Overview

Learning with Reproducing Kernel Hilbert Spaces

Yoonkyung Lee
Department of Statistics
The Ohio State University

January 22-26, 2007
Winter School at CIMAT

Outline

- Part I: Introduction to Kernel methods
- Part II: Learning with Reproducing Kernel Hilbert Spaces
- Part III: Structured learning for feature selection and prediction

Regularization in RKHS

Find \( f = \sum_{\nu=1}^{M} d_{\nu} \phi_{\nu} + h \) with \( h \in \mathcal{H}_K \) minimizing

\[
\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \|h\|_{\mathcal{H}_K}^2.
\]

- \( \mathcal{H}_K \): a reproducing Kernel Hilbert space of functions defined on an arbitrary domain
- \( J(f) = \|h\|_{\mathcal{H}_K}^2 \): penalty
- The null space spanned by \( \{\phi_{\nu}\}_{\nu=1}^{M} \)
Why consider RKHS?

- Theoretical basis for regularization methods
- Unified framework for function estimation and modeling various data
- Allow general domains for functions
- Can do much more than estimation of function values (e.g. integrals and derivatives)
- Geometric understanding

Reproducing kernel Hilbert spaces

- Consider a Hilbert space $\mathcal{H}$ of real valued functions on a domain $\mathcal{X}$.
- A Hilbert space $\mathcal{H}$ is a complete inner product linear space.
- For example, the domain $\mathcal{X}$ could be
  - $\{1, \ldots, k\}$
  - $[0, 1]$
  - $\{A, C, G, T\}$
  - $\mathbb{R}^d$
  - $S$: sphere.
- A reproducing kernel Hilbert space is a Hilbert space of real valued functions, where the evaluation functional $L_x(f) = f(x)$ is bounded in $\mathcal{H}$ for each $x \in \mathcal{X}$.

Riesz Representation Theorem

- For every bounded linear functional $L$ in a Hilbert space $\mathcal{H}$, there exists a unique $g_L \in \mathcal{H}$ such that $L(f) = (g_L, f)$, $\forall f \in \mathcal{H}$.
- $g_L$ is called the representer of $L$.

Reproducing kernel

Aronszajn (1950), *Theory of Reproducing kernels*.

- By the Riesz representation theorem, there exists $K_x \in \mathcal{H}$, the representer of $L_x(\cdot)$, such that $(K_x, f) = f(x), \forall f \in \mathcal{H}$.
- $K(x, t) = K_x(t)$ is called the reproducing kernel.
  - $K(x, \cdot) \in \mathcal{H}$ for each $x$
  - $(K(x, \cdot), f(\cdot)) = f(x)$ for all $f \in \mathcal{H}$
- Note that $K(x, t) = K_x(t) = (K_t(\cdot), K_x(\cdot)) = (K_x(\cdot), K_t(\cdot))$ (the reproducing property)
Example of RKHS

- $\mathcal{F} = \{ f \mid f : \{1, \ldots, k\} \to \mathbb{R} \} = \mathbb{R}^k$ with the Euclidean inner product $(f, g) = \sum_{j=1}^{k} f_j g_j$
- Note that $|f(j)| \leq \| f \|$. That is, the evaluation functional $L_j(f) = f(j)$ is bounded in $\mathcal{F}$ for each $j \in \mathcal{X}$.
- $L_j(f) = f(j) = (e_j, f)$ where $e_j$ is the $j$th column of $I_k$.
  Hence $K(i, j) = \delta_{ij} = 1[i = j]$ or $[K(i, j)] = I_k$.

Reproducing kernel is non-negative definite

- A bivariate function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is non-negative definite if for every $n$, and every $x_1, \ldots, x_n \in \mathcal{X}$, and every $a_1, \ldots, a_n \in \mathbb{R}^n$,
  $$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0.$$ 
  In other words, letting $a = (a_1, \ldots, a_n)^t$,
  $$a^t \left[K(x_i, x_j)\right] a \geq 0.$$
- For a reproducing kernel $K$,
  $$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) = \left\| \sum_{i=1}^{n} a_i K(x_i, \cdot) \right\|^2 \geq 0.$$

The Mercer-Hilbert-Schmidt Theorem

- If $\int_{\mathcal{X}} \int_{\mathcal{X}} K^2(s, t) dsdt < \infty$ for a continuous symmetric non-negative definite $K$, then there exists an orthonormal sequence of continuous eigenfunctions $\Phi_1, \Phi_2, \ldots$ in $L_2[\mathcal{X}]$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ with $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$ such that
  $$K(s, t) = \sum_{i=1}^{\infty} \lambda_i \Phi_i(s) \Phi_i(t)$$
- The inner product in $\mathcal{H}$ of functions $f$ with $\sum_{i=1}^{\infty} (f_i^2 / \lambda_i) < \infty$
  $$(f, g) = \sum_{i=1}^{\infty} \frac{f_i}{\lambda_i} g_i$$
  where $f_i = \int_{\mathcal{X}} f(t) \Phi_i(t) dt$.
- Feature mapping: $\Phi(x) = (\sqrt{\lambda_1} \Phi_1(x), \sqrt{\lambda_2} \Phi_2(x), \ldots)$

The Moore-Aronszajn Theorem

- For every RKHS $\mathcal{H}$ of functions on $\mathcal{X}$, there corresponds a unique reproducing kernel (RK) $K(s, t)$, which is n.n.d.
- Conversely, for every n.n.d. function $K(s, t)$ on $\mathcal{X}$, there corresponds a unique RKHS $\mathcal{H}$ that has $K(s, t)$ as its RK.
How to construct RKHS given an n.n.d. $K(s, t)$

- For each fixed $x \in \mathcal{X}$, define $K_x(\cdot) = K(x, \cdot)$.
- Taking all finite linear combinations of the form $\sum_i a_i K_{x_i}$ for all choices of $n$, $a_1, \ldots, a_n$, and $x_1, \ldots, x_n$, construct a linear space $\mathcal{M}$.
- Define an inner product via $(K_{x_i}, K_{x_j}) = K(x_i, x_j)$ and extend it by linearity.

$$\sum_i a_i K_{x_i}, \sum_j b_j K_{t_j} = \sum_{i,j} a_i b_j K(x_i, t_j).$$

- For any $f$ of the form $\sum_i a_i K_{x_i}$, $(K_{x_i}, f) = f(x)$.
- Complete $\mathcal{M}$ by adjoining all the limits of Cauchy sequences of functions in $\mathcal{M}$.

Sum of reproducing kernels

- The sum of two n.n.d. matrices is n.n.d.
- Hence, the sum of two n.n.d. functions defined on the same domain $\mathcal{X}$ is n.n.d.
- In particular, if $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, and $K_j(s, t)$ is the RK of $\mathcal{H}_j$ for $j = 1, 2$, then $K(s, t) = K_1(s, t) + K_2(s, t)$ is the RK for the tensor sum space of $\mathcal{H}_1 \oplus \mathcal{H}_2$.

Product of reproducing kernels

- Suppose that $K_1(x_1, x_2)$ is n.n.d. on $\mathcal{X}_1$ and $K_2(t_1, t_2)$ is n.n.d. on $\mathcal{X}_2$.
- Consider the tensor product of $K_1$ and $K_2$

$$K((x_1, t_1), (x_2, t_2)) = K_1(x_1, x_2) K_2(t_1, t_2)$$

on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$.
- The tensor product of two RK's $K_1$ and $K_2$ for $\mathcal{H}_1$ and $\mathcal{H}_2$ is an RK for the tensor product space of $\mathcal{H}_1 \otimes \mathcal{H}_2$ on $\mathcal{X}$.

Constructing kernels

- Use reproducing kernels on a univariate domain as building blocks.
- Tensor sums and products of reproducing kernels.
- Systematic approach to estimating multivariate functions.
- Other tricks to expand and design kernels: Haussler (1999), Convolution Kernels on Discrete Structures
- Learning kernels (n.n.d. matrices) from data: Lanckriet et al. (2004)
Cubic smoothing splines

Find \( f(x) \in W_2[0,1] \)
\( = \{ f : f, f' \) absolutely continuous, and \( f'' \in L_2 \} \) minimizing
\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx.
\]

- The null space: \( M = 2, \) \( \phi_1(x) = 1, \) and \( \phi_2(x) = x. \)
- The penalized space: \( H_K = W_2^0[0,1] \)
\( = \{ f \in W_2[0,1] : f(0) = 0, f'(0) = 0 \} \) is an RKHS with
\[ i) (f, g) = \int_0^1 f''(x)g''(x) dx \\
ii) \| f \|^2 = \int_0^1 (f''(x))^2 dx \\
iii) K(x, x') = \int_0^1 (x - u)_+ (x' - u)_+ du. \]

Representer Theorem

Kimeldorf and Wahba (1971)
- The minimizer \( f = \sum_{\nu=1}^{M} d_{\nu} \phi_{\nu} + h \) with \( h \in H_K \) of
\[
\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \| h \|^2_{H_K}
\]
has a representation of the form
\[
\hat{f}(x) = \sum_{\nu=1}^{M} d_{\nu} \phi_{\nu}(x) + \sum_{i=1}^{n} c_i K(x_i, x).
\]

- \( \| h \|^2_{H_K} = \sum_{i,j} c_i c_j K(x_i, x_j). \)

SVM in general

Find \( f(x) = b + h(x) \) with \( h \in H_K \) minimizing
\[
\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i)) + \lambda \| h \|^2_{H_K}.
\]

- The null space: \( M = 1 \) and \( \phi_1(x) = 1 \)
- Linear SVM:
\( H_K = \{ h(x) = w^T x \mid w \in \mathbb{R}^p \} \) with
\[ i) K(x, x') = x^T x' \\
ii) \| h \|^2_{H_K} = \| w \|^2 \]
- Nonlinear SVM: \( K(x, x') = (1 + x^T x')^d, \) \( \exp(-\|x - x'\|^2/2\sigma^2), \ldots \)

Sketch of the proof

- Write \( h \in H_K \) as
\[
h(x) = \sum_{i=1}^{n} c_i K(x_i, x) + \rho(x)
\]
where \( \rho(x) \) is some element in \( H_K \) perpendicular to \( K(x_i, x) \) for \( i = 1, \ldots, n. \)
- Note that \( h(x_i) = (K(x_i, \cdot), h(\cdot))_{H_K} \) does not depend on \( \rho \) and \( \| h \|^2_{H_K} = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) + \| \rho \|^2_{H_K} \)
- Then, \( \| \rho \|^2_{H_K} \) needs to be zero.
- Hence, the minimizer \( \hat{f} \) is of the form
\[
\hat{f}(x) = \sum_{\nu=1}^{M} d_{\nu} \phi_{\nu}(x) + \sum_{i=1}^{n} c_i K(x_i, x).
\]
Remarks on the Representer Theorem

- It holds for an arbitrary loss function $\mathcal{L}$.
- Minimizer of RKHS method resides in a finite dimensional space.
- So the solution is computable even if the RKHS had infinite dimension.
- The resulting optimization problems with the representation $\hat{f}$ depend on $\mathcal{L}$ and the penalty $J(f)$.

Kernelize

- Kernel trick:
  Replace $(x, x')$ with $K(x, x')$ in your linear method!
- This idea goes beyond supervised learning problems.
  - nonlinear dimension reduction and data visualization: kernel PCA
  - clustering: kernel k-means algorithm
  - ...

Summary

- Regularization in reproducing kernel Hilbert spaces is discussed.
- RKHS method provides a unified framework for statistical model building.
- Kernels are now used as a versatile tool for flexible modeling and learning in various contexts.
- Feature selection for kernel methods will be discussed in Part III based on the idea of kernel construction by sums and products.