Reproducing Kernel Hilbert Spaces

Lorenzo Rosasco

9.520 Class 03

February 15, 2012
Goal  To introduce a particularly useful family of hypothesis spaces called Reproducing Kernel Hilbert Spaces (RKHS) We will discuss several perspectives on RKHS. In particular in this class we investigate the fundamental definition of RKHS as Hilbert spaces with bounded, continuous evaluation functionals and the intimate connection with symmetric positive definite kernels.
Plan

- Part I: RKHS are Hilbert spaces with bounded, continuous evaluation functionals.
- Part II: Reproducing Kernels
- Part III: Mercer Theorem
- Part IV: Feature Maps
- Part V: Representer Theorem
The basic idea of regularization (originally introduced independently of the learning problem) is to restore well-posedness of ERM by constraining the hypothesis space $\mathcal{H}$.

A possible way to do this is considering regularized empirical risk minimization, that is we look for solutions minimizing a two term functional

$$\underbrace{\text{empirical error}}_{\text{ERR}(f)} + \lambda \underbrace{\text{regularizer}}_{R(f)}$$

the regularization parameter $\lambda$ trade-offs the two terms.
Tikhonov regularization amounts to minimize

\[
\frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \mathcal{R}(f) \quad \lambda > 0
\]

- \(V(f(x), y)\) is the loss function, that is the price we pay when we predict \(f(x)\) in place of \(y\)
- \(\mathcal{R}(f)\) is a regularizer—often \(\mathcal{R}(f) = \| \cdot \|_{\mathcal{H}}\), the norm in the function space \(\mathcal{H}\)

The regularizer should encode some notion of smoothness of \(f\).
The scheme we just described is very general and by choosing different loss functions $V(f(x), y)$ we can recover different algorithms.

The main point we want to discuss is how to choose a norm encoding some notion of smoothness/complexity of the solution.

Reproducing Kernel Hilbert Spaces allow us to do this in a very powerful way.
Different Views on RKHS

- PD Kernels
- Integral Operators
- Feature Maps
- Bounded Continuous Evaluation Functionals
Part I: Evaluation Functionals
A function space $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \to \mathbb{R}$.

A norm is a nonnegative function $\| \cdot \|$ such that $\forall f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ iff $f = 0$;
2. $\|f + g\| \leq \|f\| + \|g\|$;
3. $\|\alpha f\| = |\alpha| \|f\|$.

A norm can be defined via an inner product $\|f\| = \sqrt{\langle f, f \rangle}$.

A Hilbert space is a complete inner product space.
A function space $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \rightarrow \mathbb{R}$.

A norm is a nonnegative function $\| \cdot \|$ such that $\forall f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ iff $f = 0$;
2. $\|f + g\| \leq \|f\| + \|g\|$;
3. $\|\alpha f\| = |\alpha| \|f\|$

A norm can be defined via an inner product $\|f\| = \sqrt{\langle f, f \rangle}$.

A Hilbert space is a complete inner product space.
A **function space** $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \to \mathbb{R}$.

A **norm** is a nonnegative function $\| \cdot \|$ such that $\forall f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ iff $f = 0$;
2. $\|f + g\| \leq \|f\| + \|g\|$;
3. $\|\alpha f\| = |\alpha| \|f\|$.

A norm can be defined via a **inner product** $\|f\| = \sqrt{\langle f, f \rangle}$.

A **Hilbert space** is a complete inner product space.
A **function space** $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \to \mathbb{R}$.

A **norm** is a nonnegative function $\| \cdot \|$ such that $\forall f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ iff $f = 0$;
2. $\|f + g\| \leq \|f\| + \|g\|$;
3. $\|\alpha f\| = |\alpha| \|f\|$.

A norm can be defined via a **inner product** $\|f\| = \sqrt{\langle f, f \rangle}$.

A **Hilbert space** is a complete inner product space.
Examples

- Continuous functions $C[a, b]$: a norm can be established by defining
  \[ \| f \| = \max_{a \leq x \leq b} |f(x)| \]
  (not a Hilbert space!)

- Square integrable functions $L^2[a, b]$: it is a Hilbert space where the norm is induced by the dot product
  \[ \langle f, g \rangle = \int_a^b f(x)g(x)dx \]
Examples

- Continuous functions $C[a, b]$: a norm can be established by defining

$$
\|f\| = \max_{a \leq x \leq b} |f(x)|
$$

(not a Hilbert space!)

- Square integrable functions $L^2[a, b]$: it is a Hilbert space where the norm is induced by the dot product

$$
\langle f, g \rangle = \int_{a}^{b} f(x)g(x)dx
$$
Hypothesis Space: Desiderata

- Hilbert Space.
- Point-wise defined functions.
Hypothesis Space: Desiderata

- Hilbert Space.
- Point-wise defined functions.
An evaluation functional over the *Hilbert space of functions* $\mathcal{H}$ is a linear functional $F_t : \mathcal{H} \to \mathbb{R}$ that *evaluates* each function in the space at the point $t$, or

$$F_t[f] = f(t).$$

**Definition**

A Hilbert space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if the evaluation functionals are bounded and continuous, i.e. if there exists a $M$ s.t.

$$|F_t[f]| = |f(t)| \leq M\|f\|_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$
An evaluation functional over the *Hilbert space of functions* $\mathcal{H}$ is a linear functional $F_t : \mathcal{H} \to \mathbb{R}$ that *evaluates* each function in the space at the point $t$, or

$$F_t[f] = f(t).$$

**Definition**

A Hilbert space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if the evaluation functionals are bounded and continuous, i.e. if there exists a $M$ s.t.

$$|F_t[f]| = |f(t)| \leq M\|f\|_{\mathcal{H}} \forall f \in \mathcal{H}$$
Evaluation functionals are not always bounded. Consider $L_2[a, b]$:

- Each element of the space is an equivalence class of functions with the same integral $\int |f(x)|^2 dx$.
- An integral remains the same if we change the function in a countable set of points.
Choosing different kernels one can show that the norm in the corresponding RKHS encodes different notions of smoothness.

- **Band limited functions.** Consider the set of functions

\[ \mathcal{H} := \{ f \in L^2(\mathbb{R}) \mid F(\omega) \in [-a, a], a < \infty \} \]

with the usual $L^2$ inner product. The function at every point is given by the convolution with a sinc function $\sin(ax)/ax$. The norm

\[
\|f\|_H^2 = \int f(x)^2 \, dx = \int_a^a |F(\omega)|^2 \, d\omega
\]

Where $F(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt$ is the Fourier transform of $f$. 
Choosing different kernels one can show that the norm in the corresponding RKHS encodes different notions of smoothness.

- Band limited functions. Consider the set of functions

\[ \mathcal{H} := \{ f \in L^2(\mathbb{R}) \mid F(\omega) \in [-a, a], a < \infty \} \]

with the usual \( L^2 \) inner product. the function at every point is given by the convolution with a sinc function \( \sin(ax)/ax \). The norm

\[ \| f \|_{\mathcal{H}}^2 = \int f(x)^2 \, dx = \int_a^a |F(\omega)|^2 \, d\omega \]

Where \( F(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} \, dt \) is the Fourier transform of \( f \).
Sobolev Space: consider \( f : [0, 1] \rightarrow \mathbb{R} \) with \( f(0) = f(1) = 0 \). The norm

\[
\|f\|_{\mathcal{H}}^2 = \int (f'(x))^2 \, dx = \int \omega^2 |F(\omega)|^2 \, d\omega
\]

Gaussian Space: the norm can be written as

\[
\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi^d} \int |F(\omega)|^2 \exp \frac{\sigma^2 \omega^2}{2} \, d\omega
\]
Sobolev Space: consider $f : [0, 1] \rightarrow \mathbb{R}$ with $f(0) = f(1) = 0$. The norm

$$\|f\|_{\mathcal{H}}^2 = \int (f'(x))^2 dx = \int \omega^2 |F(\omega)|^2 d\omega$$

Gaussian Space: the norm can be written as

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi^d} \int |F(\omega)|^2 \exp \frac{\sigma^2 \omega^2}{2} d\omega$$
Sobolev Space: consider $f: [0, 1] \rightarrow \mathbb{R}$ with $f(0) = f(1) = 0$. The norm

$$\|f\|_{\mathcal{H}}^2 = \int (f'(x))^2 \, dx = \int \omega^2 |F(\omega)|^2 \, d\omega$$

Gaussian Space: the norm can be written as

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi^d} \int |F(\omega)|^2 \exp \frac{\sigma^2 \omega^2}{2} \, d\omega$$
Our function space is 1-dimensional lines

\[ f(x) = w x \]

where the RKHS norm is simply

\[ \| f \|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = w^2 \]

so that our measure of complexity is the slope of the line. We want to separate two classes using lines and see how the magnitude of the slope corresponds to a measure of complexity. We will look at three examples and see that each example requires more "complicated functions, functions with greater slopes, to separate the positive examples from negative examples."
here are three datasets: a linear function should be used to separate the classes. Notice that as the class distinction becomes finer, a larger slope is required to separate the classes.
Part II: Kernels
Different Views on RKHS

PD Kernels

Bounded Continuous Evaluation Functionals

Integral Operators

Feature Maps
Let $\mathcal{H}$ be a Hilbert space and $g \in \mathcal{H}$, then

$$\Phi_g(f) = \langle f, g \rangle, \quad f \in \mathcal{H}$$

is a continuous linear functional.

**Riesz representation theorem**

The theorem states that every continuous linear functional $\Phi$ can be written uniquely in the form,

$$\Phi(f) = \langle f, g \rangle$$

for some appropriate element $g \in \mathcal{H}$. 
If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists, by the Riesz representation theorem, a function $K_t$ in $\mathcal{H}$ (called representer) with the reproducing property

$$\mathcal{F}_t[f] = \langle K_t, f \rangle_{\mathcal{H}} = f(t).$$

Since $K_t$ is a function in $\mathcal{H}$, by the reproducing property, for each $x \in X$

$$K_t(x) = \langle K_t, K_x \rangle_{\mathcal{H}}$$

The reproducing kernel (rk) of $\mathcal{H}$ is

$$K(t, x) := K_t(x)$$
Reproducing kernel (rk)

If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists, by the Riesz representation theorem a function $K_t$ in $\mathcal{H}$ (called representer) with the reproducing property

$$\mathcal{F}_t[f] = \langle K_t, f \rangle_{\mathcal{H}} = f(t).$$

Since $K_t$ is a function in $\mathcal{H}$, by the reproducing property, for each $x \in X$

$$K_t(x) = \langle K_t, K_x \rangle_{\mathcal{H}}$$

The reproducing kernel (rk) of $\mathcal{H}$ is

$$K(t, x) := K_t(x)$$
Let $X$ be some set, for example a subset of $\mathbb{R}^d$ or $\mathbb{R}^d$ itself. A *kernel* is a symmetric function $K : X \times X \to \mathbb{R}$.

**Definition**

A kernel $K(t, s)$ is **positive definite** (pd) if

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$. 

L. Rosasco

RKHS
The following theorem relates pd kernels and RKHS

**Theorem**

a) For every RKHS there exist an associated reproducing kernel which is symmetric and positive definite

b) Conversely every symmetric, positive definite kernel $K$ on $X \times X$ defines a unique RKHS on $X$ with $K$ as its reproducing kernel
a) We must prove that the rk \( K(t, x) = \langle K_t, K_x \rangle_H \) is symmetric and pd.

- Symmetry follows from the symmetry property of dot products

\[
\langle K_t, K_x \rangle_H = \langle K_x, K_t \rangle_H
\]

- \( K \) is pd because

\[
\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) = \sum_{i,j=1}^{n} c_i c_j \langle K_{t_i}, K_{t_j} \rangle_H = \| \sum_{i,j=1}^{n} c_j K_{t_j} \|_H^2 \geq 0.
\]
Conversely, given $K$ one can construct the RKHS $\mathcal{H}$ as the completion of the space of functions spanned by the set \( \{K_x | x \in X\} \) with an inner product defined as follows. The dot product of two functions $f$ and $g$ in $\text{span}\{K_x | x \in X\}$ is by definition

\[
\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{s} \sum_{j=1}^{s'} \alpha_i \beta_j K(x_i, x_j).
\]
Examples of pd kernels

Very common examples of symmetric pd kernels are

- **Linear kernel**
  \[ K(x, x') = x \cdot x' \]

- **Gaussian kernel**
  \[ K(x, x') = e^{-\frac{||x-x'||^2}{\sigma^2}}, \quad \sigma > 0 \]

- **Polynomial kernel**
  \[ K(x, x') = (x \cdot x' + 1)^d, \quad d \in \mathbb{N} \]

For specific applications, designing an effective kernel is a challenging problem.
Examples of pd kernels

- Kernel are a very general concept. We can have kernel on vectors, string, matrices, graphs, probabilities...
- Combinations of Kernels allow to do integrate different kinds of data.
- Often times Kernel are views and designed to be similarity measure (in this case it make sense to have normalized kernels)

\[ d(x, x')^2 = \| K_x - K'_x \|^2 = 2(1 - K(x, x')). \]