Chapter 4, Part 3: RKHSs as Hypothesis Classes

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2024

rainforth@stats.ox.ac.uk

Recap

Last time we showed that:

- Kernels as inner products ($k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$),
- Positive definite functions ($\sum_i \sum_j a_i a_j k(x_i, x_j) \geq 0$),
- Reproducing kernels ($k(\cdot, x) \in \mathcal{H},\ \ \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$),

are all equivalent concepts as we can prove all three properties by assuming any one of them

We also explained how reproducing kernels have the canonical feature map $\varphi(x) = k(\cdot, x)$ that maps datapoints to functions in a reproducing kernel Hilbert space (RKHS)

Overview

This time:

- Deriving a RKHS from a kernel
- How can we interpret RKHSs? Intuitively, when is a function space an RKHS?
- Can we use an RKHS as a hypothesis class for empirical risk minimization?

An Alternative View of RKHSs

The requirements of a function space to be an RKHS turn out to be very weak: informally a function space $\mathcal{H}$ is an RKHS if $f(x)$ is finite whenever the function norm $\|f\|_{\mathcal{H}}$ is finite. More formally,

Definition 1 (RKHS, alternative defintion)

$\mathcal{H}$ is an RKHS if the evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x f = f(x)$ are continuous $\forall x \in \mathcal{X}$.
Equivalently, $\mathcal{H}$ is an RKHS if δ_x is a bounded operator,[1] that is $\exists C_x : 0 < C_x < \infty$ and

$$|\delta_x f| = |f(x)| \leq C_x \|f\|_{\mathcal{H}} \qquad \forall x \in \mathcal{X}, \forall f \in \mathcal{H}.$$

Important implication: if $\|f - g\|_{\mathcal{H}} = 0$, then $f(x) = g(x)$ $\forall x \in \mathcal{X}$

[1] Note that while we require that $C_x < \infty$ $\forall x \in \mathcal{X}$, it can be the case that $\sup_x C_x = \infty$. An example of this is the linear kernel, where C_x is finite for any given x, but $\lim_{\|x\| \to \infty} C_x = \infty$.

An Alternative View of RKHSs

- Proving this alternative definition holds from our previous definition is relatively straightforward using the reproducing property and the Cauchy–Schwarz inequality:

$$
\begin{aligned}
|\delta_x f| = |f(x)| = |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}| &\le \|f\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}} \\
&= \|f\|_{\mathcal{H}} \sqrt{k(x, x)},
\end{aligned}
$$

  so we have $|\delta_x f| \le C_x \|f\|_{\mathcal{H}}$ where $C_x = \sqrt{k(x, x)}$

- Proving our previous definition holds from this definition is also possible but somewhat messier: it uses something called the Riesz representation theorem and is beyond the scope of the course.

The Reproducing Kernel of a RKHS is Unique

Theorem 2 (Uniqueness of reproducing kernel)

Each RKHS has a unique corresponding reproducing kernel.

Proof.

Assume, for the sake of contradiction, that an RKHS $\mathcal{H}$ has two unique reproducing kernels k_1 and k_2. Using a combination of linearity and the reproducing property we have $\forall f \in \mathcal{H}, x \in \mathcal{X}$:

$$\begin{aligned} \langle f, k_1(\cdot, x) - k_2(\cdot, x) \rangle_{\mathcal{H}} &= \langle f, k_1(\cdot, x) \rangle_{\mathcal{H}} - \langle f, k_2(\cdot, x) \rangle_{\mathcal{H}} \\ &= f(x) - f(x) = 0. \end{aligned}$$

In particular, this holds for $f = k_1(\cdot, x) - k_2(\cdot, x)$, which yields $\|k_1(\cdot, x) - k_2(\cdot, x)\|_{\mathcal{H}}^2 = 0, \ \forall x \in \mathcal{X}$, which implies $k_1 = k_2$ and we have our desired contradiction. $\square$

Uniqueness of RKHS

Though we will not prove it, the inverse result also turns out to be true: the RKHS for any kernel (and thus positive definite function) is unique; we can denote the RKHS for kernel k as $\mathcal{H}_k$

Putting everything together, we have the following key ideas:

- An RKHS corresponds to a space of functions: choosing a RKHS corresponds to choosing a hypothesis class of functions
- RKHSs can be very general: most "well–behaved" function spaces are RKHSs
- There is a one–to–one correspondence between a kernel k and its RKHS $\mathcal{H}_k$

We thus see that we can directly imply powerful hypothesis classes of functions, $f \in \mathcal{H}$ through appropriate choices of kernels k

RKHSs as Hypothesis Classes

Can we use an RKHS as a hypothesis class for (regularized) empirical risk minimization (ERM)?

A typical and general setup would be that we are looking for the function f^* in the RKHS $\mathcal{H}_k$ which solves

$$f^* = \underset{f \in \mathcal{H}_k}{\arg\min}\, \hat{R}(f) + \Omega\left(\|f\|^2_{\mathcal{H}_k}\right),$$

for empirical risk $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i), x_i)$, a loss function $L\colon \mathcal{Y} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$ and any non-decreasing function Ω.

Representer Theorem

Theorem 3 (Representer Theorem)

There is always a solution to

$$f^* = \arg\min_{f \in \mathcal{H}_k} \hat{R}(f) + \Omega\left(\|f\|^2_{\mathcal{H}_k}\right) \tag{1}$$

that takes the form

$$f^* = \sum_{i=1}^{n} a_i k(\cdot, x_i), \qquad a_i \in \mathcal{R} \tag{2}$$

where x_i are our input datapoints. If Ω is strictly increasing, all solutions have this form.

Proof.

Whiteboard/notes $\square$

Representer Theorem Implications

The critical part of this result is that f^* is a linear combination of the feature mappings of our training data

- We can work with complex RKHS hypothesis classes while knowing that our solution will still take a simple form
- There is a very clear direct dependency of the functions we learn from the kernel we choose
- For a fixed kernel, the complexity of f^* is restricted for a given n, helping to prevent overfitting: we learn more complex functions as and when we see more data
- A downside is that we need to retain all our data to make predictions and this prediction will cost at least $O(n)$: can make kernel methods unsuitable for large datasets

Example: Kernel SVMs

We can express the primal problem for a kernel-SVM (fixing $b = 0$ for simplicity—we can incorporate the offset into the kernel)

$$\min_{w \in \mathcal{H}_k} \left(\frac{1}{2} \|w\|_{\mathcal{H}}^2 + C \sum_{i=1}^{n} (1 - y_i \langle w, k(x_i, \cdot) \rangle_{\mathcal{H}})_+ \right)$$

which is the form required by the representer theorem.

We know from before that this leads to the decision function $\hat{y}(x) = \mathrm{sign}(f(x))$, where, because $b = 0$,

$$f(x) = \langle w, k(x, \cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i y_i k(x, x_i)$$

and we see that this is of the required form with factors $\alpha_i y_i$

Kernel Methods are Powerful

Our results so far demonstrate a number of advantages of kernel methods:

- RKHS spaces are a general and powerful class of function spaces: virtually all “well–behaved” function spaces can be expressed as an RKHS
- We can use kernels to perform ERM with an RKHS as our hypothesis class, allowing for very wide ranges of predictors to be learned in a **nonparametric** manner
- Many kernel methods permit simple, or even even analytic, solutions to the ERM because of their basis in linear models

Kernel Methods have Drawbacks

But also some major drawbacks:

- Choosing the right kernel can be extremely important: our predictor will depend directly on this choice
- Choosing the right kernel (and thus RKHS) can be difficult: some RKHSs are actually very restrictive, while choosing an overly broad RKHS will lead to poor generalization (due to overfitting)
- They tend to have relatively poor computational scaling in the size of the data (compared with, e.g., deep learning, random forests): they are based on pairwise similarities and thus have at best $O(n^2)$ scaling at training time and $O(n)$ at test time and often much worse than this

Next time: constructing kernels