Reproducing Kernel Hilbert Spaces

Lorenzo Rosasco

9.520 Class 04

February 21, 2012

About this class

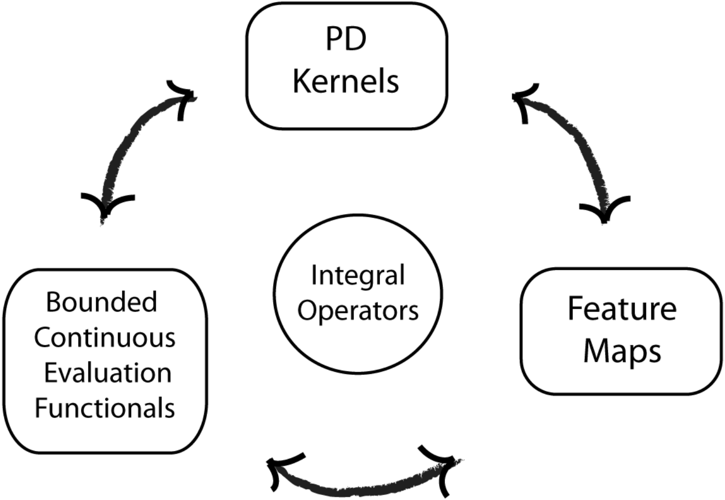
Goal In this class we continue our journey in the world of RKHS. We discuss the Mercer theorem which gives a new characterization of RKHS while introducing the concept of feature map. Then we discussed the concept of feature map and its interpretation. Finally, we show the computational implication of using RKHS by deriving the general solution of Tikhonov regularization, the so called he representer theorem.

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

Part III: Mercer Theorem

Different Views on RKHS



Integral Operator

RKH space can be characterized via the integral operator

$$L_K f(x) = \int_X K(x,s) f(s) p(s) dx$$

where $p(x)$ is the probability density on X.

The operator has domain and range in $L^2(X, p(x)dx)$ the space of functions $f : X \to \mathbb{R}$ such that

$$\langle f, f \rangle_{L^2} = \int_X |f(x)|^2 p(x) dx < \infty$$

Integral Operator

If X is a compact set and K is a **continuous** reproducing kernel (i.e. symmetric and PD) then L_K is a compact, positive and self-adjoint operator.

- There is a decreasing sequence $(\sigma_i)_i \geq 0$ such that $\lim_{i\to\infty} \sigma_i = 0$ and

$$L_K\phi_i(x) = \int_X K(x,s)\phi_i(s)p(s)ds = \sigma_i\phi_i(x),$$

where ϕ_i is an orthonormal basis in $L^2(X, p(x)dx)$.

- The action of L_K can be written as

$$L_K f = \sum_{i\geq 1} \sigma_i \langle f, \phi_i \rangle \phi_i.$$

Integral Operator

If X is a compact set and K is a **continuous** reproducing kernel (i.e. symmetric and PD) then L_K is a compact, positive and self-adjoint operator.

- There is a decreasing sequence $(\sigma_i)_i \geq 0$ such that $\lim_{i\to\infty} \sigma_i = 0$ and

$$L_K\phi_i(x) = \int_X K(x,s)\phi_i(s)p(s)ds = \sigma_i\phi_i(x),$$

  where ϕ_i is an orthonormal basis in $L^2(X, p(x)dx)$.
- The action of L_K can be written as

$$L_K f = \sum_{i\geq 1} \sigma_i \langle f, \phi_i \rangle \phi_i.$$

Integral Operator

If X is a compact set and K is a **continuous** reproducing kernel (i.e. symmetric and PD) then L_K is a compact, positive and self-adjoint operator.

- There is a decreasing sequence $(\sigma_i)_i \geq 0$ such that $\lim_{i \to \infty} \sigma_i = 0$ and

$$L_K \phi_i(x) = \int_X K(x, s) \phi_i(s) p(s) ds = \sigma_i \phi_i(x),$$

  where ϕ_i is an orthonormal basis in $L^2(X, p(x)dx)$.
- The action of L_K can be written as

$$L_K f = \sum_{i \geq 1} \sigma_i \langle f, \phi_i \rangle \phi_i.$$

Mercer Theorem

- The kernel function has the following representation

$$K(x, s) = \sum_{i \geq 1} \sigma_i \phi_i(x) \phi_i(s).$$

A symmetric, positive definite *and* continuous Kernel is called a *Mercer* kernel.

Different Definition of RKHS

It is possible to prove that:

- $$\mathcal{H} = \{f \in L^2(X, p(x)dx) | \sum_{i \geq 1} \frac{\langle f, \phi_i \rangle_{L^2}^2}{\sigma_i} < \infty\}.$$

- The scalar product in $\mathcal{H}$ is

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i \geq 1} \frac{\langle f, \phi_i \rangle_{L_2} \langle g, \phi_i \rangle_{L^2}}{\sigma_i}.$$

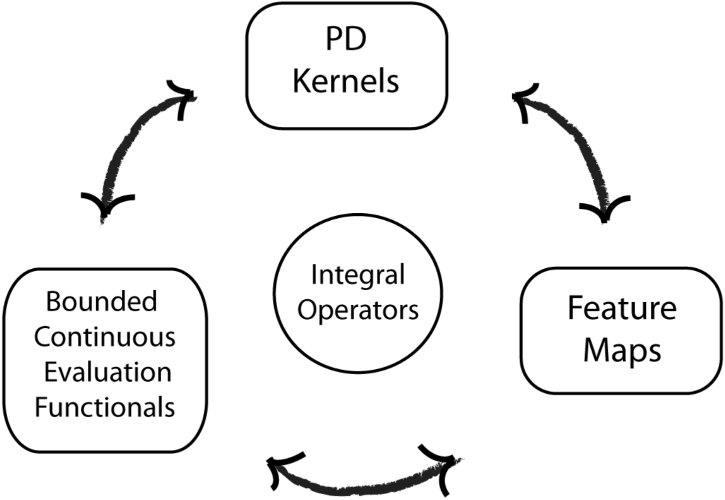
Different Definition of RKHS

It is possible to prove that:

- $$\mathcal{H} = \{f \in L^2(X, p(x)dx) | \sum_{i \geq 1} \frac{\langle f, \phi_i \rangle_{L^2}^2}{\sigma_i} < \infty\}.$$
- The scalar product in $\mathcal{H}$ is
$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i \geq 1} \frac{\langle f, \phi_i \rangle_{L_2} \langle g, \phi_i \rangle_{L^2}}{\sigma_i}.$$

Part IV: Feature Map

Different Views on RKHS



$$K(x,s) = \sum_{i \geq 1} \sigma_i \phi_i(x) \phi_i(s).$$

Let $\Phi(x) = (\sqrt{\sigma_i}\phi_i(x))_i$, then $\Phi : X \to \ell^2$ and (by definition)

$$K(x,s) = \langle \Phi(x), \Phi(x) \rangle .$$

The above is an example of **feature map** associated to K.

Feature Maps and Kernels

The above remark shows that we can associate a feature map to every kernel.
In fact, multiple feature maps can be associated to a kernel.

- Let $\Phi(x) = K_x$. Then $\Phi : X \to \mathcal{H}$.
- Let $\Phi(x) = (\psi_j(x))_j$, where $(\psi_j(x))_j$ is an orthonormal basis of $\mathcal{H}$. Then $\Phi : X \to \ell^2$.
  Why?

Feature Map and Feature Space

In general a feature map is a map $\Phi : X \to \mathcal{F}$, where $\mathcal{F}$ is a Hilbert space and is called Feature Space.
Every feature map defines a kernel via

$$K(x, s) = \langle \Phi(x), \Phi(x) \rangle .$$

Kernel from Feature Maps

Often times, feature map, and hence kernels, are defined through a dictionary of features

$$\mathcal{D} = \{\phi_j,\ i = 1, \dots, p \mid \phi_j : X \to \mathbb{R},\ \ \forall j\}$$

where $p \leq \infty$.

We can interpret the above functions as (possibly non linear) *measurements* on the inputs.

- If $p < \infty$ we can always define a feature map.
- If $p = \infty$ we need extra assumptions.
  Which ones?

Kernel from Feature Maps

Often times, feature map, and hence kernels, are defined through a dictionary of features

$$\mathcal{D} = \{\phi_j,\ i = 1, \dots, p \mid \phi_j : X \to \mathbb{R},\ \ \forall j\}$$

where $p \leq \infty$.
We can interpret the above functions as (possibly non linear) *measurements* on the inputs.

- If $p < \infty$ we can always define a feature map.
- If $p = \infty$ we need extra assumptions.
  Which ones?

Kernel from Feature Maps

Often times, feature map, and hence kernels, are defined through a dictionary of features

$$\mathcal{D} = \{\phi_j,\ i = 1, \dots, p \mid \phi_j : X \to \mathbb{R},\ \ \forall j\}$$

where $p \leq \infty$.
We can interpret the above functions as (possibly non linear) *measurements* on the inputs.

- If $p < \infty$ we can always define a feature map.
- If $p = \infty$ we need extra assumptions.
  Which ones?

Function as Hyperplanes in the Feature Space

The concept of feature map allows to give a new interpretation of RKHS.

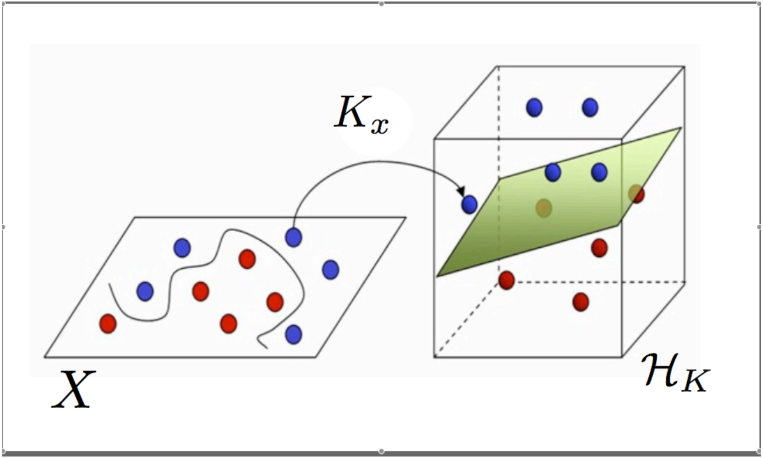
Functions can be seen as hyperplanes,

$$f(x) = \langle w, \Phi(x) \rangle .$$

This can be seen for any of the previous examples.

- Let $\Phi(x) = (\sqrt{\sigma_j}\phi_j(x))_j$.
- Let $\Phi(x) = K_x$.
- Let $\Phi(x) = (\psi_j(x))_j$.

Feature Maps Illustrated



Kernel "Trick" and Kernelization

Any algorithm which works in a euclidean space, hence requiring only inner products in the computations, can be *kernelized*

$$K(x,s) = \langle \Phi(x), \Phi(x) \rangle .$$

- Kernel PCA.
- Kernel ICA.
- Kernel CCA.
- Kernel LDA.
- Kernel...

Part V: Regularization Networks and Representer Theorem

Again Tikhonov Regularization

The algorithms (*Regularization Networks*) that we want to study are defined by an optimization problem over RKHS,

$$f_S^\lambda = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

where the *regularization parameter* λ is a positive number, $\mathcal{H}$ is the RKHS as defined by the *pd kernel* $K(\cdot,\cdot)$, and $V(\cdot,\cdot)$ is a **loss function.**

Note that $\mathcal{H}$ is possibly infinite dimensional!

Existence and uniqueness of minimum

If the positive loss function $V(\cdot, \cdot)$ is convex with respect to its first entry, the functional

$$\Phi[f] = \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

is *strictly convex* and *coercive*, hence it has exactly one local (global) minimum.
Both the squared loss and the hinge loss are convex.
On the contrary the 0-1 loss

$$V = \Theta(-f(x)y),$$

where $\Theta(\cdot)$ is the Heaviside step function, is **not** convex.

The Representer Theorem

An important result

The minimizer over the RKHS $\mathcal{H}$, f_S, of the regularized empirical functional

$$I_S[f] + \lambda \|f\|_{\mathcal{H}}^2,$$

can be represented by the expression

$$f_S^\lambda(x) = \sum_{i=1}^n c_i K(x_i, x),$$

for some n-tuple $(c_1, \ldots, c_n) \in \mathbb{R}^n$.
Hence, minimizing over the (possibly infinite dimensional) Hilbert space, *boils down to minimizing over* $\mathbb{R}^n$.

Sketch of proof

Define the linear subspace of $\mathcal{H}$,

$$\mathcal{H}_0 = \mathrm{span}(\{K_{x_i}\}_{i=1,\dots,n})$$

Let $\mathcal{H}_0^\perp$ be the linear subspace of $\mathcal{H}$,

$$\mathcal{H}_0^\perp = \{f \in \mathcal{H} | f(x_i) = 0,\ i = 1, \dots, n\}.$$

From the reproducing property of $\mathcal{H}$, $\forall f \in \mathcal{H}_0^\perp$

$$\langle f, \sum_i c_i K_{x_i} \rangle_{\mathcal{H}} = \sum_i c_i \langle f, K_{x_i} \rangle_{\mathcal{H}} = \sum_i c_i f(x_i) = 0.$$

$\mathcal{H}_0^\perp$ is the orthogonal complement of $\mathcal{H}_0$.

Sketch of proof (cont.)

Every $f \in \mathcal{H}$ can be uniquely decomposed in components along and perpendicular to $\mathcal{H}_0$: $f = f_0 + f_0^\perp$.
Since by orthogonality

$$\|f_0 + f_0^\perp\|^2 = \|f_0\|^2 + \|f_0^\perp\|^2,$$

and by the reproducing property

$$I_S[f_0 + f_0^\perp] = I_S[f_0],$$

then

$$I_S[f_0] + \lambda\|f_0\|_{\mathcal{H}}^2 \leq I_S[f_0 + f_0^\perp] + \lambda\|f_0 + f_0^\perp\|_{\mathcal{H}}^2.$$

Hence the minimum $f_S^\lambda = f_0$ *must belong to the linear space* $\mathcal{H}_0$.

Common Loss Functions

The following two important learning techniques are implemented by different choices for the loss function $V(\cdot, \cdot)$

• **Regularized least squares** (RLS)

$$V = (y - f(x))^2$$

• **Support vector machines for classification** (SVMC)

$$V = |1 - yf(x)|_+$$

where

$$(k)_+ \equiv \max(k, 0).$$

Tikhonov Regularization for RLS and SVMs

In the next two classes we will study Tikhonov regularization with different loss functions for both regression and classification. We will start with the square loss and then consider SVM loss functions.