Chapter 4, Part 2: Reproducing Kernel Hilbert Spaces (RKHSs)

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2024

rainforth@stats.ox.ac.uk

A More Formal Look at Kernels

- Kernels are inner products between feature maps that form a similarity measure between inputs
- We would like to be able to construct valid kernels directly as similarity measures in way that ensures they imply a valid feature map
- What are the conditions on a similarity measure $k\colon \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ to ensure that it is a valid such kernel?
- To answers this, we need to get a lot more technical and introduce the concept of a **reproducing kernel Hilbert spaces (RKHS)**

Inner Products

Our first step is to give a more formal definition of an inner product that we will be able to extend to infinite spaces:

Definition 1 (Inner Product)

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be an **inner product** on $\mathcal{H}$ if

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$ (**linear**),
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$ (**symmetric**),
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$,
4. $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

We can further define a **norm** using this inner product as

$$\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$$

Hilbert Spaces

Definition 2 (Hilbert Space)

A **Hilbert space** is a vector space on which we define an inner product, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and which is complete with respect to this inner product.

- Hilbert space is a generalization of Euclidean space that allows for infinite dimensionality
- The inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is part of the definition of a Hilbert space: we can dictate this definition
- Informally, "complete" means that there are no gaps, e.g. the set of rational numbers is not complete because the irrational numbers cause "gaps" in the space

Kernels

Definition 3 (Kernel)

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** if there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} .$$

- $\mathcal{H}$ is known as **a feature space** and φ **a feature map** of k
  - Neither of these is unique: $\varphi_1(x) = x$ and $\varphi_2(x) = [\sqrt{2}x, -x]^T$ both produce the kernel $k(x, x') = x^\top x'$ when using a standard inner product, but are distinct feature maps with distant Hilbert spaces
- There are **no conditions** on $\mathcal{X}$; it does not even need to be numeric (e.g. we can define features, & thus kernels, for text)

Positive Definite Functions

Definition 4 (Positive Definite Function)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is **positive definite** iff,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0 \quad \forall n \geq 1,\ \forall a_i \in \mathbb{R},\ \forall x_i \in \mathcal{X}$$

This is analogous to the notation of a positive **semi**–definite matrix: if we define $K_{i,j} = k(x_i, x_j)$ then we have $a^T K a \geq 0 \; \forall a$

The function $k(\cdot, \cdot)$ is **strictly positive definite** if for mutually distinct x_i, the equality holds only when all the a_i are zero.

Kernels are Positive Definite

Lemma 5

All kernels are positive definite functions.

Proof.

$$
\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left\langle \varphi(x_i), \varphi(x_j) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle a_i \varphi(x_i), a_j \varphi(x_j) \right\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} a_i \varphi(x_i), \sum_{j=1}^{n} a_j \varphi(x_j) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{n} a_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.
\end{aligned}
$$

Function Spaces

- We can think of **infinite dimensional** Hilbert spaces as spaces of **functions**
- We can thus think about functions as living in a Hilbert space, $f\colon \mathcal{X} \mapsto \mathbb{R},\ \forall f \in \mathcal{H}$
- By definition, $\varphi(x) \in \mathcal{H}$ and so here we can think of φ as mapping from x to functions, such that $\varphi\colon \mathcal{X} \mapsto (\mathcal{X} \mapsto \mathbb{R})$ and we can think of evaluating $(\varphi(x))(x')$
- Most infinite dimensional features spaces do not contain all possible functions; they impose constraints on **smoothness**
- The inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ now requires **explicit definition** ($\varphi(x)^\top \varphi(x')$ is no longer well–defined): we will construct it in a deliberate manner to give us the properties we want

Reproducing Kernels and Reproducing Kernel Hilbert Spaces

A particularly important class of Hilbert spaces are **reproducing kernel Hilbert spaces** (RKHS), which have the following, somewhat unusual, definition

Definition 6 (Reproducing Kernel and RKHS)

Let $\mathcal{H}$ be a **Hilbert space of functions** $f\colon \mathcal{X} \to \mathbb{R}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if it satisfies

- $\forall x \in \mathcal{X},\ \ k_x := k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H},\ \ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (**the reproducing property**).

If $\mathcal{H}$ has a reproducing kernel, it is called a **reproducing kernel Hilbert space** (RKHS).

Reproducing Kernels are Kernels

We did not explicitly introduce a feature map, but it is easy to show that **any reproducing kernel is also valid kernel** with the feature map $\varphi\colon x \mapsto k(\cdot, x)$, known as the **canonical feature map**.

This result follows by applying the reproducing property backwards with $f = k\left(\cdot, x'\right)$:

$$\begin{aligned} k(x, x') = f(x) = \langle f, k\left(\cdot, x\right)\rangle_{\mathcal{H}} &= \left\langle k\left(\cdot, x'\right), k\left(\cdot, x\right)\right\rangle_{\mathcal{H}} \\ &= \langle \varphi(x), \varphi(x')\rangle_{\mathcal{H}}. \end{aligned}$$

Note that the features $\varphi(x)$ are not specified explicitly in a vector form, but rather as **functions** on $\mathcal{X}$.

All Positive Definite Functions are Reproducing Kernels

So far we have relied on careful definitions without really saying much about when these definitions might be satisfied. The following profound result is an abrupt divergence from this:

Theorem 7 (Moore-Aronszajn)

Every positive definite function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a also a reproducing kernel with a unique corresponding RKHS.

In other words, not only does k being positive definite imply that it is also a kernel that can be written as an inner product, it can specifically be written as an inner product in a RKHS such that the reproducing property, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$, holds.

Proof: whiteboard/notes

Taking Stock

We have now shown the following:

- Any positive definite k is a reproducing kernel with feature map $\varphi(x) = k(\cdot, x)$
- All reproducing kernels are kernels in the sense of being expressible as inner products between features
- All kernels are positive definite

We have thus come full circle and see that the following concepts are all **equivalent**: (1) reproducing kernels, (2) kernels as inner products, and (3) positive definite functions

Further Reading

- An alternative (relatively gentle) lecture from Ulrike von Luxburg: `https://youtu.be/EoM_DF3VAO8`
- An alternative (less gentle but more in depth) lecture from Arthur Gretton: `https://youtu.be/alrKls6BORc` (note this goes into some things we are yet to cover and some things we will not cover at all)
- All the gory details: `http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_2014.pdf`