CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Reproducing Kernel Hilbert Spaces

Lecturer: Michael I. Jordan *Scribes: Sierra Boyd*

1 Hilbert Space

A Hilbert space is essentially an infinite-dimensional Euclidean space. It is a vector space (i.e., is closed under addition and scalar multiplication, obeys the distributive and associative laws, etc.). It is also endowed with an inner product $\langle \cdot, \cdot \rangle$; a bilinear form obeying the following conditions:

$$\begin{aligned} \langle x+y, z \rangle &= \langle x, z \rangle + \langle y, z \rangle \\ \langle \alpha x, y \rangle &= \alpha \langle x, y \rangle \\ \langle x, y \rangle &= \langle y, x \rangle \\ \langle x, x \rangle &\geq 0 \\ \langle x, x \rangle = 0 &\rightarrow x = 0 \end{aligned}$$

From $\langle \cdot, \cdot \rangle$ we get a norm $\| \cdot \|$ via $\| x \| = \langle x, x \rangle^{1/2}$. This norm allows us to define notions of convergence. Adding all limit points of Cauchy sequences to our space yields a Hilbert space—a *complete* inner product space.

1.1 Examples

- R^n: $\langle x, y \rangle = x^T y$
- L_2: $\langle x, y \rangle = \int x(t) y(t) dt$
- l_2: $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$

The Cauchy-Schwartz lemma:

$$\langle x, y \rangle \leq \| x \| \| y \|$$

is easily proved for any Hilbert space.

2 Reproducing kernel Hilbert spaces

The Hilbert space L_2 is too "big" for our purposes, containing too many non-smooth functions. One approach to obtaining restricted, smooth spaces is the Reproducing Kernel Hilbert Space (RKHS) approach. A RKHS is "smaller" than a general Hilbert space.

Given a kernel $k(x, x')$, we will construct a Hilbert space such that k is a dot product in that space. First define the *Gram matrix*. Given points $x_1, x_2, x_3, ...x_n$, define:

$$K_{ij} = k(x_i, x_j)$$

We say that the kernel k is *positive definite* if its Gram matrix is positive definite for all $x_1, x_2, ..., x_n$.

The Cauchy-Schwartz inequality holds for kernels:

$$k(x_1, x_2)^2 \leq k(x_1, x_1)k(x_2, x_2)$$

Proof: Form a Gram matrix of the two points x_1 and x_2:

$$K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{pmatrix}$$

For K to be positive definite as a matrix, the determinant of K must be nonnegative:

$$\Longrightarrow k(x_1, x_1)k(x_2, x_2) - k(x_2, x_1)k(x_1, x_2) \geq 0,$$

which implies Cauchy-Schwartz.

Define the following *reproducing kernel map*:

$$\Phi : x \longrightarrow k(\cdot, x).$$

I.e., to each point x in the original space we associate a function $k(\cdot, x)$.

Example: Gaussian kernel. Each point x maps to a Gaussian centered at that point. Intuitively this captures the similarity of x to all other points.

We now construct a vector space containing all linear combinations of the functions $k(\cdot, x)$:

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

This will be our RKHS.

We now define an inner product. Let $g(\cdot) = \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j)$, and define:

$$\langle f, g \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

We need to verify that this in fact defines an inner product. Symmetry is obvious: $\langle f, g \rangle = \langle g, f \rangle$. Linearity is easy to show. We focus on the key property: $\langle f, f \rangle = 0 \Longrightarrow f = 0$.

Note first that for any $f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$, we have:

$$\langle k(\cdot, x), f \rangle = \sum_{i=1}^{m} \alpha_i k(x_i, x) = f(x),$$

which shows that the kernel is the *representer of evaluation.*

Kernels are analogs of Dirac delta functions. Consider L_2 (which is not a RKHS). We have:

$$f(x) = \int f(t)\delta(t, x)dt,$$

where $\delta(t, x)$ is the Dirac delta function. The Dirac delta function is the representer of evaluation for L_2, but of course it is not itself in L_2. (Which is consistent with the fact that L_2 is not a RKHS).

Suppose that we plug the kernel in for f in Eq. 2:

$$\langle k(\cdot,x), k(\cdot,x')\rangle = k(x,x').$$

This is the *reproducing property* of the kernel.

From Cauchy-Schwartz we can prove the following:

$$f(x)^2 = \langle k(\cdot,x), f\rangle^2 \leq k(x,x)\langle f,f\rangle$$

and (finally) this implies that if $\langle f,f\rangle = 0$, then $f \equiv 0$.

We complete the space that we have constructed to obtain a Hilbert space. This is our RKHS.

3 Mercer's theorem and RKHS

Recall the following condition for Mercer's theorem:

$$\int k(x,x')f(x)f(x')dxdx' \geq 0.$$

Given this condition, we can expand the function $k(x,x')$ in its eigenfunctions:

$$k(x,x') = \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\psi_j(x'). \tag{1}$$

where

$$\int k(x,x')\psi(x')dx' = \lambda_j \psi_j(x);$$

i.e., $\psi_j(x')$ is an eigenfunction.

We now construct a RKHS via Mercer as a linear combination of these eigenfunctions. This is a different approach to constructing a RKHS that our earlier kernel map. We have:

$$\mathcal{H} = \left\{ \sum_{j=1}^{\infty} c_n \psi_n(x) \right\}.$$

In particular, the kernel is in this space since it is a linear combination of the eigenfunctions (cf. Eq. 1).

Define a dot product $\langle\cdot,\cdot\rangle$:

$$\left\langle \sum_n c_n \psi_n(x), \sum_n d_n \psi_n(x) \right\rangle = \sum_n \frac{c_n d_n}{\lambda_n}.$$

Note that dividing by the eigenvalue, λ_n, makes H different from l_2. Dividing by these eigenvalues in effect amounts to imposing a smoothness condition on the space; for a function to be in $\mathcal{H}$ the coefficients c_n must go to zero quickly (so that the norm $\sum_n c_n^2 \lambda_n$ is finite).

Now verify we have a RKHS:

$$\begin{aligned} \langle f(\cdot), k(\cdot,x')\rangle &= \sum_n \frac{c_n \lambda_n \psi_n(x')}{\lambda_n} \\ &= \sum_n c_n \psi_n(x') \\ &= f(x'). \end{aligned}$$

In summary, Mercer's theorem provides a concrete way to construct a RKHS. In essence, Mercer's theorem provides a coordinate basis representation of an RKHS.