than the spectrumof a matrix; for example, it can contain continuous intervals, the whole real line, limit points. If the spectrum contains none of these, i.e. consists of only isolated eigenvalues, we say the spectrum is* **discrete**.

The feature map revisited

- The consequences of this theorem are remarkable. In particular, it lets us express the kernel itself in the basis of T.

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x') \tag{19}$$

- Therefore,

$$K(x, x) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)^2. \tag{20}$$

- From here, it is easy to see that the feature map $x \mapsto K_x$ can also be written as

$$x \mapsto \left[\sqrt{\lambda_i}\psi_i(x)\right]_{i=1}^{\infty} \tag{21}$$

- And finally, the infinite sum converges uniformly

$$\lim_{n\to\infty} \sup_{x,x'} \left| K(x, x') - \sum_{i=1}^{n} \lambda_i \psi_i(x) \psi_i(x') \right| = 0 \tag{22}$$

The feature map revisited

- The consequences of this theorem are remarkable. In particular, it lets us express the kernel itself in the basis of T.

$$K(x,x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(x') \quad (19)$$

- Therefore,

$$K(x,x) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)^2. \quad (20)$$

- From here, it is easy to see that the feature map $x \mapsto K_x$ can also be written as

$$x \mapsto \left[\sqrt{\lambda_i}\psi_i(x)\right]_{i=1}^{\infty} \quad (21)$$

- And finally, the infinite sum converges uniformly

$$\lim_{n\to\infty} \sup_{x,x'} \left| K(x,x') - \sum_{i=1}^{n} \lambda_i \psi_i(x)\psi_i(x') \right| = 0 \quad (22)$$

- It's important to remember that there are 2 scalar products here. There is the scalar product induced by the kernel K on $\mathcal{H}$, defined in (10) and (11), and there is the standard scalar product on $L^2(\mathcal{X})$ defined by $\langle f, g\rangle = \int_{\mathcal{X}} f(x)g(x)dx$.
- The basis $\{\psi_i\}$ is orthonormal w.r.t. the $L^2(\mathcal{X})$ scalar product.

How to prove (19).

- ψ_j is eigenfunction, hence

$$\int_{\mathcal{X}} \psi_j(x')K(x,x')dx' = \lambda_j \psi_j(x). \quad (23)$$

- Now K_x itself has a decomposition in the basis, $K_x = \sum_i \gamma_i(x)\psi_i$, where $\gamma_i(x)$ are the coefficients.
- Let's plug this decomposition in (23)

$$\int_{\mathcal{X}} \psi_j(x')K(x,x')dx' = \int_{\mathcal{X}} \psi_j(x') \sum_i \gamma_i(x)\psi_i(x')dx' \quad (24)$$

$$= \sum_i \gamma_i(x) \int_{\mathcal{X}} \psi_j(x')\psi_i(x')dx' \quad (25)$$

$$= \gamma_j(x) = \lambda_j \psi_j(x). \quad (26)$$

- Hence $K_x(x') \equiv K(x,x') = \sum_i \lambda_i \psi_i(x)\psi(x')$. Done.