Reproducing Kernel Hilbert Spaces

Uri Shaham

January 20, 2025

1 Kernels

Motivation - the XOR problem: cannot be linearly separated in 2 dimensions, but can be in higher dimensionality. Kernels can efficiently compute dot product in infinite dimensional space, without actually transition the data to that space.

Definition 1.1 (Hilbert space)**.** *A Hilbert space is a complete space with inner product.*

Definition 1.2 (Kernel)**.** *Let $\mathcal{X}$ be a non-empty set. 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 $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$, $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.*

For example, $\mathcal{X} = \mathbb{R}$, $\phi(x) = x$.

Definition 1.3 (Positive semi-definite functions)**.** *A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive semi-definite (PSD) if for all $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$,*

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

Lemma 1.4. *Let $\mathcal{X}$ be a non-empty set, $\mathcal{H}$ be a Hilbert space and let k be a kernel function. Then k is PSD.*

Proof. Choose some $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$. Then

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

□

The converse holds as well:

Lemma 1.5. *A symmetric positive definite function is an inner product in some Hilbert space (and thus a kernel)*

Proof. We first need to define $\mathcal{H}$, its inner product, and ϕ, and then show that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. We define $\mathcal{H}$ as the space of linear combinations of functions $k(\cdot, x_i)$, i.e.,

$$\mathcal{H} = \left\{ \sum_{i=1}^{m} a_i k(\cdot, x_i),\ a_i \in \mathbb{R}, x_i \in \mathcal{X}, m \in \mathbb{N} \right\}.$$

We then define the inner product as

$$\left\langle \sum_{i=1}^{m} a_i k(\cdot, x_i), \sum_{j=1}^{n} a_j k(\cdot, x_j) \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i a_j k(x_i, x_j).$$

Note that since k is PSD, this inner product is valid. Finally, we see that by defining $\phi(x) = k(\cdot, x)$ we have $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. □

Lemma 1.6. *Sum of kernels is a kernel.*

Proof. By using Lemmas 1.4 and 1.5 we get

$$\sum_{i,j=1}^{n} a_i a_j \left(k_1(x_i, x_j) + k_2(x_i, x_j) \right) = \sum_{i,j=1}^{n} a_i a_j k_1(x_i, x_j) + \sum_{i,j=1}^{n} a_i a_j k_2(x_i, x_j) \geq 0.$$

□

Definition 1.7 (RBF kernel)**.** *The Radial Basis Function kernel (aka Gaussian kernel) is defined as*

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2} \right).$$

Lemma 1.8. *The RBF kernel is a valid kernel, whose corresponding feature map is infinite-dimensional.*

Proof. See homework.

□

2 Reproducing Kernel Hilbert Spaces

Definition 2.1 (RKHS)**.** *Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is called a reproducing kernel Hilbert space if k satisfies:*

1. *For every $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$*
2. *The reproducing property: for every $x \in \mathcal{X}$ and $f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.*

In particular, $\langle k(\cdot, y), k(\cdot, x) \rangle_{\mathcal{H}} = k(x, y)$, hence a reproducing kernel is a valid kernel. $\phi(x) = k(\cdot, x)$ is often called the canonical feature map. The following theorem says the converse.

Theorem 2.2 (Moore-Aronszajn)**.** *Every symmetric, PSD kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines a RKHS $\mathcal{H}$, for which k is the reproducing kernel.*

Proof. Define $\mathcal{H}_0 = \text{span}\{\phi(x) : x \in \mathcal{X}\}$, with the inner product

$$\left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{m} a_j \phi(x_j) \right\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i a_j k(x_i, x_j),$$

hence $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}0} = k(x, y)$. To make $\mathcal{H}_0$ a Hilbert space, we need to consider its completion $\mathcal{H}$, which is composed of elements of the form $f = \sum_{i=1}^{\infty} a_i \phi(x_i)$, where the sum converges. We can now verify the reproducing property holds:

$$\langle f, \phi(x) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{\infty} a_i \langle \phi(x_i), \phi(x) \rangle = \sum_{i=1}^{\infty} a_i k(x_i, x) = \sum_{i=1}^{\infty} a_i \phi(x_i)(x) = f(x).$$

It remains to show that $\mathcal{H}$ is unique. Let $\mathcal{G}$ be an RKHS for which k is a reproducing kernel. Then for every $x, y \in \mathcal{X}$, $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{G}}$. Hence, by linearity, the inner products in $\mathcal{H}$ and $\mathcal{G}$ equal on $\text{span}\{\phi(x) : x \in \mathcal{X}\}$. Then $\mathcal{H} \subseteq \mathcal{G}$, since G is complete and contains $\mathcal{H}_0$. We will show that $\mathcal{G} \subseteq \mathcal{H}$. Let $f \in \mathcal{G}$ and write $f = f_{\mathcal{H}} + f_{\mathcal{H}^\perp}$, where $f_{\mathcal{H}} \in \mathcal{H}$ and $f_{\mathcal{H}^\perp} \in \mathcal{H}^\perp$. Then

$$f(x) = \langle \phi(x), f \rangle_{\mathcal{G}} = \langle \phi(x), f_{\mathcal{H}} \rangle + \langle \phi(x), f_{\mathcal{H}^\perp} \rangle = \langle \phi(x), f \rangle_{\mathcal{H}} = f_{\mathcal{H}}(x),$$

since $\phi(x) \in \mathcal{H}$, so $\langle \phi(x), f_{\mathcal{H}^\perp} \rangle = 0$. Then $f \in \mathcal{H}$ and hence $\mathcal{H} = \mathcal{G}$, which concludes the proof. □

The representer theorem shows that the minimizer of the empirical risk (i.e., train loss) over an RKHS can be obtained as a linear combination of feature maps of training points. This is a significant result, as it simplifies the search for optimal solutions to a linear program.

Theorem 2.3 (Representer thm)**.** *Let k be a kernel function and $\mathcal{H}$ be the corresponding RKHS. We are provided with training data $(x_1, y_1), \ldots (x_n, y_n)$, an error function $E : \mathbb{R}^2 \to \mathbb{R}$ and a strictly increasing regularizer function $g : [0, \infty) \to \mathbb{R}$. Let f^* be a minimizer of the regularized empirical risk, i.e.,*

$$f^* = \arg\min_f \left(E(f(x_1), y_1), \ldots, E(f(x_n), y_n)\right) + g(\|f\|).$$

Then $f^* = \sum_{i=1}^{n} a_i \phi(x_i)$, for some a_i's.

Proof. We decompose every function $f \in \mathcal{H}$ to a component in $\text{span}\{\phi(x_1), \ldots, \phi(x_n)\}$ and an orthogonal component: $f = \sum_{i=1}^{n} a_i \phi(x_i) + v$, where $\langle \phi(x_i), v \rangle = 0$ for all $i = 1, \ldots, n$. Then by the reproducing property,

$$f(x_j) = \left\langle \sum_{i=1}^{n} a_i \phi(x_i) + v, \phi(x_j) \right\rangle = \sum_{i=1}^{n} a_i k(x_i, x_j).$$

Hence the values of f on the training data do not depend on v, and consequently the errors $E(f(x_i), y_i)$. Finally, considering the regularization term,

$$\begin{aligned} g(\|f\|) &= g\left(\left\|\sum_{i=1}^{n} a_i \phi(x_i) + v\right\|\right) \\ &= g\left(\sqrt{\left\|\sum_{i=1}^{n} a_i \phi(x_i)\right\|^2 + \|v\|^2}\right) \\ &\geq g\left(\left\|\sum_{i=1}^{n} a_i \phi(x_i)\right\|\right), \end{aligned}$$

where we have used orthogonality and the fact that g is increasing. Therefore $v = 0$ does not affect the training error and strictly reduces the regularization penalty. Therefore $v = 0$, so $f^* = \sum_{i=1}^n a_i \phi(x_i)$. $\square$

3 Application: kernel ridge regression

Given train data (x_i, y_i) $i = 1, \dots n$, we assume a model $y = f(x) + \epsilon$, and seek for f^* such that $y_i = f^*(x_i) + \epsilon_i$ for all i. Let $\mathcal{H}$ be a RKHS with kernel k. Since f can be arbitrarily expressive, we need to regularize it. The optimization is therefore

$$\arg\min_{f \in \mathbb{H}} \sum_{i=1}^n ((y_i) - f(x_i))^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2.$$

By the representer theorem, we know that $f = \sum_{j=1}^n a_j \phi(x_j)$, for some $a = (a_1, \dots, a_n)^T$, where $\phi(x_i) = k(\cdot, x_i)$. In vector notation, we define $y = (y_1, \dots, y_n)^T$, and the kernel matrix K, such that $k_{ij} = k(x_i, x_j)$. Then the miminization problem becomes

$$\arg\min_a \|y - Ka\|^2 + + \frac{\lambda}{2} a^T K a.$$

Taking gradient wrt a, using the fact that K is symmetric, and equating to zero, we get

$$K^2 a - Ky + \lambda K a = 0.$$

Rearranging, we get

$$K(K + \lambda I)a = Ky.$$

Assuming k is PD, and multiplying from the left by K^{-1}, we get

$$\hat{a} = (K + \lambda I)^{-1} y.$$

For prediction at a new test point x we then have

$$\hat{y}(x) = a^T \phi(x_i)(x) = \sum_{i=1}^n a_i k(x_i, x) = y^T (K + \lambda I)^{-1} k(x)$$

where $k(x) = (k(x, x_1), \dots, k(x, x_n))^T$.