Kernels are all you Need

Thomas Walker

December 2023

Abstract

Machine learning is fundamentally an optimization problem. Albeit a challenging optimisation problem as it is often situated in high dimensional spaces along highly convex surfaces. Consequently, efforts to theoretically ground this process have been explored. One of the most natural structures to embed this optimization problem is that of Hilbert spaces. For a detailed introduction to the theory of Hilbert spaces refer to [2]. Hilbert spaces are regarded as having a simple and highly regular nature, along with a capacity to learn on group structured input or output space [1]. Much of the following exposition will be based on the content of [1], which provides a comprehensive theoretical grounding of machine learning techniques. The contextualisation of machine learning techniques into Hilbert spaces has been a prominent mechanism by which researchers have come to understand techniques from a theoretical perspective. The idea is to recast optimisation problems within a Hilbert space using kernels, and then study the kernels.

Contents

Part I
The Theory of Kernels

1 Introducing Kernels

1.1 Kernels on a Hilbert Space

Simply put, a Hilbert space is a vector space with an inner product that has desirable convergence properties. Hilbert spaces can be vector spaces over arbitrary fields, however, to study machine learning it suffices to only consider the case when the field is $\mathbb{R}$. A prototypical example of a Hilbert space is $\mathbb{R}^n$, where the inner product is the usual dot product between vectors. That is,

$$\langle x, y \rangle = \sum_{k=1}^{n} x_k y_k$$

for $x, y \in \mathbb{R}^n$. Any finite dimension real Hilbert space is equivalent to $\mathbb{R}^n$ for some n. This is useful for in this setting we can apply the regular tools of linear algebra to study these spaces. However, the utility of the Hilbert space formalisation is its capacity to investigate infinite-dimensional Hilbert space whose theory is more exotic than the finite-dimensional case. Henceforth, we will let H be a real Hilbert space, and consider $X \subset H$.

Definition 1.1.1. *For a map $\varphi : X \to H$, the corresponding kernel is $k : X \times X \to \mathbb{R}$ given by*

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle ,$$

where $\langle \cdot, \cdot \rangle$ is the inner product defined on H.

In $\mathbb{R}^2$ it is well-understood that the inner product measures some sort of similarity between vectors. In particular, the similarity is related to the angular separation of the vectors. Generalising this to an arbitrary Hilbert space, one senses that a kernel is measuring the similarity between the output of a function over points in a set. Recall that an inner product is bilinear, symmetric, and positive definite. Consequently, a kernel has the following properties.

- A kernel k is symmetric.
- A kernel is positive semi-definite. That is, for any m, $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, and $x_1, \ldots, x_m \in X$ we have

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0. \tag{1.1}$$

  - If (1.1) holds with a strict inequality, we say that k is a strictly positive kernel. On the other hand, if (1.1) holds under the condition that $\sum_{i=1}^{m} \alpha_i = 0$, then we say that k is conditionally positive definite.

Other functions $k' : X \times X \to \mathbb{R}$ may also satisfy the symmetry and positive-definiteness properties detailed above, without being explicitly defined as in Definition 1.1.1. It turns out that symmetry and definiteness are necessary and sufficient conditions for a function to be a kernel of some Hilbert space.

1.2 Building Spaces for Kernels

Let $k : X \times X \to \mathbb{R}$ be a symmetric and positive definite function. That there is no reference to a Hilbert space. We can now construct a Hilbert space on which k defines a kernel.

1. Let V be the vector space defined with the basis $\{k_x\}_{x \in X}$, where $k_x := k(x, \cdot)$.
2. For the basis vectors define the inner product as $\langle k_x, k_{x'} \rangle = k(x, x')$. Extend this definition to V using linearity.
3. Form the Hilbert space H as the completion of V.

A space constructed in such a way is known as the reproducing kernel Hilbert space for k. Conversely, using the Riesz representation theorem, it follows that any Hilbert space in which the maps φ_x given by $f \mapsto f(x)$ are continuous is a reproducing kernel Hilbert space.

2 Regularising Using Kernels

2.1 Statistical Learning Framework

Suppose that a set of data $X \times Y$ has some probability distribution $\mathcal{D}$ defined over it. Typically we refer to X as the input space and Y as the output space. If this distribution is unknown then we can use the framework of statistical learning to construct a function $f : X \to Y$ to approximate the distribution. The construction requires a set of samples from the distribution and a method of quantifying the quality of the approximation. Using a loss $L : Y \times Y \to \mathbb{R}$ we can compute $\mathbb{E}_{(x,y)\sim\mathcal{D}}(L(f(x), y))$ to understand the discrepancy between our approximation f, and the true distribution $\mathcal{D}$. Empirically, one tries to minimize this quantity, using a set of samples $\{(x_i, y_i)\}_{i=1}^m$, by considering the optimisation problem,

$$f_{\text{emp}} = \operatorname{argmin}_{f\in\mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i) \right). \tag{2.1}$$

Here $\mathcal{F}$ denotes a hypothesis space of approximations. The optimisation problem of (2.1) is known as the empirical risk minimisation problem. Ideally, our sample will provide a useful proxy of the distribution $\mathcal{D}$ such that optimising (2.1) will yield a representative approximation for $\mathcal{D}$. However, this is typically not the case and some capacity control is required to achieve effective learning. Specifically, one instead optimises the regularised risk

$$R_{\text{reg}} = \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i) + \lambda\Omega(f), \tag{2.2}$$

where Ω is a regularising term which penalises f based on a selected measure that correlates with overfitting. The parameter $\lambda > 0$ dictates the strength of the regularisation. Deciding a suitable form for Ω is critical to understanding its impact on the optimisation process. We now formulate Ω in the language of Hilbert spaces to facilitate the investigation of this term.

2.2 Optimisation Problems in Hilbert Spaces

Specifically, we let

$$R_{\text{reg}} = \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i) + \lambda\langle f, f\rangle. \tag{2.3}$$

and optimise over functions f in some Hilbert space.

Theorem 2.2.1. *Let X be an input space, Y an output space, $T = \{(x_i, y_i)\}_{i=1}^m$ a training sample, $L : Y \times Y \to \mathbb{R}$ a loss function, and H a reproducing kernel Hilbert space for a kernel $k : X \times X \to \mathbb{R}$. Then the minimizer of (2.3), f_T, is of the form*

$$f_T(x) = \sum_{i=1}^{m} \alpha_i k_{x_i}(x)$$

for some $\alpha_i \in \mathbb{R}$.

Remark 2.2.2. *The success of Theorem 2.2.1 is that it reduces the optimisation from a large Hilbert space to the space of finite linear combinations. Although finite, the linear combinations may still be large, however, a suitably precise solution to the optimisation problem can be obtained by truncating the sums at a pre-defined limit. Consequently, we can consider finding an approximate solution on a finite search space.*

2.3 Choosing the Kernel

Despite the statements of Remark 2.2.2 holding for arbitrary kernels used in Theorem 2.2.1, there may be benefits in judiciously choosing the kernel. For example, different kernels may have different regularisation powers and extract functions which represent canonical features of the distribution. Resulting in more efficient optimisation problems and simpler solutions. To investigate the effect of different kernels on the optimisation problem we compare kernels to the usual inner product defined on $L^2(X)$. Specifically, we assume X to be a Lebesgue measurable space. We let F be the reproducing kernel Hilbert space for a kernel k, with the corresponding inner product denoted $\langle \cdot, \cdot \rangle_F$. We then compare $\langle \cdot, \cdot \rangle_F$ to the inner product

$$\langle f, g \rangle_{L^2(X)} = \int_X f(x) g(x)\, dx.$$

Definition 2.3.1. *For a kernel k with reproducing kernel Hilbert space F, define the kernel operator $K : L^2(X) \to L^2(X)$ as*

$$(K(f))(x) = \int k(x, x')\, f(x')\, dx'.$$

For simplicity, we assume that the kernel operator is invertible.

Definition 2.3.2. *For K the kernel operator of some kernel k, we define the regularisation operator $\Upsilon = K^{-\frac{1}{2}}$.*

It follows that

$$\langle f, g \rangle_F = \langle \Upsilon f, \Upsilon g \rangle_{L^2(X)},$$

thus regularising f with k is equivalent to employing L^2 regularisation of Υf.

Example 2.3.3. *A ubiquitous kernel is the Gaussian kernel, given by*

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

where σ is the variance parameter. The reason for its popularity is its connection to Fourier transforms. With $\hat{f}$ denoting the Fourier transform of f and $\hat{\Upsilon}$ denoting the frequency space regularisation operator, given by $\hat{\Upsilon}\hat{f} = \widehat{\Upsilon f}$, it follows that

$$\left(\hat{\Upsilon}\hat{f} \right)(\omega) = \exp\left(|\omega|^2 \sigma^2 \right).$$

Using our discussion above, it follows that regularisation by the Gaussian kernel is equivalent to penalising high-frequency components in f by a factor of $|\omega|^2$.

The kernels used in machine learning are often motivated by their capacity to identify similarities between objects. On the one hand, kernels can be used to facilitate the optimisation process. On the other hand, kernels can be used to expound the underlying processes of the optimisation process, for instance, the neural tangent kernel.

Part II
The Application of Kernels

3 Kernel on Groups

A training sample will likely contain various symmetries which any approximation f should maintain. Such symmetries can be represented in any approximation function by requiring the function to have specific properties. Thus we can perpetuate these symmetries throughout the optimisation procedure by using kernels which induce a Hilbert space of functions possessing the desired properties. As groups are the inherent structure to represent symmetries, it is natural to understand how the theory of groups can be intertwined with kernels and Hilbert spaces more generally.

3.1 Groups Acting on Sets

Definition 3.1.1. *A group G acts on a set X if every group element $g \in G$ we can associate a function $T_g : X \to X$ in such a way that $T_e(x) = x$ for all $x \in X$ and $T_{g_1 g_2} = T_{g_1}(T_{g_2}(x))$ for all $g_1, g_2 \in G$ and $x \in X$.*

We will now consider a group G acting on our input space X by $x \mapsto T_g(x)$. Our interest lies in finding kernels that induce Hilbert spaces of functions that are invariant to this action. That is, $f(T_g(x)) = f(x)$ for all $g \in G$. A canonical example of illustrating why we want to achieve is that of classifying images. Suppose an image classifier is constructed to identify hand-written digits, then one would ideally want this classifier to produce the same output even if the image is slightly translated. In other words, function approximation should possess translation symmetry.

Definition 3.1.2. *Let G be a group acting on the input space X by $x \mapsto T_g(x)$ for all $g \in G$. Then a positive definite kernel $k : X \times X \to \mathbb{R}$ is invariant with respect to this action if*

$$k(x, x') = k(T_g(x), T_g(x'))$$

for all $x, x' \in X$ and $g \in G$.

Example 3.1.3. *Suppose $X \subset \mathbb{R}^n$, and consider the group $\mathbb{R}^n$ acting on X through the action $T_a(x) = x + a$. Then the Gaussian kernel is invariant under this action.*

Theorem 3.1.4. *Let G be a group acting on the input space X by $x \mapsto T_g(x)$ for all $g \in G$. Let k be a positive definite kernel on X inducing the reproducing kernel Hilbert space H of functions satisfying $f(x) = f(T_g(x))$ for $x \in X$ and $g \in G$. Then k is invariant with respect to the action $x \mapsto T_g(x)$.*

Remark 3.1.5. *Theorem 3.1.4 shows that an induced space of invariant functions, with respect to some action, implies that the reproducing kernel is invariant. However, a kernel being invariant is not sufficient to ensure the induced Hilbert space consists of invariant functions.*

Theorem 3.1.6. *Let G be a finite group acting on the space X by $x \mapsto T_g(x)$ with $g \in G$. Let k be a positive definite kernel on X that is invariant to this action and reproducing kernel Hilbert space H. Then*

$$k^G(x, x') = \frac{1}{|G|} \sum_{g \in G} k(x, T_g(x'))$$

is a positive definite function whose reproducing kernel Hilbert space H^G is the subspace of H consisting of functions satisfying $f(x) = f(T_h(x))$ for all $x \in X$ and $h \in G$.

Remark 3.1.7. *To ensure that the induced Hilbert space consists of functions invariant to the action in question Theorem 3.1.6 capitalises on the finiteness of the group to define a positive definite function that is also symmetric.*

3.2 Groups Acting on Homogeneous Spaces

Above we considered groups acting on sets, however, now we would like to further utilise the algebraic structure of a group G and consider groups acting on homogeneous spaces.

Definition 3.2.1. *Let G act on a set X. Then the orbit of $x \in X$ is given by $O(x) := \{T_g(x)\}_{g \in G}$.*

Orbits partition X into disjoint subsets. If an $x_0 \in X$ exists such that $O(x_0) = X$, then G is said to act transitively on X. Moreover, in such a case we call X a homogeneous space of G. By considering G acting on homogeneous spaces, we are ensuring that the algebraic structure of X matches that of G. Henceforth, for simplicity, we will consider G acting on itself through the action $x \mapsto zx$ for $z, h \in G$.

Definition 3.2.2. *For a given G, and a definite kernel $k : G \times G \to \mathbb{R}$ is left-invariant on G if $k(x, y) = k(zx, zy)$ for all $x, y, z \in G$. Similarly, k is right invariant if $k(x, y) = k(xz, yz)$ for all $x, y, z \in G$.*

Remark 3.2.3.

1. *The notion of a definite kernel being right-invariant is only possible as the group is acting on itself.*
2. *If a kernel is both left and right invariant then it is said to be bi-invariant.*

Note that if k is right-invariant then $r(x) = k(x, e)$ fully specifies k as $k(x, y) = k\left(xy^{-1}, e\right)$. Similarly, $r(x) = k(x, e)$ fully specifies left-invariant kernels. Such functions $r : G \to \mathbb{R}$ are positive definite kernels and are called the positive definite functions of G

Theorem 3.2.4. *Let r be a positive definite function on a group G and let $k(x, y) = r\left(xy^{-1}\right)$ be the corresponding right-invariant kernel. Then the following are equivalent.*

1. *k is bi-invariant.*
2. *r is a class function.*
3. *$r(xy) = r(yx)$ for any $x, y \in G$.*
4. *$k\left(x^{-1}, y^{-1}\right) = k(x, y)$.*

Using a kernel k on G, we can define an inner product $\langle e_x, e_y \rangle = k(x, y)$ where e_x denotes the action of x on G. Moreover, we can identify it with the reproducing kernel Hilbert space of k by $k_x(y) = \langle e_x, e_y \rangle$. When we are dealing with a homogeneous space H of G we have the identification

$$e_{xH} = \frac{1}{|H|} \sum_{h \in H} e_{xh}.$$

One can show that if k is right-invariant on G and H is a homogeneous space of G, then

$$k\left(xH, x'H\right) = \frac{1}{|H|} \sum_{h \in H} k\left(x, x'h\right).$$

Note the similarity to the kernel constructed in Theorem 3.1.6.

4 Neural Tangent Kernel

References

[1] Risi Kondor. "Group Theoretical Methods in Machine Learning". PhD thesis. Columbia University, 2008. 111 pp. URL: https://people.cs.uchicago.edu/~risi/papers/KondorThesis.pdf (visited on 12/07/2023).

[2] Thomas Walker. *Function Spaces and Applications*. 2023. 72 pp.