From Zero to Reproducing Kernel Hilbert Spaces in Twelve Pages or Less

Hal Daumé III

11 February 2004

1 Introduction

Reproducing Kernel Hilbert Spaces (RKHS) have been found incredibly useful in the machine learning community. Their theory has been around for quite some time and has been used in the statistics literature for at least twenty years. More recently, their application to perceptron-style algorithms, as well as new classes of learning algorithms (specially large-margin or other regularization machines) has lead to the proliferation of algorithms and software that depend on their nature.

Despite this burgeoning of practical uses, the theory and structure behind the use of the "kernel trick" is often glossed over. This tutorial attempts to take the reader from a very basic understanding of fields through Banach spaces and Hilbert spaces, into Reproducing Kernel Hilbert Spaces. This is very much a "RKHSs without the magic (with the math)" style paper, but every effort has been put in to making the math as clear as possible.

For more information on the use of kernels in machine learning, the reader is referred to the well-known tutorials on support vector machines [3] and gaussian processes [9, 12]. Both SVMs and GPs belong to the class regularization learning machines that take advantage of the "kernel trick" to perform linear learning in non-linear spaces.

2 Fields

The beginning of our voyage will be the field. A field is perhaps the most complex (in terms of operations) basic algebraic structure. A field $\mathbb{F}$ is a structure $\langle F, +, \cdot, 0, 1\rangle$; it consists of a universe F, an addition operation $(+)$, a multiplication operation $(\cdot)$, an identity for addition 0, and an identity for multiplication 1. Furthermore, an inverse operation $-$ must exist for addition and an inverse operation $(\cdot)^{-1}$ must exist for multiplication (for non-zero elements). In addition, these operations must satisfy the following axioms (the Field Axioms), for all elements $a, b, c \in F$:

- Associative Laws:

$$a + (b + c) = (a + b) + c$$

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c$$

- Commutative Laws:

$$a + b = b + a$$

$$a \cdot b = b \cdot a$$

- Distributive Laws:
$$a \cdot (b + c) = a \cdot b + a \cdot c$$
$$(a + b) \cdot c = a \cdot c + b \cdot c$$

- Identity Laws:
$$a + 0 = 0 + a = a$$
$$a \cdot 1 = 1 \cdot a = a$$

- Inverse Laws:
$$a + (-a) = (-a) + a = 0$$
$$a \cdot a^{-1} = a^{-1} \cdot a = 1 \text{ when } a \neq 0$$

Typically the $(\cdot)$ will be omitted for multiplication, as is standard. Furthermore, we will write $a - b$ for $a + (-b)$ and a/b or $\frac{a}{b}$ for $a \cdot (b^{-1})$.

2.1 Examples of Fields

The "smallest" well-known field is the rational field, $\mathbb{Q}$. The reals $\mathbb{R}$ form a larger field than $\mathbb{Q}$ and the complex numbers also form a field. Most of our examples here will have to do with the reals, since they have other nice properties that we will find useful. Note that the integers are not a field, as they do not have a multiplicative inverse.

2.2 Ordered Fields

An ordered field $\mathbb{F}$ is a field with a binary relation $(\leq)$ that is a linear order. In particular, for any $a, b, c \in F$, $(<)$ must obey:

- Reflexive: $a \leq a$
- Antisymetric: $a \leq b \wedge b \leq a \Rightarrow a = b$
- Transitive: $a \leq b \wedge b \leq c \Rightarrow a \leq c$

As is standard, when $a \leq b \wedge a \neq b$ we will write $a < b$.

All of the fields mentioned above ($\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$) are ordered fields with the usual ordering.

2.3 Complete Ordered Fields

A complete ordered field is the first not completely straightforward structure we will encounter. Completeness is important to us because it enables us to define the notion of a limit. Informally, a space is complete if every (infinite) sequence of its elements that approaches a particular value has this value as it's "limit" and this limit is in the space itself.

The formal definition is that X is a complete space if every Cauchy sequence in X is convergent.

A Cauchy sequence is a sequence $\langle x_i \rangle_{i=0}^{\infty}$ such that for every real number $\epsilon > 0$ we can find a natural number N such that $d(x_n, x_m) < e$ whenever $n, m > N$. Here, d is a distance metric on X; for instance, when X is $\mathbb{R}$,

$d(x, y) = |x - y|$. This basically says that we can take an arbitrarily small value for ϵ and are guarnteed that after some point (N), all later values of x are no further apart than ϵ.

The definition of a convergent sequence is very similar. A sequence $\langle x_i \rangle_{i=0}^{\infty}$ is convergent in X if there is a point $x \in X$ such that for every real number $\epsilon > 0$ we can find a natural number N such that $d(x, x_n) < \epsilon$ for all $n > N$. This says that for any convergent sequence, we can find some value x that is *in the original space* that is arbitrarily close to x_n for all n after a certain point.

By way of example, $\mathbb{Q}$ is *not* a complete ordered field. For instance, the sequence represnting the decimal expansion of $\sqrt{2}$ is a Cauchy sequence in $\mathbb{Q}$ which does not converge in $\mathbb{Q}$. On the other hand, $\mathbb{R}$ is a complete ordered field.

2.4 Isomorphisms

The concept of an isomorphism is prevalent in all areas of math. It is essentially a mapping between two objects that preserves all the relevant properties of those objects.

In the case of fields $\mathbb{F}$ and $\mathbb{G}$, we say that ϕ is an isomorphism between $\mathbb{F}$ and $\mathbb{G}$ if ϕ is a function from $F \to G$ and ϕ obeys certain properties:

- Injective (one-to-one): for all $f, f' \in F$, $\phi(f) = \phi(f')$ implies that $f = f'$ (i.e., there is at most one element in F which maps to a single element in G).

- Surjective (onto): for all $g \in G$ there exists $f \in F$ such that $\phi(f) = g$ (i.e., there is at least one element in F which maps to a single element in G). The combination of these first two properties states that ϕ is a bijection.

- Preservation: basically, ϕ preserves operations. That is, for example, $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(ab) = \phi(a)\phi(b)$. The image of the indentities of F must be the identities of G.

Extending the notion of isomorphism to ordered fields simply involves adding a clause to the preservation statement that ϕ preserves relative ordering. It is worth noting that $\mathbb{R}$ is the unique (up to isomorphism) complete ordered field (i.e., for any complete ordered field $\mathbb{F}$, we can find an isomorphism ϕ between $\mathbb{F}$ and $\mathbb{R}$).

2.5 Further Reading

Almost any book on algebra will describe the above topics in much more detail, together with various theorems and proofs about them. I am rather partial to [5] as a reference, but [7] (and others) will also contain this information.

For the more foundationally-inclined, several books on set theory and logic will describe the construction of $\mathbb{R}$ as a complete field, beginning from axiomatic set theory (typically the Zermelo-Frankel axioms). My personal favorite is [6], but [10] also describes this process in the first few sections.

3 Vector Spaces

A vector space is, in short, a space that contains elements called "vectors" and supports two kinds of operations: addition of vectors and multiplication by scalars. The scalars are drawn from some field and the vector space is a vector space *over* that field.

More formally, let $\mathbb{F}$ be a field. $\mathbb{V}$ is a vector space over $\mathbb{F}$ if $\mathbb{V}$ is a structure of the form $\langle V, \mathbb{F}, \oplus, \otimes, \ominus, 0_{\mathbb{V}}\rangle$ consisting of a universe V, a vector addition operation $\oplus$, a scalar multiplication operation $\otimes$, a unary additive inverse operation $\ominus$ and an identity element $0_{\mathbb{V}}$. This structure must obey the following axioms for any $u, v, w \in V$ and $a, b \in F$:

- Associative Law: $(u \oplus v) \oplus w = u \oplus (v \oplus w)$
- Commutative Law: $u \oplus v = v \oplus u$
- Inverse Law: $u \oplus (\ominus u) = 0_{\mathbb{V}}$
- Identity Laws:
$$0_{\mathbb{V}} \oplus u = u$$
$$1 \otimes u = u$$
- Distributive Laws:
$$a \otimes (b \otimes u) = (ab) \otimes u$$
$$(a + b) \otimes u = a \otimes u \oplus b \otimes u$$

As is standard, we will omit $\otimes$ and write $a \otimes u$ as au and will write $\oplus$ as $+$ and $\ominus$ as $-$. We will write 0 for $0_{\mathbb{V}}$ (context should be sufficient to distinguish this from $0 \in \mathbb{F}$) and $u \oplus (\ominus v)$ as $u - v$.

3.1 Examples

The simplest example of a vector space is just $\mathbb{R}$ itself, which is a vector space of $\mathbb{R}$. Vector addition and scalar multiplication are just addition and multiplication on $\mathbb{R}$. A slightly more complex, but still very familiar example is $\mathbb{R}^n$, n-dimensional vectors of real numbers. Addition is defined point-wise and scalar multiplication is defined by multiplying each element in the vector by the scalar (using standard multiplication in $\mathbb{R}$).

Another example is $\mathbb{R}^{\mathbb{R}}$, the set of functions from $\mathbb{R} \to \mathbb{R}$. We can define addition and multiplication point-wise, by $(fg)(x) = f(x)g(x)$ and $(af)(x) = af(x)$ for $f, g \in \mathbb{R}^{\mathbb{R}}$ and $a \in \mathbb{R}$. This is therefore a vector space over $\mathbb{R}$. In fact, for any reasonable space, X (i.e., one which has a distance metric d defined – a so-called metric space), the set $\mathbb{R}^X$ of functions from $X \to \mathbb{R}$ is a vector space over $\mathbb{R}$ with addition and multiplication defined as above. In particular, X could be $\mathbb{R}^n$.

Taking all functions from X to $\mathbb{R}$ is often admitting too many "weird" things, so we will typically limit ourselves to continuous functions. Recall that a function $f : X \to \mathbb{R}$ is continuous at a point $x_0 \in X$ if for every $\epsilon > 0$ we can find a $\delta > 0$ such that $d(x, x_0) < \delta$ implies $|f(x) - f(x_0)| < \epsilon$. A function is continuous if it is continuous everywhere. The set of continuous functions from X to $\mathbb{R}$ is commonly denoted $C(X)$ and forms a vector space over $\mathbb{R}$ using the usual definitions of addition and scalar multiplication.

3.2 Further Reading

Most books on algebra or analysis will cover the definition of vector spaces (sometimes called *modules*). From an algebraic perspective, [5] is a good start, though vector spaces are introduced very late. From an analytical perspective, I am rather fond of [8], though there are of course many other great texts.

4 Banach Spaces

A Banach space is a complete vector space $\mathbb{V}$ endowed with a method of calculating the "size" of vectors in V. This is called the norm and the norm of $v \in V$ is written $||v||$. Often a single vector space will have multiple norms, in which case we will write $||v||_{(\cdot)}$ with $(\cdot)$ replaced with a reasonable identifier to specify which norm we are talking about. If $\mathbb{V}$ is finite, there is at most one norm, but for infinite $\mathbb{V}$ there are often many.

4.1 Complete Vector Spaces

Just as we have defined a field to be complete if every Cauchy sequence is convergent, so we also define a complete vector space. The definition is the same, with the exception that the distance metric $d(x, y)$ is replaced by $||x-y||$, where $||\cdot||$ is a suitable norm.

4.2 Norms

A norm is a function on a vector space $\mathbb{V}$ over $\mathbb{R}$ from V to $\mathbb{R}$ satisfying the following properties for all $u, v \in V$ and all $a \in \mathbb{R}$:

- Non-negative: $||u|| \geq 0$
- Strictly positive: $||u|| = 0$ implies $u = 0$
- Homogenous: $||au|| = |a|||u||$
- Triangle inequality: $||u + v|| \leq ||u|| + ||v||$

4.3 Examples

As we have seen, $\mathbb{R}$ is a complete vector space over $\mathbb{R}$, which makes it a suitable candidate for being a Banach space. The most common norm for $\mathbb{R}$ is the absolute norm: $||x|| = |x|$.

It is easy to see that $\mathbb{R}^n$ is a complete vector space over $\mathbb{R}$ for any $n > 0$. There are several (actually an infinite number) of norms we can define on $\mathbb{R}^n$. The Euclidean-norm (also called the 2-norm) is defined as:

$$||\langle x_i \rangle_{i=1}^n||_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

In general, we define the p-norm (for $p \geq 1$) in $\mathbb{R}^n$ by:

$$||\langle x_i \rangle_{i=1}^n||_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

The 1 norm is also called the sum norm. The maximum norm, or the ∞ norm, is defined by:

$$||\langle x_i \rangle_{i=1}^{n}||_{\infty} = max\{|x_1|, |x_2|, \ldots, |x_n|\}$$

4.3.1 Infinite Sequences

Certain sequences of infinite length, often denoted $\mathbb{R}^{<\omega}$ or $\mathbb{R}^{\infty}$ also form a Banach space. Norms here are defined similarly to the case of sequences of finite length; however, in order to ensure completeness, we need to make sure that these sequences don't diverge under summing. For a positive integer p, we define the l_p space as:

$$l_p = \left\{ \langle x_i \rangle_{i=0}^{\infty} : \sum_{i=0}^{\infty} |x_i|^p < \infty \right\}$$

Given this definition, we define a norm on l_p ("little ell p") by:

$$||\langle x_i \rangle_{i=0}^{\infty}||_{l_p} = \left(\sum_{i=0}^{\infty} |x_i|^p \right)^{\frac{1}{p}}$$

4.3.2 Norms on Function Spaces

For continuous functions from X to $\mathbb{R}$ (i.e., $C(x)$), the natural norm (called the uniform norm or the sup norm) is defined by:

$$||f||_{sup} = \sup_{x \in X} |f(x)|$$

I.e., this is the highest (supremum) value that f takes on all of X. This is analogous to the ∞ norm defined above for sequences and is also denoted $||\cdot||_{\infty}$.

We will define the notion of an L_p ("ell p") over functions from $\mathbb{R}^n$ to $\mathbb{R}$. A more general notion is possible, but needs too much in the way of background knowledge, which we elect not to go in to. Using the standard notion of integration of $\mathbb{R}^n$, we define:

$$L_p = \{(f : \mathbb{R}^n \to \mathbb{R}) : \int_{-\infty}^{\infty} |f^p(x)| \mathrm{d}x < \infty\}$$

We define a norm on L_p by:

$$||f||_{L_p} = \left(\int_{-\infty}^{\infty} |f^p(x)| \mathrm{d}x \right)^{\frac{1}{p}}$$

It is relatively easy to see that this satisfies the needed requirements.

4.4 Further Reading

The definition and properties of l_p and L_p spaces can be found in most texts on analysis; I favor [8], but [1] also contains the relevant information. These books will all define the norm general notion of L_p, based on the theory of measures, which essentially extends integration to metric spaces other than $\mathbb{R}$. A good general reference for measure theory is [4].

5 Hilbert Spaces

A Hilbert space is a Banach space further endowed with a dot-product operation. We will typically denote Hilbert spaces by $\mathcal{H}$. For elements $u, v \in \mathcal{H}$, we will write the dot product of u and v either as $\langle u, v\rangle_{\mathcal{H}}$ or, when it is clear by context that the dot product is taking place in $\mathcal{H}$, either $\langle u, v\rangle$ or simply $u \cdot v$. For when $\mathcal{H}$ is a vector space over $\mathbb{F}$, then the result of the dot product will be an element in F. Since we will typically deal with $\mathbb{F} = \mathbb{R}$, the result of dot products will be real numbers. The dot product operation must satisfy a few properties, for all $u, v, w \in \mathcal{H}$ and all $a \in \mathbb{F}$:

- Associative: $(au) \cdot v = a(u \cdot v)$
- Commutative: $u \cdot v = v \cdot u$
- Distributive: $u \cdot (v + w) = u \cdot v + u \cdot w$

Given a complete vector space $\mathbb{V}$ with a dot product $\langle \cdot, \cdot\rangle_{\mathbb{V}}$, we can easily define a norm on $\mathbb{V}$ by $||u||_{\mathbb{V}} = \sqrt{\langle u, u\rangle}$, thus making this space into a Banach space and therefore into a full Hilbert space. It is not the case (as we shall shortly see) that all Banach spaces can be made into Hilbert spaces.

5.1 Examples

As usual, $\mathbb{R}$ and $\mathbb{R}^n$ are both Hilbert spaces. In the latter case, the dot product is defined by:

$$\langle x, y\rangle_{\mathbb{R}^n} = \sum_{i=1}^{n} x_i y_i$$

A similar definition is given for infinite sequences. We can similarly define a dot product of functions from $\mathbb{R}^n$ to $\mathbb{R}$:

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x) g(x) \mathrm{d}x$$

As you can see, both of these dot product definitions induce the 2 norm on their respective spaces. However, it can be seen (though we won't show it) that there is no dot product corresponding to l_p or L_p spaces for $p \neq 2$.

5.2 Further Reading

Any text which discusses Banach spaces will also discuss Hilbert spaces; as before, see [8].

6 Reproducing Kernel Hilbert Spaces

This section is the goal of this tutorial and will thus be the most in-depth of all. In case you skipped directly here, or if you need a quick refresher, let's first recall what we've done up until now.

6.1 Refresher

We first defined a field, which is a space that supports the usual operations of addition, subtraction, multiplication and division. We imposed an ordering on the field and described what it means for a field to be complete (this will become important soon – if you missed this section, reread it). We then defined vector spaces over fields, which are spaces that interact in a friendly way with their associated fields. We defined complete vector spaces and extended them to Banach spaces by adding a norm. Banach spaces were then extended to Hilbert spaces with the addition of a dot product.

6.2 Reproducing Kernels

A Reproducing Kernel Hilbert Space (RKHS) again builds on a Hilbert space $\mathcal{H}$ and requires that all Dirac evaluation functionals in $\mathcal{H}$ are bounded and continuous (though one implies the other). We will define these concepts in order.

For now, we will assume that $\mathcal{H}$ is the $L_2(X)$ space of functions from X to $\mathbb{R}$ for some measurable X (typically X will be $\mathbb{R}^n$ for some n). For an element $x \in X$, a Dirac evaluation functional at x is an functional $\delta_x \in \mathcal{H}$ such that $\delta_x(f) = f(x)$. In our case, x will be some vector a real numbers and f will be a function from this vector space into $\mathbb{R}$. Then δ_x is simply a function which maps f to the value f has at x. Thus, δ_x is a function from $(\mathbb{R}^n \to \mathbb{R})$ into $\mathbb{R}$.

To say that δ_x is bounded means that there is a constant $M > 0 \in \mathbb{R}$ such that for all $f \in \mathcal{H}$, $||\delta_x f|| \leq M||f||$. Note that the first norm is the norm in $\mathbb{R}$ and thus is just the absolute value. The second norm is the norm in L_2 and is thus the integral equation from before. Expanding this by the definition of δ_x, we get that we require that there is an M such that for all $x \in \mathbb{R}^n$, for all $f \in \mathcal{H}$, $|f(x)| \leq M \int f(x)\mathrm{d}x$.

This is a somewhat technical definition, but it is easy to verify for the spaces we care about. The importance of it is due to the Riesz representation theorem, which states that if ϕ is a bounded linear functional (conditions satisfied by the Dirac evaluation functionals) on a Hilbert space $\mathcal{H}$, then there is a unique vector u in $\mathcal{H}$ such that $\phi f = \langle f, u \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

Translating this theorem back into Dirac evaluation functionals, we get that for each δ_x, there is a unique vector (which we will denote k_x) in $\mathcal{H}$ such that $\delta_x f = f(x) = \langle f, k_x \rangle_{\mathcal{H}}$. Using this, we can define the *Reproducing Kernel* K for $\mathcal{H}$ by: $K(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$, where k_x and $k_{x'}$ are respectively the unique representatives of δ_x and $\delta_{x'}$.

The property of reproducing kernels that we need is that $\langle f, K(x, x') \rangle_{\mathcal{H}} = f(x')$. Furthermore, k_x is defined to be the function $y \mapsto K(x, y)$ and thus $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$. This is why they are called reproducing kernels.

6.3 Positive Definiteness

We will take a short break from reproducing kernels and define what it means for a function to be positive definite. This is a simple extension of the definition with the same name for matrices. A symmetric function K is positive definite if for any L_2 function f (other than the zero function), we have that:

$$\int \int f(x) K(x, x') f(x') \mathrm{d}x \mathrm{d}x' > 0$$

This generalizes the definition for matrices since it implies that for any finite subset of X, we get that the matrix $\mathcal{K}$ defined by $(\mathcal{K})_{i,j} = K(x_i, x_j)$ is positive definite. It turns out that all reproducing kernels are positive definite.

6.4 Creating RKHSs

We have seen that all reproducing kernels are positive definite; in fact, any positive definite function *is* a reproducing kernel for some RKHS, a concept we will make more formal now.

Essentially, we will assume we have a p.d. kernel K and will show how to fashion a RKHS $\mathcal{H}_K$ such that K is the reproducing kernel on $\mathcal{H}$. It turns out that for a given K, $\mathcal{H}_K$ is unique up to isomorphism. The construction is fairly straightforward. First we define the universe V of $\mathcal{H}$ and then define a dot product on it. This will induce a norm and thus give us an RKHS. It will be easy to show that K is the reproducing kernel of this RKHS. We assume that K is a kernel over X.

We define the universe V of $\mathcal{H}_K$ first by taking $S = \{k_x : x \in X\}$, where k_x is the function such that $k_x(y) = K(x, y)$ and then defining V to be the set of all linear combinations of elements from S. Therefore, each element of V, the universe of $\mathcal{H}_K$, can be written as $\sum_i \alpha_i k_x$. We define the dot product on $\mathcal{H}_K$ by:

$$\langle k_x, k_y \rangle_{\mathcal{H}_K} = \left\langle \sum_i \alpha_i k_{x_i}, \sum_i \beta_i k_{y_i} \right\rangle_X$$

for some vectors α and β. Due to the reproducing property of K, we can write this dot product in X by:

$$\sum_i \sum_j \alpha_i \beta_j K(x_i, y_j)$$

We should note that V is not necessarily complete. However, we can force it to be complete by simply taking all Cauchy sequences over V and adding their limits. Importantly, we can write the differences of functions from V pointwise, as can be seen:

$$|f_n(x) - f_m(x)| = |K(x, \cdot), f_n - f_m| \leq K(x, x)||f_n - f_m||_2$$

Because of this, we can just take pointwise limits and add them to V.

We now need to define the dot product operation on $\mathcal{H}_K$. In order to do this, we need to ensure that K is continuous and doesn't diverge, namely:

$$\int \int K^2(x, x')\mathrm{d}x\mathrm{d}x' < \infty$$

This property is known as finite trace. It is worth noting, though beyond the scope here, that if K does not have a finite trace, we can restrict ourselves to a specific subset of the space X and ensure that K has finite trace on that subspace.

In order to define the dot product, we need to introduce the concept of eigenfunctions.

6.4.1 Eigenfunctions

An eigenfunction is the functions-space equivalent of an eigenvector. Recall from linear algebra that an eigenvector of a matrix M is a vector v such that $vA = \lambda v$ for some scalar λ. $lambda$ is then called the corresponding eigenvector.

A similar statement can be made about functions. Suppose K is a kernel; then, ϕ is an eigenfunction of K if:

$$\int K(x, x')\phi(x')\mathrm{d}x' = \lambda\phi(x')$$

for all x. In dot product notation, this corresponds to the notion that $\langle K(x, \cdot), \phi\rangle_X = \lambda\phi$, which is nearly an identical statement to the one for matrices.

The Mercer-Hilbert-Schmit theorems state that if K is a positive definite kernel (that is continuous with finite trace), then there exists an infinite sequence of eigenfunctions $\langle\phi_i\rangle_{i=0}^{\infty}$ and eigenvalues λ_i with $\lambda_1 \geq \lambda_2 \geq \ldots$ of K, and that we can write K as:

$$K(x, x') = \sum_{i=0}^{\infty} \lambda_i \phi_i(x)\phi_i(x')$$

This is analogous to the expression of a matrix in terms of of its eigenvectors and eigenvalues, except in this case we have functions and an infinity of them.

6.4.2 Defining the Dot Product

We now will assume that for our kernel K, we have out set of eigenfunctions ϕ_i and eigenvalues λ_i. For $y \in L_2$, we will denote y in terms of its coefficients in the eigenfunctions:

$$y_i = \langle y, \phi_i\rangle_{L_2} = \int y(x)\phi_i(x)\mathrm{d}x$$

It is a basic result of Fourier analysis that such a representation exists and is unique. Given all this, we are ready to define our inner product:

$$\langle y, y'\rangle_{\mathcal{H}_K} = \sum_{i=0}^{\infty} \frac{y_i y_i'}{\lambda_i}$$

Though it may seem very round-about, this is very similar to the definition of a dot product in $\mathbb{R}^n$. In $\mathbb{R}^n$ we will choose n-many basis vectors, e_i, such that all entries in e_i are zero, except for the ith component, which is 1. This is the standard orthonormal basis for $\mathbb{R}^n$. In this case, the corresponding eigenvalues are all $\lambda_i = 1$. Just as above, we can identify any vector y in $\mathbb{R}^n$ as a linear combination of the e_is, with coefficients y_i as we did above for eigenfunctions. The dot product expression in $\mathbb{R}^n$ is the identical to the above, with the exception that the sum is now finite, from 1 to n.

6.5 Feature Spaces

We have seen just now that given a p.d. kernel K over X, we can find a Hilbert space $\mathcal{H}$ with reproducing kernel K. We will now show that in addition to finding $\mathcal{H}$, we can find a feature function $\Phi : X \to \mathcal{H}$ such that:

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

This property is really the one we have been looking for and is the main result we will use. This says, very simply, that **given a symmetric p.d. function K, there exists a function Φ such that the evaluation of the kernel at points x and x' is equivalent to taking the dot product between $\Phi(x)$ and $\Phi(x')$ in some (perhaps unknown) Hilbert space.**

We will think of Φ as a mapping from an input space X to a large, possibly infinite feature space $\mathcal{H}$, where we can compute dot products simply by computing K. This enables us to perform the kernel trick, in which dot products are replaced by kernel products (i.e., evaluations of kernels). Doing so is well-motivated by this result, since we can just say that we're not actually taking a kernel product; instead, we transforming the inputs into $\mathcal{H}$ and then taking the dot product as before.

We should recall that we have seen how we can explicitly construct $\mathcal{H}$, and that $\mathcal{H}$ is unique up to isomorphism. This means in turn that Φ is not absolutely unique, but it is just as unique as $\mathcal{H}$ is. In fact, we will show the two most common constructions of Φ, which are more or less equivalent.

First, we will note that no matter how we choose Φ, it will always be injective (one-to-one). Otherwise, two different values x and x' will yield the same value $K(x, \cdot)$ and $K(x', \cdot)$. This is not possible, since we require the matrix defined over x, x' using K to be p.d., which it will not be if Φ is not injective.

The first definition we give of Φ will use the space $\mathcal{H}_K$ as the feature space. We then simply define $\Phi(x) = K(x, \cdot)$. By the reproducing property of the kernel, we get:

$$\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}_K} = \langle K(x, \cdot), K(x', \cdot) \rangle_{\mathcal{H}_K} = K(x, x')$$

which satisfies our requirements for Φ.

We can instead ignore our explicitly constructed $\mathcal{H}_K$ all together and use l_2 as the feature space. This construction uses the eigenfunctions ϕ_i and eigenvalues λ_i and defines Φ by:

$$\Phi(x) = \left\langle \sqrt{\lambda_i}\phi_i(x) \right\rangle_{i=0}^{\infty}$$

Now, we can calculate the dot product by:

$$\begin{aligned}
\langle \Phi(x), \Phi(x') \rangle_{l_2} &= \left\langle \left\langle \sqrt{\lambda_i}\phi_i(x) \right\rangle_i, \left\langle \sqrt{\lambda_i}\phi_i(x') \right\rangle_i \right\rangle_{l_2} \\
&= \sum_i \sqrt{\lambda_i}\phi_i(x)\sqrt{\lambda_i}\phi_i(x') \\
&= \sum_i \lambda_i \phi_i(x)\phi_i(x') \\
&= K(x, x')
\end{aligned}$$

also as desired.

6.6 Further Reading

Good references on RKHSs are hard to find, which is part of the motivation for this document. The Riesz representation theorem is described in most books on functional analysis, including [8]. The original reference for reproducing kernels is [2], but is highly technical and hard to find. Other references include [1] and [11].

References

[1] N. Akhiezer and I. Glazman. *Theory of Linear Operators in Hilbert Space*. Ungar, New York, 1963.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematics Society*, 68:337 – 404, 1950.

[3] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[4] Joseph Doob. *Measure Theory*. Springer Verlag, 1994.

[5] David S. Dummit and Richard M. Foote. *Abstract Algebra*. John Wiley and Sons, second edition edition, 1999.

[6] Karel Hrbacek and Thomas Jech. *Introduction to Set Theory*. Marcel Dekker, third edition edition, 1984.

[7] Thomas W. Hungerford. *Algebra*. Springer Verlag, eighth edition edition, 1997.

[8] John K. Hunter and B. Nachtergaele. *Applied Analysis*. World Scientific Publishing Co., 2001.

[9] D.J.C. MacKay. Introduction to Gaussian Processes. In C.M. Bishop, editor, *Neural Networks and Machine Learning*, volume F 168, pages 133 – 165. Springer, 1998.

[10] Yiannis N. Moschovakis. *Notes on Set Theory*. Springer Verlag, 1994.

[11] Matthias Seeger. Relationships between gaussian processes, support vector machines an smoothing splines. Technical report, University of California at Berkeley, 1999.

[12] Christopher K. I. Williams and Carl Edward Rasmussen. Gaussian processes for regression. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8. MIT Press, 1995.