Reproducing kernel Hilbert spaces

Nuno Vasconcelos

ECE Department, UCSD
Classification

- a classification problem has **two types of variables**
  - \( X \) - vector of **observations** (features) in the world
  - \( Y \) - state (class) of the world

- **Perceptron**: classifier implements the **linear decision rule**
  \[
  h(x) = \text{sgn}[g(x)] \quad \text{with} \quad g(x) = w^T x + b
  \]

- appropriate when the **classes are linearly separable**

- to deal with **non-linear separability**, we introduce a kernel
Kernel summary

1. $D$ not linearly separable in $\mathcal{X}$, apply feature transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$, such that $\dim(\mathcal{Z}) >> \dim(\mathcal{X})$

2. computing $\Phi(x)$ too expensive:
   - write your learning algorithm in dot-product form
   - instead of $\Phi(x_i)$, we only need $\Phi(x_i)^T \Phi(x_j) \; \forall_{ij}$

3. instead of computing $\Phi(x_i)^T \Phi(x_j) \; \forall_{ij}$, define the “dot-product kernel”
   \[ K(x_i, z) = \Phi(x_i)^T \Phi(z) \]
   and compute $K(x_i, x_j) \; \forall_{ij}$ directly
   - note: the matrix
     \[
     K = \begin{bmatrix}
     \vdots \\
     \cdots K(x_i, x_j) \cdots \\
     \vdots 
     \end{bmatrix}
     \]
     is called the “kernel” or Gram matrix

4. forget about $\Phi(x)$ and use $K(x, z)$ from the start!
Polynomial kernels

- this makes a significant difference when $K(x,z)$ is easier to compute that $\Phi(x)^T \Phi(z)$

- e.g., we have seen that

\[
K(x, z) = \left(x^T z\right)^2 = \Phi(x)^T \Phi(z)
\]

with

\[
\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d^2}
\]

\[
\begin{pmatrix}
  x_1 \\
  \vdots \\
  x_d
\end{pmatrix} \rightarrow (x_1 x_1, x_1 x_2, \ldots, x_1 x_d, \ldots, x_d x_1, x_d x_2, \ldots, x_d x_d)^T
\]

- while $K(x, z)$ has complexity $O(d)$, $\Phi(x)^T \Phi(z)$ is $O(d^2)$

- for $K(x, z) = (x^T z)^k$ we go from $O(d)$ to $O(d^k)$
Question

in practice:

- pick a kernel from a library of known kernels
- we talked about
  - the linear kernel $K(x, z) = x^T z$
  - the Gaussian family
    $$K(x, z) = e^{\frac{-\|x-z\|^2}{\sigma}}$$
  - the polynomial family
    $$K(x, z) = (1 + x^T z)^k, \quad k \in \{1, 2, \ldots\}$$

what if this is not good enough?

- how do I know if a function $k(x, y)$ is a dot-product kernel?
Dot-product kernels

Let’s start by the definition

**Definition:** a mapping

\[ k: \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{H} \]

\[ (x,y) \rightarrow k(x,y) \]

is a dot-product kernel if and only if

\[ k(x,y) = \langle \Phi(x), \Phi(y) \rangle \]

where

\[ \Phi: \mathcal{X} \rightarrow \mathcal{H}, \]

\( \mathcal{H} \) is a vector space,

\( \langle ., . \rangle \) is a dot-product in \( \mathcal{H} \)

Note that both \( \mathcal{H} \) and \( \langle ., . \rangle \) can be abstract, not necessarily \( \mathbb{R}^d \).
Dot product vs positive definite kernels

that is pretty abstract. how do I turn it into something I can compute?

**Theorem:** $k(x, y)$, $x, y \in \mathcal{X}$, is a dot-product kernel if and only if it is a positive definite kernel.

this can be checked

let’s start by the definition

**Definition:** $k(x, y)$ is a positive definite kernel on $\mathcal{X} \times \mathcal{X}$ if $\forall l$ and $\forall \{x_1, \ldots, x_l\}$, $x_i \in \mathcal{X}$, the Gram matrix

$$[K]_{ij} = k(x_i, x_j)$$

is positive definite.
Positive definite matrices

- recall that (e.g. Linear Algebra and Applications, Strang)

**Definition:** each of the following is a necessary and sufficient condition for a real symmetric matrix $A$ to be positive definite:

1. $x^T A x \geq 0, \quad \forall x \neq 0$
2. all eigenvalues of $A$ satisfy $\lambda_i \geq 0$
3. all upper-left submatrices $A_k$ have non-negative determinant
4. there is a matrix $R$ with independent rows such that $A = R^T R$

**upper left submatrices:**

$$
\begin{align*}
A_1 &= a_{1,1} \\
A_2 &= \begin{bmatrix} a_{1,1} & a_{1,2} \\
a_{2,1} & a_{2,2} \end{bmatrix} \\
A_3 &= \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\
a_{2,1} & a_{2,2} & a_{2,3} \\
a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \\
& \quad \quad \quad \quad \quad \quad \vdots
\end{align*}
$$
Dot product vs positive definite kernels

equivalence between dot product and positive definite automatically proves some simple properties

• dot-product kernels are non-negative functions
  \[ k(x, x) \geq 0 \quad \forall x \in X \]

• dot-product kernels are symmetric
  \[ k(x, y) = k(y, x), \quad \forall x, y \in X \]

• Cauchy-Schwarz inequality for dot-product kernels
  \[ k(x, y)^2 \leq k(x, x)k(y, y), \quad \forall x, y \in X \]

• note that this is just
  \[ -1 \leq \frac{k(x, y)}{\sqrt{k(x, x)k(y, y)}} \leq 1 \quad \Leftrightarrow \quad -1 \leq \frac{\langle x, y \rangle}{\|x\|\|y\|} \leq 1 \]
Dot product vs positive definite kernels

The proof actually gives insight on what the kernel does

- We defined $\mathcal{H}$ as the space spanned by linear combinations of $k(.,x_i)$

$$
\mathcal{H} = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(.,x_i), \quad \forall m, \forall x_i \in X \right\}
$$

- E.g. the space of all linear combinations of Gaussians when we adopt the Gaussian kernel

$$
K(.,x_i) = e^{-\frac{\|.-x_i\|^2}{\sigma^2}}
$$

- Note: this is a function of the first argument, $x_i$ is fixed
Dot product vs positive definite kernels

• if \( f(.) \) and \( g(.) \) ∈ \( \mathcal{H} \), with

\[
f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i) \quad g(.) = \sum_{j=1}^{m'} \beta_j k(., x'_j)
\]

• we defined the dot-product \(<.,.>_{*}\) on \( \mathcal{H} \) as

\[
\langle f, g \rangle_{*} = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)
\]

• for the Gaussian kernel this is

\[
\langle f, g \rangle_{*} = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j e^{-\frac{\|x_i-x'_j\|^2}{\sigma^2}}
\]

• a dot product in \( \mathcal{H} \)

• a non-linear measure of similarity in \( X \), somewhat related to likelihoods
The dot-product $<.,.>^*$

- it is not hard to show that $\langle k(., x_i), k(., x_j) \rangle^* = k(x_i, x_j)$
- this means that

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle^*$$

with

$$\Phi: \mathcal{X} \rightarrow \mathcal{H}$$

with $x \rightarrow k(.,x)$

- the feature transformation associated with the kernel $k(x,y)$ sends the points $x_i$ to the functions $k(.,x_i)$
- furthermore, $<.,.>^*$ is itself a dot-product kernel on $\mathcal{H} \times \mathcal{H}$
The reproducing property

- with this definition of $\mathcal{H}$ and $\langle ., . \rangle_*$

$$\forall f \in \mathcal{H}, \quad \langle k(.,x), f(.) \rangle_* = f(x)$$

- this is called the reproducing property

- an analogy is to think of linear time-invariant systems
  - the dot product as a convolution
  - $k(.,x)$ as the Dirac delta
  - $f(.)$ as a system input
  - the equation above is the basis of all linear time invariant systems theory

- we will see that it also plays a fundamental role in the theory (and use) of reproducing Kernel Hilbert Spaces
The big picture

- when we use the Gaussian kernel $K(x, x_i) = e^{-\|x-x_i\|^2/\sigma}$
  - the point $x_i \in \mathcal{X}$ is mapped into the Gaussian $G(x, x_i, \sigma)$
  - $\mathcal{H}$ is the space of all functions that are linear combinations of Gaussians
  - the kernel is a dot product in $\mathcal{H}$, and a non-linear similarity on $\mathcal{X}$
  - reproducing property on $\mathcal{H}$: analogy to linear systems
Hilbert spaces

- play an important role in functional analysis
- we will do very informal review
- $\mathcal{H}$ is a vector space with a dot product:
  - this is known as a “dot-product space” or a “pre-Hilbert” space
- the difference to a Hilbert space is mostly technical
- **Definition:** a Hilbert space is a complete dot-product space
- what do we mean by completeness?
- **Definition:** $\mathcal{S}$ is complete if all Cauchy sequences in $\mathcal{S}$ converge
- this is useful mostly to prove convergence results
Cauchy sequences

just for completeness (no pun intended)

**Definition:** a sequence \( \{x_1, x_2, \ldots\} \) in a normed space is a Cauchy sequence if

\[
\forall \varepsilon \geq 0, \exists n \in \mathbb{N} \quad \text{s.t.} \quad \forall n', n'' > n, \|x_{n'} - x_{n''}\| \leq \varepsilon
\]

you can picture this as

why might this not converge? well, the limit point could be outside the space

why is this important? complete \( \Rightarrow \) convergent = Cauchy
reproducing kernel Hilbert spaces

to turn our pre-Hilbert space $\mathcal{H}$ into a Hilbert space we have to “complete it” with respect to the norm

$$\|f\|_* = \sqrt{\langle f, f \rangle_*}$$

this is just adding to $\mathcal{H}$ the limit points of all its Cauchy sequences

we represent the completion of $\mathcal{H}$ by $\overline{\mathcal{H}}$

more than an Hilbert space, $\overline{\mathcal{H}}$ becomes a reproducing kernel Hilbert space (RKHS)
reproducing kernel Hilbert spaces

**Definition:** Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$. $\mathcal{H}$ is a RKHS endowed with dot-product $<.,.>^*$ if there exists $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that

1. $k$ spans $\mathcal{H}$, i.e., $\exists \{x_i\}, \{\alpha_i\}$, such that

$$\mathcal{H} = \text{span}\{k(.,x_i)\} = \left\{ f(.) \mid f(.) = \sum \alpha_i k(., x) \right\}$$

2. $<f(.),k(.,x)>^* = f(x)$, $\forall f \in \mathcal{H}$,

**in summary,**

- the $\mathcal{H}$ we built can easily be transformed into a RKHS by adding to it the limit points (functions) of all Cauchy seqs

**Question:** is there a one-to-one mapping between kernels and RKHSs?
kernels vs RKHSs

- answer: RKHS does indeed specify kernel uniquely
- proof:
  - assume that kernels $k_1$ and $k_2$ span $\mathcal{H}$
  - using reproducing property
    - $k_1(x',x) = \langle k_1(.,x), k_2(.,x') \rangle^*$
    - $k_2(x,x') = \langle k_2(.,x'), k_1(.,x) \rangle^*$
  - hence, from symmetry of dot product $\langle .,. \rangle^*$ we must have
    $$k_1(x',x) = k_2(x,x')$$
  - and from symmetry of kernel it follows that
    $$k_1(x,x') = k_2(x,x')$$
  - since this holds for all $x$ and $x'$, the kernels are the same
kernels vs RKHSs

- however, it is not clear that a kernel uniquely specifies the RKHS
  - there might be multiple feature transforms and dot-products that are consistent with a kernel

- to study this we need to introduce Mercer kernels

**Definition:** a symmetric mapping \( k: \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) such that

\[
\int \int k(x, y)f(x)f(y)dx\,dy \geq 0,
\]

\[\forall f(x) \text{ s.t. } \int f(x)^2 \, dx < \infty \ (\star)\]

is a Mercer kernel
Mercer kernels

why do we care about them? Two reasons

1) **Theorem:** a kernel is positive definite (and dot-product) if and only if it is a Mercer kernel

proof that Mercer implies PD:

- consider any sequence \( \{x_1, ..., x_n\} \), \( x_i \in X \). Let
  \[
  f(x) = \sum_{i=1}^{n} w_i \delta(x - x_i), \quad s.t. \sum_{i=1}^{n} w_i^2 < \infty
  \]
- since \( \int f(x)^2 \, dx < \infty \), if \( k(x,y) \) is Mercer, it follows from (*) that
  \[
  0 \leq \int \int k(x, y) f(x) f(y) \, dx \, dy
  = \int \int k(x, y) \sum_{ij} w_i w_j \delta(x - x_i) \delta(x - x_j) \, dx \, dy
  \]
Mercer kernels

\[ 0 \leq \sum_{ij} w_i w_j \int \int k(x, y) \delta(x - x_i) \delta(x - x_j) \, dx \, dy \]

\[ = \sum_{ij} w_i w_j k(x_i, x_j) = w^T Kw \]

- since this holds for any \( w \), \( k(x, y) \) is PD
Mercer kernels

proof that PD implies Mercer:

- suppose there is a $g(x)$ s.t.
  \[ \iint k(x, y)g(x)g(y)dx\,dy = \varepsilon < 0 \]
- consider a partition of $X \times X$ fine enough that
  \[ \left| \sum_{ij} k(x_i, x_j)g(x_i)g(x_j)\Delta_x\Delta_y - \iint k(x, y)g(x)g(y)dx\,dy \right| \leq \frac{\varepsilon}{2} \]
- hence
  \[ \sum_{ij} k(x_i, x_j)g(x_i)g(x_j)\Delta_x\Delta_y < 0 \iff u^TKu < 0 \]
  with $u = (g(x_1), \ldots, g(x_n))^T$
- $K$ is not PD and $k(x,y)$ not a PD kernel
- the Theorem follows by contradiction
Mercer kernels

2) they have the following property (without proof)

**Theorem:** Let $k: \mathcal{X} \times \mathcal{X} \to R$ be a Mercer kernel. Then, there exists an orthonormal set of functions

$$\int \phi_i(x) \phi_j(x) dx = \delta_{ij}$$

and a set of $\lambda_i \geq 0$, such that

1) $\sum_{i=1}^{\infty} \lambda_i^2 = \int \int k^2(x, y) dx dy$

2) $k(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$ (**)

**intuition:** think of this as a 2D Fourier series (+Parseval)
Eigenfunctions

Note: if we define the operator

\[ (Tf)(x) = \int k(x, y) f(y) dy \]

then

\[ (T\phi_i)(x) = \int k(x, y)\phi_i(y) dy \]

\[ = \sum_k \lambda_k \int \phi_k(x)\phi_k(y)\phi_i(y) dy \]

\[ = \sum_k \lambda_k \phi_k(x) \underbrace{\int \phi_k(y)\phi_i(y) dy}_{\delta_{ij}} = \lambda_i \phi_i(x) \]

The functions \( \phi_i(x) \) are the eigenfunctions of \( (Tf)(x) \).

Again, a connection to linear systems (LTI if \( k(x,y) = h(x-y) \))
 Mercer kernels

the eigenfunction decomposition gives us another way to design the feature transformation

\[ \Phi : \quad X \rightarrow l_2^d \]
\[ x \rightarrow \left( \sqrt{\lambda_1} \phi_1 (x), \sqrt{\lambda_2} \phi_2 (x), \ldots \right)^T \]

where \( l_2^d \) is the space of vectors s.t. \( \sum_i a_i^2 < \infty \) and \( d \) the number of non-zero eigenvalues \( \lambda_i \)

clearly

\[ k(x, y) = \Phi(x)^T \Phi(y) \]

i.e. there is a vector space \( l_2^d \) other than \( \mathcal{H} \) s.t. \( k(x, y) \) is a dot product in that space
The picture

- this is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^d$

- much more like to a multi-layer Perceptron than before

- the kernelized Perceptron as a neural net
In summary

**Reproducing kernel map**

\[ H_K = \left\{ f(\cdot) \mid f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \right\} \]

\[ \langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) \]

\[ \Phi : x \rightarrow k(\cdot, x) \]

**Mercer kernel map**

\[ H_M = L_2^d = \left\{ x \mid \sum_{i=1}^{d} x_i^2 < \infty \right\} \]

\[ \langle f, g \rangle_* = f^T g \]

\[ \Phi : x \rightarrow \left( \sqrt{\lambda_1} \phi_1(x), \sqrt{\lambda_2} \phi_2(x), \cdots \right)^T \]

where \( \lambda_i, \phi_i \) are the eigenvalues and eigenfunctions of \( k(x, y) \)

- two very different pictures of what the kernel does
- are the two spaces really that different?
RK vs Mercer maps

- note that for $\mathcal{H}_M$ we are writing

$$\Phi(x) = \sqrt{\lambda_1}\phi_1(x)\vec{e}_1 + \cdots + \sqrt{\lambda_d}\phi_d(x)\vec{e}_d$$

but, since the $\phi_i(.)$ are orthonormal, there is a 1-1 map

$$\Gamma: \ 1^d_2 \rightarrow \text{span}\{\phi_k(.)\}$$

$$\vec{e}_k \rightarrow \sqrt{\lambda_k}\phi_k(.)$$

and we can write

$$(\Gamma \circ \Phi)(x) = \lambda_1\phi_1(x)\phi_1(.) + \cdots + \lambda_d\phi_d(x)\phi_d(.) \quad (***)$$

$$= k(.,x) \quad \text{from (**)$$

- hence $k(.,x)$ maps $x$ into $\mathcal{M} = \text{span}\{\phi_k(.)\}$
RK vs Mercer maps

- define the dot product in $\mathcal{M}$ so that
  \[
  \langle \phi_j, \phi_k \rangle = \frac{1}{\lambda_j} \int \phi_j(x)\phi_k(x)dx = \frac{1}{\lambda_j} \delta_{ij}
  \]

- then $\{\phi_k(.)\}$ is a basis, $\mathcal{M}$ is a vector space, any function in $\mathcal{M}$ can be written as
  \[
  f(x) = \sum_k \alpha_k \phi_k(x)
  \]
  and
  \[
  \langle f(\cdot), k(\cdot, x) \rangle = \left( \sum_i \alpha_i \phi_i(\cdot), \sum_j \lambda_j \phi_j(\cdot) \phi_j(x) \right) = \sum_{ij} \alpha_i \lambda_j \phi_j(x) \langle \phi_i, \phi_j \rangle = \sum_i \alpha_i \phi_i(x) = f(x)
  \]
  i.e., $k$ is a reproducing kernel on $\mathcal{M}$
RK vs Mercer maps

- furthermore, since $k(.,x) \in \mathcal{M}$, any functions of the form

  $$f(.) = \sum_i \alpha_i k(.,x_i) \quad g(.) = \sum_j \beta_j k(.,x_j)$$

- are in $\mathcal{M}$ and

  $$\langle f, g \rangle^{**} = \left\langle \sum_i \alpha_i k(.,x_i), \sum_j \beta_j k(.,x_j) \right\rangle^{**}$$

  $$= \sum_{ij} \alpha_i \beta_j \left\langle \sum_l \lambda_l \phi_l(.,x_i), \sum_m \lambda_m \phi_m(.,x_j) \right\rangle^{**}$$

  $$= \sum_{ij} \alpha_i \beta_j \sum_{lm} \lambda_l \lambda_m \phi_l(.,x_i) \phi_m(.,x_j) \langle \phi_l(.,.), \phi_m(.,.) \rangle^{**}$$

  $$= \sum_{ij} \alpha_i \beta_j \sum_l \lambda_l \phi_l(.,x_i) \phi_l(.,x_j) = \sum_{ij} \alpha_i \beta_j k(x_i, x_j)$$

- note that $f,g \in \mathcal{H}$ and this is the dot product we had in $\mathcal{H}$
In summary

- \( \mathcal{H} \subset \mathcal{M} \) and \(<.,.>_{\star} \) in \( \mathcal{H} \) is the same as \(<.,.>_{\star\star} \) in \( \mathcal{M} \)

**Question:** is \( \mathcal{M} \subset \mathcal{H} \)?

- need to show that any \( f(x) = \sum_k \alpha_k \phi_k(x) \in \mathcal{H} \)
- from (***)

\[
k(., x) = \lambda_1 \phi_1(x)\phi_1(.) + \cdots + \lambda_d \phi_d(x)\phi_d(.)
\]

and, for any sequence \( \{x_1, ..., x_d\} \),

\[
\begin{bmatrix}
k(., x_1) \\ \vdots \\ k(., x_d)
\end{bmatrix}
= \begin{bmatrix}
\lambda_1 \phi_1(x_1) & \cdots & \lambda_d \phi_d(x_1) \\
\vdots & \ddots & \vdots \\
\lambda_1 \phi_1(x_d) & \cdots & \lambda_d \phi_d(x_d)
\end{bmatrix}
\begin{bmatrix}
\phi_1(.) \\
\vdots \\
\phi_d(.)
\end{bmatrix}
\]

- if there is an invertible \( P \), then \( \phi_k(x) = \sum_k \alpha_k k(., x_k) \in \mathcal{H} \) and \( \mathcal{M} \subset \mathcal{H} \)
In summary

since $\lambda_i > 0$

$$P = \begin{bmatrix}
\phi_1(x_1) & \phi_d(x_1) \\
\vdots & \ddots \\
\phi_1(x_d) & \phi_d(x_d)
\end{bmatrix}
\begin{bmatrix}
\lambda_1 \\
\vdots \\
0
\end{bmatrix}$$

is invertible when $\Pi$ is. If $\Pi$ is not invertible, then

$$\forall i \exists \alpha \neq 0 \text{ s.t. } \sum_k \alpha_k \phi_k(x_i) = 0$$

if there is no sequence for which $\Pi$ is invertible then

$$\forall x \exists \alpha \neq 0 \text{ s.t. } \sum_k \alpha_k \phi_k(x) = 0$$

the $\phi_k(x)$ cannot be orthonormal. Hence there must be invertible $P$ and $\mathcal{M} \subset \mathcal{H}$. ■
In summary

- $\mathcal{H} = \mathcal{M}$ and $<.,.>_{\ast}$ in $\mathcal{H}$ is the same as $<.,.>_{\ast\ast}$ in $\mathcal{M}$
- the reproducing kernel map and the Mercer kernel map lead to the same RKHS

**Reproducing kernel map**

$$\mathcal{H}_K = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i) \right\}$$

$$\langle f, g \rangle_{\ast} = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

$$\Phi_r : x \rightarrow k(., x)$$

**Mercer kernel map**

$$\mathcal{H}_M = l_2^d = \left\{ x \mid \sum_{i=1}^{d} x_i^2 < \infty \right\}$$

$$\langle f, g \rangle_{\ast} = f^T g$$

$$\Phi_M : x \rightarrow \left( \sqrt{\lambda_1} \phi_1(x), \sqrt{\lambda_2} \phi_2(x), \ldots \right)^T$$

$$\Gamma \circ \Phi_M = \Phi_r$$