

1 RHKS for a general kernel

Let k be a kernel on the input space \mathcal{X} . We want to construct a Hilbert space \mathcal{F} and a mapping $\phi : \mathcal{X} \rightarrow \mathcal{F}$, such that

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle \quad (1)$$

for all $x_1, x_2 \in \mathcal{X}$. Let us start by defining a function f_x for all $x \in \mathcal{X}$ by

$$f_x : \mathcal{X} \rightarrow \mathbb{R}, \quad f_x(x') = k(x, x') \quad (2)$$

If we think of the kernel as a similarity measure between points in the input space, then f_x measures how similar the other points in the input space is to x . We can now consider the set of functions

$$F(\mathcal{X}, k) = \{f_x \mid x \in \mathcal{X}\} \quad (3)$$

and ask ourself, if it is a Hilbert space regardless of which kernel, we used to define it? Not necessarily. In fact, we do not even know, if it is a vector space. We can turn it into a vector space by taking its span:

$$S(\mathcal{X}, k) = \text{span}(F(\mathcal{X}, k)) \quad (4)$$

$$= \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f = \sum_{i=1}^m a_i f_{x_i}, \quad m \in \mathbb{N}, \quad a_i \in \mathbb{R}, \quad x_i \in \mathcal{X}\} \quad (5)$$

$$= \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f(x) = \sum_{i=1}^m a_i k(x_i, x), \quad m \in \mathbb{N}, \quad a_i \in \mathbb{R}, \quad x_i \in \mathcal{X}\} \quad (6)$$

$S(\mathcal{X}, k)$ consists of any real function on the input space \mathcal{X} that can be written as a weighted average of f_{x_1}, \dots, f_{x_m} for some finite sequence of points x_1, \dots, x_m in the input space \mathcal{X} . If we think of the kernel as a similarity measure between points in the input space, then $S(\mathcal{X}, k)$ consists of any real function f on the input space \mathcal{X} , where the value $f(x)$ for all $x \in \mathcal{X}$ can be written as a weighted average of how similar x is to each of x_1, \dots, x_m for some finite sequence of points in \mathcal{X} .

By definition of $S(\mathcal{X}, k)$ we know that some subset of $F(\mathcal{X}, k)$ is a basis of $S(\mathcal{X}, k)$. If $F(\mathcal{X}, k)$ is a linearly independent set, then $F(\mathcal{X}, k)$ is a basis of $S(\mathcal{X}, k)$. If $F(\mathcal{X}, k)$ is infinite and linearly independent, then we have that $S(\mathcal{X}, k)$ is infinite dimensional. If $F(\mathcal{X}, k)$ is finite, or some finite subset of $F(\mathcal{X}, k)$ still spans $S(\mathcal{X}, k)$, then we have that $S(\mathcal{X}, k)$ is finite dimensional.

$S(\mathcal{X}, k)$ is a vector space, and we can turn it into an inner product space by defining for all $f, g \in S(\mathcal{X}, k)$ that if

$$f(x) = \sum_{i=1}^m a_i k(x_i, x) \quad (7)$$

and

$$g(x) = \sum_{j=1}^n b_j k(x'_j, x) \quad (8)$$

are representations of f and g ¹, then

$$\langle f, g \rangle = \sum_i^m \sum_j^n a_i b_j k(x_i, x'_j) \quad (9)$$

Lets call inner product space we get by imposing this inner product on $S(\mathcal{X}, k)$ for $\tilde{S}(\mathcal{X}, k)$. Since $\tilde{S}(\mathcal{X}, k)$ is an inner product space, Cauchy sequences are defined in $\tilde{S}(\mathcal{X}, k)$. We get the completion of $\tilde{S}(\mathcal{X}, k)$ by including the limit of any Cauchy sequence as a point in the completed space:

$$\mathcal{F} = \overline{\tilde{S}(\mathcal{X}, k)} \quad (10)$$

This completed space is now a Hilbert space. We can define a function $\phi : \mathcal{X} \rightarrow \mathcal{F}$ by simply saying

$$\phi(x) = f_x \quad (11)$$

By construction we now get that for all $x, x' \in \mathcal{X}$:

$$\langle \phi(x), \phi(x') \rangle = \langle f_x, f_{x'} \rangle = \sum_{i=1}^1 \sum_{j=1}^1 1 \cdot 1 \cdot k(x, x') = k(x, x') \quad (12)$$

since we know by construction that f_x can be written as

$$f_x(y) = k(x, y) \quad (13)$$

and $f_{x'}$ can be written as

$$f_{x'}(y) = k(x', y) \quad (14)$$

¹Although the definition of the inner product uses specific representations of f and g , its easy to proof that the value of $\langle f, g \rangle$ is independent of the concrete representations, we use to calculate the value.

2 RKHS for a Gaussian kernel

Let $\mathcal{X} = \mathbb{R}^n$ for some $n \in \mathbb{N}$. Let k_γ be the Gaussian kernel with spread parameter γ , that is for all $x, x' \in \mathcal{X}$:

$$k_\gamma(x, x') = e^{-\gamma\|x-x'\|^2} \quad (15)$$

We define for all $x \in \mathcal{X}$:

$$f_x : \mathcal{X} \rightarrow \mathbb{R}, \quad f_x(x') = k_\gamma(x, x') = e^{-\gamma\|x-x'\|^2} \quad (16)$$

and the set of functions

$$F(\mathcal{X}, k_\gamma) = \{f_x \mid x \in \mathcal{X}\} = \{f(x') = e^{-\gamma\|x-x'\|^2} \mid x \in \mathcal{X}\} \quad (17)$$

In other words, $F(\mathcal{X}, k_\gamma)$ is the set of n -dimensional Gaussian density functions with all covariances set to 0, all variances set to $\frac{1}{\gamma}$, and some $x \in \mathbb{R}^n$ as mean.

We can now consider the vector space:

$$S(\mathcal{X}, k_\gamma) = \text{span}(F(\mathcal{X}, k_\gamma)) \quad (18)$$

$$= \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f(x) = \sum_{i=1}^m a_i k_\gamma(x_i, x), \ m \in \mathbb{N}, \ a_i \in \mathbb{R}, \ x_i \in \mathcal{X}\} \quad (19)$$

$$= \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f(x) = \sum_{i=1}^m a_i e^{-\gamma\|x-x_i\|^2}, \ m \in \mathbb{N}, \ a_i \in \mathbb{R}, \ x_i \in \mathcal{X}\} \quad (20)$$

My guess is that this must be the vector space of all n -dimensional Gaussian density distributions with all covariances set to 0 and equal variances in all n dimensions. However, I am not sure that this is the case.

Since the function set $F(\mathcal{X}, k_\gamma)$ is a linearly independent, spanning subset of $S(\mathcal{X}, k_\gamma)$, and $F(\mathcal{X}, k_\gamma)$ is infinite, we get that $S(\mathcal{X}, k_\gamma)$ is infinite dimensional. The intuitive reason that $F(\mathcal{X}, k_\gamma)$ is linearly independent is that all the Gaussian density functions in $F(\mathcal{X}, k_\gamma)$ have the same covariance matrix and different means. If we should build one density function by a weighted average of some of the rest of the density functions, we would need to add the weighted, chosen density functions, which means that the resulting function would have a covariance matrix with larger variances. Therefore it would not be in $F(\mathcal{X}, k_\gamma)$ at all. In other words, $F(\mathcal{X}, k_\gamma)$ is linearly independent, because the weighted sum of any of its members falls outside of $F(\mathcal{X}, k_\gamma)$.

$S(\mathcal{X}, k_\gamma)$ is a vector space, and we can turn it into an inner product space by defining for all $f, g \in S(\mathcal{X}, k_\gamma)$ that if

$$f(x) = \sum_{i=1}^m a_i e^{-\gamma \|x - x_i\|^2} \quad (21)$$

and

$$g(x) = \sum_{j=1}^n b_j e^{-\gamma \|x - x_j\|^2} \quad (22)$$

are representations of f and g , then

$$\langle f, g \rangle = \sum_i^m \sum_j^n a_i b_j e^{-\gamma \|x_j - x_i\|^2} \quad (23)$$

3 Why god, why?

Why is this way of viewing kernels interesting, when we never use anything from \mathcal{F} in our learning algorithms except for the value of the inner products calculated via the kernel. One reason is that we can prove that certain properties are true for our learning algorithm, if the dot product on \mathbb{R}^n is replaced by any inner product on some feature space \mathcal{F} . Therefore, we need to know that the kernel taken on x and x' is an inner product on the transformations of these points in some transformed feature space. To make this very general, we need to use function spaces as our feature space instead of just scalar spaces.