MLSS 2019 London

Kernels
Part II: Reproducing Kernel Hilbert Spaces (RKHS)

Lorenzo Rosasco
MaLGa- Machine learning Genova Center
Universitá di Genova
MIT
IIT

Linear models $\mapsto$features $\mapsto$kernels

$$\mathbf{x}^\top \mathbf{x}' \quad \mapsto \quad \Phi(\mathbf{x})^\top \Phi(\mathbf{x}') \quad \mapsto \quad \mathbf{k}(\mathbf{x}, \mathbf{x}')$$

"The kernel trick"

Outline

PD kernels

RKHS

A bit more than a trick...

- Feature maps
- PD kernels
- RKHS

Can we start from a kernel?

Can we start from a kernel?

What is a kernel?

$$\mathbf{k}(\mathbf{x}, \mathbf{x}')$$

Can we start from a kernel?

What is a kernel?

$$\mathbf{k}(\mathbf{x}, \mathbf{x}')$$

Two words:

Positive definite

Positive definite kernels

$k : X \times X \to \mathbb{R}$ is **positive definite** if:

1. for all $x_1, \dots, x_N \in X$, the $N \times N$ matrix $\widehat{K}$ with entries

$$\widehat{K}_{ij} = \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j)$$

   is positive semidefinite (non negative eigenvalues)

Positive definite kernels

$k : X \times X \to \mathbb{R}$ is **positive definite** if:

1. for all $x_1, \ldots, x_N \in X$, the $N \times N$ matrix $\widehat{K}$ with entries
$$\widehat{K}_{ij} = k(x_i, x_j)$$
   is positive semidefinite (non negative eigenvalues)
2. Equivalently for all $x_1, \ldots, x_N \in X$, and $a \in \mathbb{R}^N$
$$a^\top \widehat{K} a \geqslant 0, \qquad \forall a \in \mathbb{R}^N.$$

Positive definite kernels

$k : X \times X \to \mathbb{R}$ is **positive definite** if:

1. for all $x_1, \dots, x_N \in X$, the $N \times N$ matrix $\widehat{K}$ with entries
$$\widehat{K}_{ij} = k(x_i, x_j)$$
is positive semidefinite (non negative eigenvalues)
2. Equivalently for all $x_1, \dots, x_N \in X$, and $a \in \mathbb{R}^N$
$$a^\top \widehat{K} a \geqslant 0, \qquad \forall a \in \mathbb{R}^N.$$
3. Equivalently, for all $a_1, \dots, a_N \in \mathbb{R}$, $x_1, \dots, x_N \in X$,
$$\sum_{i,j=1}^{N} k(x_i, x_j) a_i a_j \geqslant 0.$$

Positive definite kernels

$k : X \times X \to \mathbb{R}$ is **positive definite** if:

1. for all $x_1, \dots, x_N \in X$, the $N \times N$ matrix $\widehat{K}$ with entries
$$\widehat{K}_{ij} = k(x_i, x_j)$$
is positive semidefinite (non negative eigenvalues)
2. Equivalently for all $x_1, \dots, x_N \in X$, and $a \in \mathbb{R}^N$
$$a^\top \widehat{K} a \geqslant 0, \qquad \forall a \in \mathbb{R}^N.$$
3. Equivalently, for all $a_1, \dots, a_N \in \mathbb{R}$, $x_1, \dots, x_N \in X$,
$$\sum_{i,j=1}^{N} k(x_i, x_j) a_i a_j \geqslant 0.$$

Symmetry is also always assumed.

Inner product kernels are PD

For $\Phi : X \to \mathbb{R}^p$, $p \leqslant \infty$ let

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\top \Phi(\mathbf{x}')$$

UniGe | MaLGa

Inner product kernels are PD

For $\Phi : X \to \mathbb{R}^p$, $p \leqslant \infty$ let

$$k(x, x') = \Phi(x)^\top \Phi(x')$$

- Then

$$\sum_{i,j=1}^{n} k(x_i, x_j) a_i a_j = \sum_{i,j=1}^{n} \Phi(x_i)^\top \Phi(x_j) a_i a_j = \|\sum_{i=1}^{n} \Phi(x_i) a_i\|^2 \geqslant 0.$$

Inner product kernels are PD

For $\Phi : X \to \mathbb{R}^p$, $p \leqslant \infty$ let

$$k(x, x') = \Phi(x)^\top \Phi(x')$$

- Then

$$\sum_{i,j=1}^{n} k(x_i, x_j) a_i a_j = \sum_{i,j=1}^{n} \Phi(x_i)^\top \Phi(x_j) a_i a_j = \|\sum_{i=1}^{n} \Phi(x_i) a_i\|^2 \geqslant 0.$$

- Clearly k is symmetric.

Kernel properties

Let $K_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_3 : \mathbb{R}^t \times \mathbb{R}^t$ be kernels, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^t$ and $\alpha, \beta > 0$ then the following are also kernels

1. $\alpha K_1(\mathbf{x}, \mathbf{x}') + \beta K_2(\mathbf{x}, \mathbf{x}')$
2. $K_1(\mathbf{x}, \mathbf{x}')K_2(\mathbf{x}, \mathbf{x}')$

Kernel properties

Let $K_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_3 : \mathbb{R}^t \times \mathbb{R}^t$ be kernels, $x, x' \in \mathbb{R}^d$, $z, z' \in \mathbb{R}^t$ and $\alpha, \beta > 0$ then the following are also kernels

1. $\alpha K_1(x, x') + \beta K_2(x, x')$
2. $K_1(x, x')K_2(x, x')$
3. $p(K_1(x, x'))$ for any p a function whose polynomial expansion has only non-negative coefficients
4. $f(x)K_1(x, x')f(x')$ for any $f : \mathbb{R}^d \to \mathbb{R}$
5. $\frac{K_1(x,x')}{\sqrt{K_1(x,x)K_1(x',x')}}$
6. $K_3(\psi(x), \psi(x))$ for any $\psi : \mathbb{R}^d \to \mathbb{R}^t$
7. $\alpha K_1(x, x') + \beta K_3(z, z')$
8. $K_1(x, x')K_3(z, z')$

There are many PD kernels

All the examples seen so far are based on inner products.

- linear $k(x, x') = x^\top x'$
- polynomial $k(x, x') = (x^\top x' + 1)^s$
- Gaussian $k(x, x') = e^{-\|x-x'\|^2 \gamma}$

X need not be $\mathbb{R}^d$, one can consider

- kernels on probability distributions
- kernels on strings
- kernels on functions
- kernels on groups
- kernels graphs
- ...

Convolution kernels

- Sets $X, X_1, \dots, X_D$ and $X^D = X_1 \times X_2 \times \dots X_D$.
- Pd kernels $k_1 : X_1 \times X_1 \to \mathbb{R}, \quad \dots \; k_D : X_D \times X_D \to \mathbb{R}$
- Relation $R : X^D \times X \to \{0, 1\}$ and

$$R^{-1}(x) = \{x^D \in X^D \mid R(x^D, x) = 1\}$$

Interpretation

- $x^D = (x_1, \dots, x_D) \in X^D$ are the possible parts of x.
- $R(x^D, x) = 1$ if x^D are the parts of x

Example $X = X_1 = X_2$ strings over a finite alphabet.

$$R((x_1, x_2), x) = 1 \quad \text{iff} \quad x_1 \circ x_2 = x.$$

Convolution kernels

- Sets $X, X_1, \dots, X_D$ and $X^D = X_1 \times X_2 \times \dots X_D$.
- Pd kernels $k_1 : X_1 \times X_1 \to \mathbb{R}, \ \dots \ k_D : X_D \times X_D \to \mathbb{R}$
- Relation $R : X^D \times X \to \{0, 1\}$ and

$$R^{-1}(x) = \{x^D \in X^D \mid R(x^D, x) = 1\}$$

Interpretation

- $x^D = (x_1, \dots, x_D) \in X^D$ are the possible parts of x.
- $R(x^D, x) = 1$ if x^D are the parts of x

Convolution kernel

$$k(x, z) = \sum_{x^D \in R^{-1}(x), z^D \in R^{-1}(z)} \prod_{j=1}^{D} k_j(x_j, z_j)$$

A bit more than a trick...

- Feature maps
- PD kernels
- RKHS

Outline

PD kernels

RKHS

Function spaces

$$\Phi \quad \mapsto \quad \mathbf{f}(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x})$$

$$\mathbf{k} \quad \mapsto \quad \mathbf{f}(\mathbf{x}) = \quad ?$$

Function spaces

$$\Phi \quad \mapsto \quad \mathcal{H}_\Phi = \{\mathsf{f} : \mathsf{X} \to \mathbb{R} \mid \exists!\, \mathsf{w} \in \mathcal{F} \ \text{ s.t. } \ \mathsf{f}(\mathbf{x}) = \mathsf{w}^\top \Phi(\mathbf{x}), \ \ \forall \mathbf{x} \in \mathcal{X}\}$$

$$\mathbf{k} \quad \mapsto \quad \mathcal{H}_\mathbf{k} = ?$$

Hilbert spaces

Hilbert space H

- Linear space (closed under sum/multiplication with reals)

$$h_1, h_2 \in \mathcal{H}, a, b \in \mathbb{R} \qquad \Rightarrow \qquad ah_1 + bh_2 \in H$$

[1]There are no holes!

Hilbert spaces

Hilbert space H

- Linear space (closed under sum/multiplication with reals)

$$h_1, h_2 \in \mathcal{H}, a, b \in \mathbb{R} \qquad \Rightarrow \qquad ah_1 + bh_2 \in H$$

- With an inner product (positive and symmetric bilinear form)

$$h_1^\top h_2$$

[1]There are no holes!

Hilbert spaces

Hilbert space H

- Linear space (closed under sum/multiplication with reals)

$$h_1, h_2 \in \mathcal{H}, a, b \in \mathbb{R} \qquad \Rightarrow \qquad ah_1 + bh_2 \in H$$

- With an inner product (positive and symmetric bilinear form)

$$h_1^\top h_2$$

- complete (Cauchy sequences converge)[1]

$$(h_j)_j \quad i, j > N, \quad \|h_j - h_i\| \leqslant \epsilon \qquad \Rightarrow \qquad \lim_{j \to \infty} h_j = h \in H$$

[1] There are no holes!

Function spaces

$$\Phi \quad \mapsto \quad \mathcal{H}_\Phi = \{\mathbf{f} : \mathbf{X} \to \mathbb{R} \mid \exists!\, \mathbf{w} \in \mathcal{F} \;\; \text{s.t.} \;\; \mathbf{f}(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}), \;\; \forall \mathbf{x} \in \mathcal{X}\}$$

It's a Hilbert space since $\mathcal{F}$ is!

$$\mathbf{k} \quad \mapsto \quad \mathcal{H}_\mathbf{k} = ?$$

From PD kernels to function spaces

For a set X and PD kernel k:

- define the linear the space of functions[2]

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

for any $a_1, \ldots, a_N \in \mathbb{R}$, $x_1, \ldots, x_N \in X$ and any $N \in \mathbb{N}$.

[2] Any set of inputs not just data.

From PD kernels to function spaces

For a set X and PD kernel k:

- define the linear the space of functions[2]

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

  for any $a_1, \ldots, a_N \in \mathbb{R}$, $x_1, \ldots, x_N \in X$ and any $N \in \mathbb{N}$.

- define the inner product

$$\langle f, f' \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N'} k(x_i, x_j') a_i a_j'.$$

[2]Any set of inputs not just data.

From PD kernels to function spaces

For a set X and PD kernel k:

- define the linear the space of functions[2]

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

  for any $a_1, \dots, a_N \in \mathbb{R}$, $x_1, \dots, x_N \in X$ and any $N \in \mathbb{N}$.

- define the inner product

$$\langle f, f' \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N'} k(x_i, x_j') a_i a_j'.$$

$\mathcal{H}_k$ is the completion to a Hilbert space.

[2] Any set of inputs not just data.

A key result

Theorem

Given a PD kernel $\mathbf{k}$ there exists Φ s.t. $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}_{\mathbf{k}}}$ and

$$\mathcal{H}_{\Phi} \simeq \mathcal{H}_{\mathbf{k}}$$

A key result

Theorem

Given a PD kernel k there exists Φ s.t. $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}_{\mathbf{k}}}$ and

$$\mathcal{H}_{\Phi} \simeq \mathcal{H}_{\mathbf{k}}$$

Proof: Not so easy...

A key result

Theorem

Given a PD kernel k there exists Φ s.t. $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}_k}$ and

$$\mathcal{H}_\Phi \simeq \mathcal{H}_k$$

Proof: Not so easy...

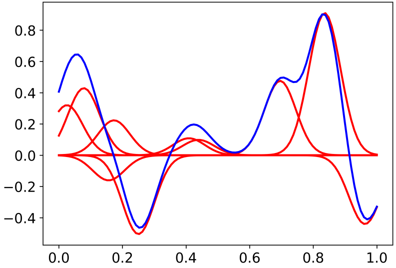
Roughly speaking

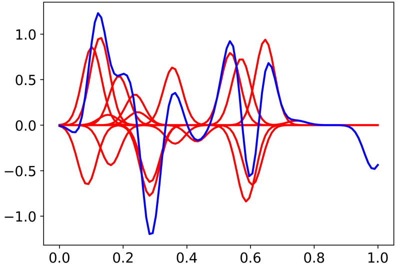
$$f(x) = w^\top \Phi(x) \qquad \simeq \qquad f(x) = \sum_{i=1}^{N} k(x, x_i) a_i$$

An illustration

Functions defind by Gaussian kernels with large and small widths.

$$f(x) = \sum_{i=1}^{N} k(x, x_i) a_i = \sum_{i=1}^{N} e^{-\frac{\|x - x_i\|^2}{2\sigma^2}} a_i$$





RKHS and representer

Functions in $\mathcal{H}_k$ are (limit of) combinations

$$f(x) = \sum_{i=1}^{N} k(x, x_i')c_i$$

for a sequence $x_1', \ldots, x_N'$.

The representer theorem ensures we can consider kernel combinations on data

$$f(x) = \sum_{i=1}^{n} k(x, x_i)c_i$$

From features and kernels to RKHS and beyond

$\mathcal{H}_k$ has many properties, characterizations, connections:

From features and kernels to RKHS and beyond

$\mathcal{H}_k$ has many properties, characterizations, connections:

- **Reproducing property**
- **Reproducing kernel Hilbert spaces (RKHS)**
- Mercer theorem (Karhunen Loéve expansion)
- Gaussian processes
- Stochastic calculus Cameron-Martin spaces
- Physics: POVM
- Harmonic analysis: group representation and wavelet transforms

Reproducing property

By definition of $\mathcal{H}_k$

- $k_x = k(x, \cdot) \in \mathcal{H}_k$
- Reproducing property
$$f(x) = \langle f, k_x \rangle$$
for all $f \in \mathcal{H}_k$, $x \in X$.

- Note that
$$|f(x) - f'(x)| \leqslant \|k_x\| \|f - f'\|, \qquad \forall x \in X.$$

The above observations have a converse.

RKHS

Definition

A RKHS $\mathcal{H}$ is a Hilbert space of functions where $\exists\, k : X \times X \to \mathbb{R}$ s.t.

- $k_x = k(x, \cdot) \in \mathcal{H}_k$,
- and

$$f(x) = \langle f, k_x \rangle \ .$$

Theorem

If $\mathcal{H}$ is a RKHS then k is pos. def.

Proof hint: Let $\Phi(x) = k_x$...

Equivalent RKHS definition by evaluation functionals

If $\mathcal{H}$ is a RKHS then the evaluation functionals

$$e_x(f) = f(x)$$

are continuous. i.e.

$$|e_x(f) - e_x(f')| \lesssim \|f - f'\|, \qquad \forall x \in X$$

since

$$e_x(f) = \langle f, k_x \rangle \; .$$

Note that $L^2(\mathbb{R}^d)$ or $C(\mathbb{R}^d)$ don't have this property[3]!

UniGe | MaLGa

[3] Why?

Alternative RKHS definition

Turns out the previous property also characterizes a RKHS.

Theorem

A Hilbert space of functions with continuous evaluation functionals is a RKHS.

Proof hint: direct application of Riesz lemma.

Feature maps & kernel/RKHS

- RKHS and PD kernels are in one to one relation.
- Feature map and kernel/RKHS are not.

Examples

- $\Phi(x) = k_x$ is a feature map (Aronzajin).
- If $(v_i)_i$ is a o.n.b. then $\Phi(x) = (v_1(x), v_2(x), \dots)$ is a feature map.
- ...

Mercer theorem

A famous feature map.

Let k bounded kernel and ρ probability distribution, and

$$L_k f(x) = \int d\rho(x) k(x, x') f(x') d\rho(x').$$

Theorem (Mercer theorem)

If $(\lambda_j, \psi_j)_j$ is the eigensystem [4] of L_k, then

$$k(x, x') = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(x')$$

where the series converges absolutely and pointwise.

- $\Phi(x) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \dots)$ is a feature map.

[4] $L_k \psi_j = \lambda_j \psi_j$

End of the tour

- Feature maps
- PD kernels
- RKHS

My view

"It's hard to find a useful function space which is not a RKHS".

L. Rosasco

The one exception are neural nets. But even there RKHS are useful for understanding (something...).