Notes on RKHS

immediate

Wednesday 26th September, 2018

1 Introduction

Good references include Wahba (1990), Gu (2013), Andreas Christmann (2008).

Functional analysis: Vershynin (2010). Functional principal component analysis: Phd dissertation Amini (2013).

2 Positive semi-definite kernel

This section is adapted from Hofmann et al. (2008).

We consider a general index set $\mathcal{T}$. A *kernel* is a function $R(\cdot, \cdot)$ from $\mathcal{T} \times \mathcal{T}$ to $\mathbb{R}$ which is symmetric: $R(s,t) = R(t,s)$. A kernel $R(\cdot, \cdot)$ is said to be *positive semi-definite* (PSD) if, for any real $a_1, \ldots, a_n$ and $t_1, \ldots, t_n \in \mathcal{T}$,

$$\sum_{i,j=1}^{n} a_i a_j R(t_i, t_j) \geq 0.$$

In other words, for any $t_1, \ldots, t_n \in \mathcal{T}$ the matrix

$$\begin{pmatrix} R(t_1,t_1) & \ldots & R(t_1,t_n) \\ \ldots & \ldots & \ldots \\ R(t_n,t_1) & \ldots & R(t_n,t_n) \end{pmatrix}$$

is PSD.

Let $(\mathbb{H}, (\cdot,\cdot))$ be an inner product space. It is clear that for any function $\phi : \mathcal{T} \to \mathbb{H}$, $R(s,t) := (\phi(s), \phi(t))$ is a PSD kernel. If $R(\cdot,\cdot)$ is so defined, $\phi(\cdot)$ is called its *feature map* and $\mathbb{H}$ is called the *feature space*. We note that the feature map of a kernel is not unique. We shall return to this point later.

The following proposition gives other ways to construct PSD kernels.

Proposition 1. *Suppose $R_i(\cdot,\cdot)$, $i = 1, 2, \ldots$, are PSD kernels on $\mathcal{T} \times \mathcal{T}$. Then*

(i) If $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 R_1(\cdot,\cdot) + \alpha_2 R_2(\cdot,\cdot)$ is a PSD kernel.

(ii) *If $R(s,t) := \lim_{n\to\infty} R_n(s,t)$ exists for all $s,t \in \mathcal{T}$, then $R(\cdot,\cdot)$ is a PSD kernel.*

(iii) *The pointwise product $R_1 \circ R_2(s,t) := R_1(s,t)R_2(s,t)$ is a PSD kernel.*

(iv) *Assume that for $i = 1, 2$, $R_i(\cdot,\cdot)$ is a PSD kernel on $\mathcal{T}_1 \times \mathcal{T}_2$. Then the tensor product, which is a function defined on $(\mathcal{T}_1 \times \mathcal{T}_2) \times (\mathcal{T}_1 \times \mathcal{T}_2)$ as*

$$R_1 \otimes R_2\left((s_1,s_2),(t_1,t_2)\right) := R_1(s_1,t_1)R_2(s_2,t_2)$$

is a PSD kernel. The direct sum

$$R_1 \oplus R_2\left((s_1,s_2),(t_1,t_2)\right) := R_1(s_1,t_1) + R_2(s_2,t_2)$$

is a PSD kernel.

Proof. These results follows directly from the analogous results in matrix theory. For example, (iii) follows from that the Hadamard product of two PSD matrix is PSD. As for (iv), for any real $a_1,\ldots,a_n$ and $(t_{11},t_{12}),\ldots,(t_{n1},t_{n2}) \in \mathcal{T}_1 \times \mathcal{T}_2$,

$$\begin{aligned}\sum_{i,j=1}^{n} a_i a_j R_1 \oplus R_2((t_{i1},t_{i2}),(t_{j1},t_{j2})) &= \sum_{i,j=1}^{n} a_i a_j \left(R_1(t_{i1},t_{j1}) + R_2(t_{i2},t_{j2})\right) \\ &= \sum_{i,j=1}^{n} a_i a_j R_1(t_{i1},t_{j1}) + \sum_{i,j=1}^{n} a_i a_j R_2(t_{i2},t_{j2}) \geq 0.\end{aligned}$$

On the other hand

$$\sum_{i,j=1}^{n} a_i a_j R_1 \otimes R_2((t_{i1},t_{i2}),(t_{j1},t_{j2})) = \sum_{i,j=1}^{n} a_i a_j R_1(t_{i1},t_{j1})R_2(t_{i2},t_{j2}) \geq 0$$

since it corresponds to the Hadamard product of

$$\begin{pmatrix} R_1(t_{11},t_{11}) & \ldots & R_1(t_{11},t_{n1}) \\ \ldots & \ldots & \ldots \\ R_1(t_{n1},t_{11}) & \ldots & R_1(t_{n1},t_{n1}) \end{pmatrix}$$

and

$$\begin{pmatrix} R_2(t_{12},t_{12}) & \ldots & R_2(t_{12},t_{n2}) \\ \ldots & \ldots & \ldots \\ R_2(t_{n2},t_{12}) & \ldots & R_1(s_{n2},t_{n2}) \end{pmatrix}.$$

□

Example 1 (Gaussian kernel)**.** *Let $\mathcal{T} = \mathbb{R}^p$. Then $R(s,t) = s^\top t$ is a PSD kernel. By (i) and (iii) of Proposition 1,*

$$R_n(s,t) := \sum_{i=0}^{n} \frac{R(s,t)^n}{n!}$$

is a PSD kernel. Then by (ii) of Proposition 1, $\exp(R(s,t)) := \exp(s^\top t)$ *is a PSD kernel. By (v) of Proposition 1,* $\exp(-\|s\|^2 - \|t\|^2)$ *is a PSD kernel. Thus, again by (iii) of Proposition 1,* $\exp(-\|t-s\|^2)$ *is a PSD kernel.*

A PSD kernel $R(\cdot,\cdot)$ *is call* radial *if* $R(x,y) = g(\|x-y\|)$ *for some function* $g : [0,+\infty) \to \mathbb{R}$.

Example 2 (Polynomial kernels)**.** *Let* $\mathcal{T} = \mathbb{R}^p$*. From (iii) of Proposition 1 it is clear that homogeneous polynomial kernels* $R(s,t) = (s^\top t)^n$ *are PSD for* $n \in \mathbb{N}$*. We can also explicitly give the corresponding feature map:*

$$R(s,t) = (s^\top t)^n = (\sum_{i=1}^p s_i t_i)^n = \sum_{i_1=1}^p \sum_{i_2=1}^p \cdots \sum_{i_n=1}^p (s_{i_1} \cdots s_{i_n}) \cdot (t_{i_1} \cdots t_{i_n}) = (C_{n,p}(s), C_{n,p}(t)),$$

where $C_{n,p}(\cdot)$ *maps* $t \in \mathbb{R}^p$ *to a* p^n *dimensional vector whose entries are all possible nth degree ordered products of the entries of* t*. Other useful kernels include the inhomogeneous polynomial* $R(s,t) = (s^\top t + c)^n$ *where* $n \in \mathbb{N}$ *and* $c \geq 0$*.*

3 Reproducing kernel Hilbert space

This section is adapted from Wahba (1990), Andreas Christmann (2008), Hsing and Eubank (2015). A *Hilbert space* $(\mathbb{H}, (\cdot,\cdot))$ is a complete vector space with an inner product. An important example of Hilbert space is the class of all square integrable measurable functions $L^2(\mathbb{X}, \mathcal{B}, \mu)$ on a measurable space $(\mathbb{X}, \mathcal{B}, \mu)$. A continuous linear functional (or bounded linear functional) L is a linear map from $\mathbb{H}$ into $\mathbb{R}$ such that

$$|Lf| \leq M\|f\| \text{ for all } f \in \mathcal{H}.$$

For each $y \in \mathbb{H}$, the map $x \mapsto (x,y)$ is a continuous linear functional, denoted as $(\cdot, y)$. A fundamental result in real analysis says all continuous linear functional can be represented by $(\cdot, y)$ for some $y \in \mathbb{H}$.

Theorem 1 (Riesz-Fréchet)**.** *A map* L *from a Hilbert space* $\mathbb{H}$ *into* $\mathbb{R}$ *is a continuous linear functional if and only if for some* $y \in \mathbb{H}$*,* $Lx = (x,y)$ *for all* $x \in \mathbb{H}$*. If so, then* y *is unique.*

Let $\mathbb{R}^\mathcal{T}$ denote the space of all real functions from $\mathcal{T}$ to $\mathbb{R}$. Suppose $\mathbb{H}$ is a subset of $\mathbb{R}^\mathcal{T}$ and $(\mathbb{H}, (\cdot,\cdot)_\mathbb{H})$ is a Hilbert space. Then it can be seen that the for each $t \in \mathcal{T}$, the coordinate projection L_t, defined as $L_t f(\cdot) := f(t)$, is a linear functional. Note that L_t is not necessarily continuous. In fact, if $\mathbb{H} = L^2(\mathbb{R}, \mathcal{B}, \mu)$, then L_t is not continuous since $f(t)$ can be arbitrarily defined.

Definition 1. Suppose $\mathbb{H}$ is a subset of $\mathbb{R}^\mathcal{T}$ and $(\mathbb{H}, (\cdot,\cdot)_\mathbb{H})$ is a Hilbert space. Then $\mathbb{H}$ is called a reproducing kernel Hilber space (RKHS) if for each $t \in \mathcal{T}$, the coordinate projection L_t is a continuous linear functional.

Every RKHS has a unique reproducing kernel, as stated by the following theorem.

Theorem 2. *Let $\mathbb{H} \subset \mathbb{R}^{\mathcal{T}}$ be an RKHS. Then there is a unique function $R(\cdot,\cdot) : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, such that*

1. *for every $t \in \mathcal{T}$, $R(\cdot, t) \in \mathbb{H}$ and*
2. *R satisfies the reproducing property that for every $f(\cdot) \in \mathbb{H}$ and $t \in \mathcal{T}$*

$$f(t) = (f(\cdot), R(\cdot, t))_{\mathbb{H}}.$$

Furthermore, the function $R(\cdot,\cdot)$ is a PSD kernel. $R(\cdot,\cdot)$ is called the reproducing kernel *of $\mathbb{H}$.*

Proof. For each $t \in \mathcal{T}$ there exists, by Theorem 1, a unique element $R_t(\cdot) \in \mathbb{H}$ with the property

$$L_t f(\cdot) = (f(\cdot), R_t(\cdot))_{\mathbb{H}} = f(t), \quad \forall f \in \mathbb{H}.$$

Let $f(\cdot) = R_s(\cdot)$ in the above equality, we have

$$(R_s(\cdot), R_t(\cdot))_{\mathbb{H}} = R_s(t), \quad \forall s, t \in \mathcal{T}.$$

Since the inner product is symmetric, $(R_s(\cdot), R_t(\cdot))_{\mathbb{H}} = (R_t(\cdot), R_s(\cdot))_{\mathbb{H}}$. Hence

$$(R_s(\cdot), R_t(\cdot))_{\mathbb{H}} = R_t(s), \quad \forall s, t \in \mathcal{T}.$$

Thus, $R_s(t) = R_t(s)$. For any real $a_1, \ldots, a_n$ and $t_1, \ldots, t_n \in \mathcal{T}$,

$$\sum_{i,j=1}^{n} a_i a_j R_{t_i}(t_j) = \sum_{i,j=1}^{n} a_i a_j \left(R_{t_i}(\cdot), R_{t_j}(\cdot)\right)_{\mathbb{H}} = \left(\sum_{i}^{n} a_i R_{t_i}(\cdot), \sum_{i}^{n} a_j R_{t_j}(\cdot)\right)_{\mathbb{H}} \geq 0.$$

It follows that $R_{\cdot}(\cdot)$ is a positive semi-definite kernel. Since $R_{\cdot}(\cdot)$ is symmetric, it can be written as $R(\cdot,\cdot)$. □

Conversely, given a positive-definite kernel $R(\cdot,\cdot)$, there is a unique RKHS with R as its reproducing kernel.

Theorem 3 (Moore-Aronszajn)**.** *Suppose that $R(\cdot,\cdot) : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a PSD kernel. Then there is a unique RKHS $\mathbb{H} \subset \mathbb{R}^{\mathcal{T}}$ with $R(\cdot,\cdot)$ as its reproducing kernel.*

Proof. Set

$$\mathbb{H}_0 := \operatorname{span}\{R(\cdot, t) : t \in \mathcal{T}\} = \left\{ \sum_{i=1}^{n} a_i R(\cdot, t_i) \Big| n = 1, 2, \ldots, a_i \in \mathbb{R}, t_i \in \mathcal{T} \right\}.$$

Clearly $\mathbb{H}_0$ is a linear space. Define inner product $(\cdot,\cdot)_{\mathbb{H}_0}$ on $\mathbb{H}_0$ as

$$\left(\sum_{i=1}^{n_1} a_i R(\cdot, s_i), \sum_{j=1}^{n_2} b_j R(\cdot, t_j)\right)_{\mathbb{H}_0} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} a_i b_j R(s_i, t_j). \tag{1}$$

This definition is indeed feasible since if $\sum_{i=1}^{n_1} a_i R(\cdot, s_i) = \sum_{i=1}^{n_1'} a_i' R(\cdot, s_i')$ and $\sum_{j=1}^{n_2} b_j R(\cdot, t_j) = \sum_{j=1}^{n_2'} b_j' R(\cdot, t_j')$, then

$$\begin{aligned}\left(\sum_{i=1}^{n_1'} a_i R(\cdot, s_i'), \sum_{j=1}^{n_2'} b_j' R(\cdot, t_j')\right)_{\mathbb{H}_0} &= \sum_{i=1}^{n_1'}\sum_{j=1}^{n_2'} a_i' b_j' R(s_i', t_j') = \sum_{j=1}^{n_2'} b_j' \left(\sum_{i=1}^{n_1'} a_i' R(s_i', t_j')\right) \\ &= \sum_{j=1}^{n_2'} b_j' \left(\sum_{i=1}^{n_1} a_i R(s_i, t_j')\right) = \sum_{i=1}^{n_1} a_i \left(\sum_{j=1}^{n_2'} b_j' R(s_i, t_j')\right) \\ &= \sum_{i=1}^{n_1} a_i \left(\sum_{j=1}^{n_2} b_j R(s_i, t_j)\right) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} a_i b_j R(s_i, t_j).\end{aligned}$$

Now we check that $(\cdot,\cdot)_{\mathbb{H}_0}$ so defined is indeed an inner product. Clearly, $(\cdot,\cdot)_{\mathbb{H}_0}$ is bilinear and symmetric. The assumption that $R(\cdot,\cdot)$ is PSD ensures that $(f,f)_{\mathbb{H}_0} \geq 0$ for $f \in \mathbb{H}_0$. Hence $(\cdot,\cdot)_{\mathbb{H}_0}$ is a semi inner product. So it suffices to verify that $(f,f)_{\mathbb{H}_0} = 0$ implies $f = 0$. Note that the definition (1) implies that $(f, R(\cdot,t))_{\mathbb{H}_0} = f(t)$ for every $t \in \mathcal{T}$. Then for every $t \in \mathcal{T}$,

$$|f(t)| = |(f, R(\cdot,t))_{\mathbb{H}_0}| \leq \sqrt{(f,f)_{\mathbb{H}_0}(R(\cdot,t), R(\cdot,t))_{\mathbb{H}_0}} = 0.$$

Thus, $(\mathbb{H}_0, (\cdot,\cdot)_{\mathbb{H}_0})$ is an inner product space with reproducing kernel $R(\cdot,\cdot)$. However, $\mathbb{H}_0$ itself may not be complete and hence may not be a Hilbert space.

Now we proceed to complete $\mathbb{H}_0$. Suppose $\{f_n(\cdot)\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathbb{H}_0$. Since

$$|f_n(t) - f_m(t)| = |(f_n - f_m, R(\cdot,t))_{\mathbb{H}_0}| \leq \|f_n - f_m\|_{\mathbb{H}_0} \|R(\cdot,t)\|_{\mathbb{H}_0},$$

$\{f_n(t)\}$ is a Cauchy sequence in $\mathbb{R}$. Therefore, $\{f_n\}_{n=1}^{\infty}$ has a pointwise limit. Define

$$\mathbb{H} = \{f(\cdot) |\ f \text{ is the pointwise limit of some Cauchy sequence in } \mathbb{H}_0\}.$$

Clearly $\mathbb{H}$ is a linear space. Let $f(\cdot), g(\cdot) \in \mathbb{H}$, then there exist Cauchy sequences $\{f_n(\cdot)\}$ and $\{g_n(\cdot)\}$ such that $f(t) = \lim f_n(t)$, $g(t) = \lim g_n(t)$, for all $t \in \mathcal{T}$. Define $(f(\cdot), g(\cdot))_{\mathbb{H}} = \lim_{n\to\infty} (f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0}$. This limit exists since $(f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0}$ is a Cauchy sequence in $\mathbb{R}$:

$$\begin{aligned}&\left|(f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0} - (f_m(\cdot), g_m(\cdot))_{\mathbb{H}_0}\right| \\ =&\left|(f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0} - (f_n(\cdot), g_m(\cdot))_{\mathbb{H}_0}\right| + \left|(f_n(\cdot), g_m(\cdot))_{\mathbb{H}_0} - (f_m(\cdot), g_m(\cdot))_{\mathbb{H}_0}\right| \\ \leq&\|f_n(\cdot)\|_{\mathbb{H}_0}\|g_n - g_m\|_{\mathbb{H}_0} + \|g_m(\cdot)\|_{\mathbb{H}_0}\|f_n - f_m\|_{\mathbb{H}_0}.\end{aligned}$$

Also, this limit only depends on the limits $f(\cdot)$ and $g(\cdot)$. To see this, suppose there exist other Cauchy sequences $\{f_n'(\cdot)\}$ and $\{g_n'(\cdot)\}$ such that $f(t) = \lim f_n'(t)$, $g(t) = \lim g_n'(t)$, for all $t \in \mathcal{T}$. Let $f_n^{(d)}(t) = f_n(t) - f_n'(t)$. Then $\{f_n^{(d)}(\cdot)\}_{n=1}^{\infty}$ is also a Cauchy sequence in $\mathbb{H}_0$ and $f_n^{(d)}(t) \to 0$ for all $t \in \mathcal{T}$. We would like to show that $\|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0} \to 0$. In fact,

$$\limsup_{n\to\infty} \limsup_{m\to\infty} \|f_n^{(d)}(\cdot) - f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2 \to 0.$$

But

$$\begin{aligned}\limsup_{m\to\infty}\|f_n^{(d)}(\cdot)-f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2 &= \limsup_{m\to\infty}\left(\|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}^2+\|f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2-2(f_n^{(d)}(\cdot),f_m^{(d)}(\cdot))_{\mathbb{H}_0}\right)\\ &= \|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}^2+\limsup_{m\to\infty}\|f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2 \qquad (2)\\ &\geq \|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}^2,\end{aligned}$$

where the second last equality holds since for fixed $f_n^{(d)}(\cdot):=\sum_{i=1}^k a_i R(\cdot,t_i)\in\mathbb{H}_0$,

$$\limsup_{m\to\infty}(f_n^{(d)}(\cdot),f_m^{(d)}(\cdot))_{\mathbb{H}_0}=\limsup_{m\to\infty}\sum_{i=1}^k a_i(R(\cdot,t_i),f_m^{(d)}(\cdot))_{\mathbb{H}_0}=\limsup_{m\to\infty}\sum_{i=1}^k a_i f_m^{(d)}(t_i)=0.$$

Let n tends to infinity in (2), we have $\|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}=\|f_n-f_n'\|_{\mathbb{H}_0}\to 0$. Similarly, we have $\|g_n-g_n'\|_{\mathbb{H}_0}\to 0$. Thus,

$$|(f_n(\cdot),g_n(\cdot))_{\mathbb{H}_0}-(f_n'(\cdot),g_n'(\cdot))_{\mathbb{H}_0}|\leq\|f_n(\cdot)\|_{\mathbb{H}_0}\|g_n-g_n'\|_{\mathbb{H}_0}+\|g_n'(\cdot)\|_{\mathbb{H}_0}\|f_n-f_n'\|_{\mathbb{H}_0}\to 0.$$

Hence

$$\lim_{n\to\infty}(f_n(\cdot),g_n(\cdot))_{\mathbb{H}_0}=\lim_{n\to\infty}(f_n'(\cdot),g_n'(\cdot))_{\mathbb{H}_0}$$

and $(f(\cdot),g(\cdot))_{\mathbb{H}}$ is well defined.

Its not hard to verify that $(\cdot,\cdot)_{\mathbb{H}}$ is an inner product in $\mathbb{H}$. By definition, for any $f(\cdot)\in\mathbb{H}$, there exsits a Cauchy sequence $\{f_n(\cdot)\}_{i=1}^\infty$ in $\mathbb{H}_0$ such that $f_n(t)\to f(t)$ for all $t\in\mathcal{T}$. We would like to show that $\|f_n-f\|_{\mathbb{H}}\to 0$. In fact

$$\begin{aligned}\|f_n-f\|_{\mathbb{H}}^2 =&\|f_n\|_{\mathbb{H}_0}^2+\|f\|_{\mathbb{H}}^2-2(f_n(\cdot),f(\cdot))_{\mathbb{H}}\\ =&\|f_n\|_{\mathbb{H}_0}^2+\lim_{m\to\infty}\|f_m\|_{\mathbb{H}_0}^2-2\lim_{m\to\infty}(f_n(\cdot),f_m(\cdot))_{\mathbb{H}_0}\\ =&\lim_{m\to\infty}\|f_n-f_m\|_{\mathbb{H}_0}^2.\end{aligned}$$

Let n tends to infty, we have $\lim_{n\to\infty}\|f_n-f\|_{\mathbb{H}}^2=0$.

For each Cauchy sequence $\{f_n(\cdot)\}_{i=1}^\infty$ in $\mathbb{H}$, there exists, by the above argument, $\{f_n'(\cdot)\}_{i=1}^\infty$ in $\mathbb{H}_0$ such that $\|f_n'-f_n\|\leq n^{-1}$. Hence $\{f_n'(\cdot)\}_{i=1}^\infty$ is also Cauchy, and thus converges to some $f(\cdot)$ in $\|\cdot\|_{\mathbb{H}}$. And thus $\{f_n(\cdot)\}_{i=1}^\infty$ also converges to $f(\cdot)$ in $\|\cdot\|_{\mathbb{H}}$.

We have proved that $(\mathbb{H},(\cdot,\cdot)_{\mathbb{H}})$ is complete. Its easy to verify that for every $f(\cdot)\in\mathbb{H}$, $(f(\cdot),R(\cdot,t))=f(t)$ for $t\in\mathcal{T}$. Thus $R(\cdot,\cdot)$ is the reproducing kernel of $\mathbb{H}$.

We turn to the uniqueness of $\mathbb{H}$. Any RKHS $\tilde{\mathbb{H}}$ must contain $R(\cdot,t)$, hence contain $\mathbb{H}_0$. $\mathbb{H}$ is the closure of $\mathbb{H}_0$ in $\tilde{\mathbb{H}}$. Hence also $\mathbb{H}\subset\tilde{\mathbb{H}}$. We have $\tilde{\mathbb{H}}=\mathbb{H}\oplus\mathbb{H}^\perp$. If $\mathbb{H}^\perp\neq\emptyset$, let $f\in\mathbb{H}^\perp$. Then $f(t)=(f,R(\cdot,t))_{\mathbb{H}}=0$ for every $t\in\mathbb{T}$, a contradiction. This completes the proof. □

We have shown that there is a one-to-one correspondence between kernels and RKHSs. For any kernel $R(\cdot,\cdot):\mathcal{T}\times\mathcal{T}\to\mathbb{R}$, let $\mathbb{H}$ be its RKHS. Define $\phi(\cdot):\mathcal{T}\to\mathbb{R}$ as $\phi(t)=R(\cdot,t)$. Then by reproducing property, $(\phi(s),\phi(t))_{\mathbb{H}}=R(s,t)$. That is, $\phi(\cdot)$ is a feature map of kernel $R(\cdot,\cdot)$. We call $\phi(\cdot)$ the *cononical feature map* Andreas Christmann (2008).

4 Kernel methods in machine learning

The earliest application of kernels and RKHS in machine learning is *kernel support vector machine*. A rich and comprehensive reference book of this topic is Andreas Christmann (2008). Since the introduction of kernel support vector machine, many kernels for specific learning tasks have been developed. See Hofmann et al. (2008) for a review.

4.1 Kernel principal component analysis

Principal component analysis (PCA) is a powerful technique for extracting structure from possibly high-dimensional data sets. Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be a sample drawn from distribution P with zero mean and covariance matrix Σ. Then $\mathbf{S} = n^{-1} \sum_{i=1}^n X_i X_i^\top$ is an estimator of Σ. Denote by λ_i $(i = 1, \ldots, p)$ the ith largest eigenvalue of $\mathbf{S}$, and by v_i the corresponding eigenvector. We have

$$\mathbf{S} v_i = \lambda_i v_i \quad i = 1, \ldots, p.$$

Then the first few eigenvectors $v_1, \ldots, v_k$ $(k < p)$ capture the most variation. For a new data X_{new}, its principal component is $(v_1^\top X_{\text{new}}, \ldots, v_k^\top X_{\text{new}})^\top \in \mathbb{R}^k$.

The kernel PCA algorithm is proposed by Schölkopf et al. (1998). It compute principal components in feature spaces. Consider the (measurable) feature map $\phi : \mathbb{R}^p \to \mathbb{H}$ where $\mathbb{H}$ is a (possibly infinite-dimensional) Hilbert space. The sample covariance function of $\phi(X_1), \ldots, \phi(X_n)$ is defined as $\mathbf{S} = n^{-1} \sum_{i=1}^n \phi(X_i) \otimes \phi(X_i)$. By definition, the tensor product $\phi(X_i) \otimes \phi(X_i)$ is an operator from $\mathbb{H}$ to $\mathbb{H}$ and $\phi(X_i) \otimes \phi(X_i)(x) = (\phi(X_i), x)_{\mathbb{H}} \phi(X_i)$ for every $x \in \mathbb{H}$. The image of $\mathbf{S}$ is finite dimensional and hence $\mathbf{S}$ is a self-adjoint compact operator Hsing and Eubank (2015). It follows that $\mathbf{S}$ has an eigen decomposition

$$\mathbf{S} = \sum_{i=1}^p \lambda_i v_i \otimes v_i,$$

where $\lambda_1 \geq \cdots \geq \lambda_p$ and v_i's are orthonormal: $(v_i, v_j)_{\mathbb{H}} = 0$ if $i \neq j$ and 1 if $i = j$. For a new data X_{new}, its principal component is $(v_1^\top \phi(X_{\text{new}}), \ldots, v_k^\top \phi(X_{\text{new}}))^\top \in \mathbb{R}^k$.

Since $\mathbf{S}$ is infinite dimensional, it is hard to directly compute its eigen decomposition. Note that for $i = 1, \ldots, p$,

$$(\mathbf{S}, v_i)_{\mathbb{H}} = \lambda_i v_i. \tag{3}$$

The left hand of (3) is

$$n^{-1} \sum_{k=1}^n (\phi(X_k) \otimes \phi(X_k), v_i)_{\mathbb{H}} = n^{-1} \sum_{k=1}^n (\phi(X_k), v_i)_{\mathbb{H}} \phi(X_k)$$

which is a linear combinition of $\phi(X_k)$. Hence we can write $v_i = \sum_{j=1}^n a_{ij} \phi(X_j)$. Then

$$(\phi(X_j), (\mathbf{S}, v_i)_{\mathbb{H}})_{\mathbb{H}} = \lambda_i (\phi(X_j), v_i)_{\mathbb{H}} \quad j = 1, \ldots, p.$$

4.2 Kernel Fisher discriminent analysis

Muller et al. (2001)

5 Mercer's thoerem

6 Functional data analysis

Hsing and Eubank (2015)

7 Gaussian processes

Example 3 (RKHS of gaussian processes)**.**

Mercer's theorem.

8 Infinite dimensional exponential model

References

Amini, A. A. (2013). High-dimensional principal component analysis. *Dissertations & Theses - Gradworks*, 61(3):464–473.

Andreas Christmann, I. S. a. (2008). *Support Vector Machines*. Information science and statistics. Springer-Verlag New York, 1 edition.

Gu, C. (2013). *Smoothing Spline ANOVA Models*. Springer Series in Statistics 297. Springer-Verlag New York, 2 edition.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Muller, K. ., Mika, S., Ratsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201.

Schölkopf, B., Smola, A., and Mller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Vershynin, R. (2010). Lectures in functional analysis. Technical report, Department of Mathematics, University of Michigan.

Wahba, G. (1990). *Spline Models for Observational Data (CBMS-NSF Regional Conference Series in Applied Mathematics)*. SIAM: Society for Industrial and Applied Mathematics.