REPRODUCING KERNEL HILBERT SPACES

RKHSs are spaces of functions that have an interesting mathematical theory, and that have been applied is statistics and more recently machine learning.

We will see that some familiar supervised kernel methods, such as KRR and SVMs, can be cast in terms of RKHSs.

New algorithms can also be derived from this general framework.

[The "Mathematical Background" section can be skipped]

Mathematical Background

Hilbert Spaces

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.

This inner product induces a norm

$$\|v\|_{\mathcal{H}} = \sqrt{\langle v, v \rangle_{\mathcal{H}}}$$

and distance

$$d_{\mathcal{H}}(u, v) = \|u - v\|_{\mathcal{H}}.$$

With this distance we can define convergent and Cauchy sequences.

Defs Let $v_1, v_2, \ldots \in \mathcal{H}$.

1) We say the sequence converges if $\exists\, v \in \mathcal{H}$ s.t.

$$\lim_{n \to \infty} d(v_n, v) = 0$$

2) We say the sequence is Cauchy if $\forall \epsilon > 0$,

$\exists\, N$ s.t. $m, n \geq N \Rightarrow d(v_m, v_n) < \epsilon$.

3) $\mathcal{H}$ is a Hilbert space if every Cauchy sequence in $\mathcal{H}$ converges. (more generally, a metric space is complete iff every Cauchy sequence converges)

Loosely speaking: every sequence that looks like it converges actually converges (to a point in $\mathcal{H}$)

Examples

- $\mathcal{H} = \mathbb{R}^d$, $\langle u, v \rangle = \sum_{i=1}^{d} u^{(i)} v^{(i)}$

  is complete

- $\mathbb{Q}^d = \{\text{rational numbers}\}^d$

  is not complete

- $L^2(\mathbb{R}^d) = \{ f : \mathbb{R}^d \to \mathbb{R} \mid \int f(x)^2 dx < \infty \}$

  $\langle f, g \rangle = \int f(x) g(x) dx$

  is complete

- $C(\mathbb{R}^d) \subseteq L^2(\mathbb{R}^d)$ is not complete

  ↖ continuous functions

Linear Functions

A function $L : \mathcal{H} \to \mathbb{R}$ is linear iff

$$L(au + bv) = aL(u) + bL(v)$$

$\forall a, b \in \mathbb{R}$, $u, v \in \mathcal{H}$.

Theorem (Riesz Representation Theorem)

Suppose $\mathcal{H}$ is a Hilbert space and $L: \mathcal{H} \to \mathbb{R}$ linear. Then L is continuous $\iff \exists\, u \in \mathcal{H}$ such that

$$L(v) = \langle u, v \rangle \qquad \forall v \in \mathcal{H}.$$

RKHSs

Now suppose $\mathcal{H}$ is a Hilbert space whose elements are functions $f: \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^d$.

Def $\mathcal{H}$ is a RKHS $\iff \forall x \in \Omega$, the linear function

$$L_x(f) := f(x)$$

is continuous.

By the Riesz Representation Theorem, $\mathcal{H}$ is a RKHS $\iff \forall x \in \Omega, \exists\, \Phi_x \in \mathcal{H}$ s.t.

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

$\forall f \in \mathcal{H}$.

The function $k(x', x) := \Phi_x(x')$ is called the reproducing kernel of $\mathcal{H}$.

It can easily be seen that k is a PSD kernel.

- symmetry:

$$\begin{aligned} k(x', x) &= \Phi_x(x') && \text{[def of } k\text{]} \\ &= \langle \Phi_{x'}, \Phi_x \rangle && \text{[Riesz rep. of } \Phi_x\text{]} \\ &= \langle \Phi_x, \Phi_{x'} \rangle && \text{[symmetry of } \langle \cdot, \cdot \rangle\text{]} \\ &= \Phi_{x'}(x) && \text{[Riesz rep. of } \Phi_{x'}\text{]} \\ &= k(x, x') && \text{[def of } k\text{]} \end{aligned}$$

- PSD: For any $x_1, \ldots, x_n \in \Omega$,

$$K = [k(x_i, x_j)] = [\langle \Phi_{x_i}, \Phi_{x_j} \rangle]$$

is a Gram matrix and therefore a PSD matrix.

Conversely, for every PSD kernel k, there is an associated RKHS.

- Let $k(\cdot, x)$ denote the function $x' \mapsto k(x', x)$.
- Let $\mathcal{H}_0 = \{ f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \mid x_i \in \Omega, \alpha_i \in \mathbb{R}, n \in \mathbb{N} \}$
- For $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x)$, $f' = \sum_{i=1}^{n'} \alpha_i' k(\cdot, x_i')$, set

$$\langle f, f' \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \alpha_j' k(x_i, x_j').$$

It can be shown that this is an inner product on $\mathcal{H}_0$.

- Now let $\mathcal{H}$ be the <u>completion</u> of $\mathcal{H}_0$.

Then $\mathcal{H}$ is a RKHS, and <u>its reproducing kernel is k</u>

<u>Note</u> | The above construction shows that for any PSD kernel, $\exists$ a Hilbert space $\mathcal{H}$ and a feature map $x \mapsto \Phi_x$ s.t. $k(x, x') = \langle \Phi_x, \Phi_{x'} \rangle$.

In particular, $\Phi_x = k(\cdot, x)$.

<u>Ref</u> | Steinwart and Christmann, Support Vector Machines, Springer, 2008.

RKHS Overview

A RKHS is an inner product space of real-valued functions on $\mathbb{R}^d$, defined in terms of a PSD kernel as follows:

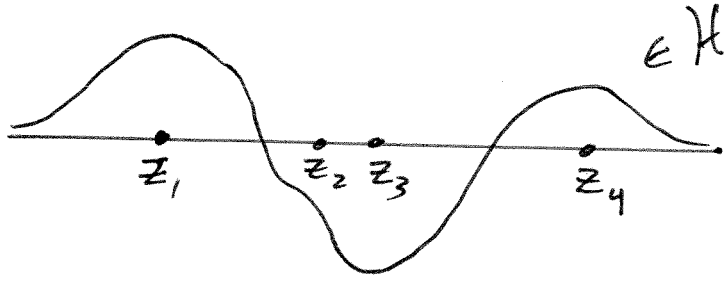
- $\mathcal{H}_0 = \left\{ \sum_{i=1}^{m} \alpha_i \, k(\cdot, z_i) \;\middle|\; \alpha_i \in \mathbb{R},\ z_i \in \mathbb{R}^d,\ m \in \mathbb{N} \right\}$

- $\left\langle \sum_{i=1}^{m} \alpha_i \, k(\cdot, z_i),\ \sum_{j=1}^{n} \beta_j \, k(\cdot, z_j') \right\rangle := \sum_{i,j} \alpha_i \beta_j \, k(z_i, z_j')$

  $= \alpha^T K \beta, \quad K = [k(z_i, z_j')]$

- $\mathcal{H}$ = $\mathcal{H}_0$ plus all its limit points (with finite norm)

Example | Gaussian kernel



Remarks |

- For every $f \in \mathcal{H}$ and $x \in \mathbb{R}^d$, $f(x) = \langle f, k(\cdot, x) \rangle$. Proof is obvious for $f \in \mathcal{H}_0$. Known as the reproducing property.

- The above construction is how one shows that every PSD kernel is an IP kernel. $\mathcal{H}$ is the feature space and $\Phi(x) := k(\cdot, x)$ is the feature mapping (note $\langle \Phi(x), \Phi(x') \rangle = k(x, x')$ by definition of $\langle \cdot, \cdot \rangle$).

The Representer Theorem

RKHSs are nice because certain optimization problems over infinite-dimensional RKHSs actually reduce to finite dimensional optimization problems.

Theorem Let $\mathcal{H}$ be a RKHS of functions f defined on a domain $\Omega \subseteq \mathbb{R}^d$. Consider an optimization problem of the form

$$\min_{f \in \mathcal{H}} \quad J(f) \qquad (*)$$

where J can be expressed

$$J(f) = L(f(x_1), \ldots, f(x_n)) + \Lambda(\|f\|_{\mathcal{H}}^2)$$

(L: arbitrary; Λ: nondecreasing)

for some $x_1, \ldots, x_n \in \Omega$.
Then there exists a minimizer of $(*)$ of the form

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

where k is the reproducing kernel of $\mathcal{H}$ and $\alpha_i \in \mathbb{R}$.

Proof Consider the subspace $S \subseteq \mathcal{H}$ given by

$$S = \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \;\middle|\; \alpha_i \in \mathbb{R},\ i = 1, \ldots, n \right\}$$

This is a finite dimensional subspace, and therefore

$$\mathcal{H} = S \oplus S^{\perp}.$$

For any $f \in \mathcal{H}$, let us write $f = f_{\parallel} + f_{\perp}$ where $f_{\parallel} \in S$, $f_{\perp} \in S^{\perp}$. Since

$$\begin{aligned} f(x_i) &= \langle f, k(x_i, \cdot) \rangle \\ &= \langle f_{\parallel}, k(x_i, \cdot) \rangle + \langle f_{\perp}, k(x_i, \cdot) \rangle \\ &= \langle f_{\parallel}, k(x_i, \cdot) \rangle \\ &= f_{\parallel}(x_i), \end{aligned}$$

the value of $L(f(x_1), \ldots, f(x_n))$ depends only on $f_{\parallel}$. Let us write $L(f)$ for $L(f(x_1), \ldots, f(x_n))$. Now

$$\begin{aligned} J(f) &= L(f) + \Lambda(\|f\|_{\mathcal{H}}^2) \\ &= L(f_{\parallel}) + \Lambda(\|f\|_{\mathcal{H}}^2) \\ &\geq L(f_{\parallel}) + \Lambda(\|f_{\parallel}\|_{\mathcal{H}}^2) \\ &= J(f_{\parallel}) \end{aligned}$$

since Λ is nondecreasing and $\|f\|^2 = \|f_{\parallel}\|^2 + \|f_{\perp}\|^2$.

Therefore, if f is a minimizer of J, then so is $f_{\|}$. Since $f_{\|} \in S$, it has the desired form ∎

Note! From the proof we see that if Λ is strictly increasing, then $J(f) > J(f_{\|})$ unless $f_{\perp} = 0$. Therefore all minimizers of J have the stated form.

Kernel Ridge Regression

Let's apply the preceding with

$$L(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\Lambda(\|f\|^2) = \lambda \|f\|^2$$

By the Representer Theorem, the solution of

$$\min_{f \in \mathcal{H}} \sum (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2$$

has the form

$$f = \sum \alpha_i k(\cdot, x_i).$$

Therefore it suffices to solve

$$\min_{\alpha} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{n} \alpha_i k(x_i, x_j)\Big)^2 + \lambda \Big\| \sum_j \alpha_j k(\cdot, x_j) \Big\|_K^2$$

Now

$$\Big\| \sum_{j=1}^{n} \alpha_j k(\cdot, x_j) \Big\|^2 = \sum_{ij} \alpha_i \alpha_j k(x_i, x_j)$$

$$= \alpha^T K \alpha$$

and

$$\sum_{i=1}^{n} \Big(y_i - \sum_j \alpha_j k(x_i, x_j)\Big)^2 = \| y - K\alpha \|^2$$

$$= y^T y - 2 y^T K \alpha + \alpha^T K^2 \alpha,$$

so we need to minimize

$$\alpha^T (K^2 + \lambda K) \alpha - 2 y^T K \alpha$$

$$\Rightarrow \quad 2(K^2 + \lambda K)\alpha - 2Ky = 0$$

$$\Rightarrow \quad \alpha = (K + \lambda I)^{-1} y \qquad \text{[assuming } K \text{ invertible]}$$

This is identical to kernel ridge regression. Easy!

Support Vector Machines

Now consider

$$L(f(x_1), \ldots, f(x_n)) = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i f(x_i))$$

$$\Lambda(\|f\|^2) = \frac{\lambda}{2}\|f\|^2$$

Once again, by the representer theorem, the minimizer has the form

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

$r_i \in \mathbb{R}$. This reduces the optimization problem to

$$\min_{r} \quad \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \cdot \sum_{j=1}^{n} r_j \, k(x_i, x_j)\right) + \frac{1}{2} r^T K r$$

$$\left[c = \frac{1}{\lambda}\right]$$

$$\iff \quad \min_{r, \xi} \quad \frac{1}{2} r^T K r + \frac{c}{n} \sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_i \cdot \sum_{j=1}^{n} r_j \, k(x_i, x_j) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

The Lagrangian is

$$L(r, \xi, \alpha, \beta) = \frac{1}{2} r^T K r + \frac{C}{n} \sum \xi_i - \sum_i \alpha_i \left[y_i \left(\sum_j r_j k(x_i, x_j) \right) - 1 + \xi_i \right] - \sum \beta_i \xi_i$$

We can write

$$\sum_i \alpha_i y_i \left(\sum_j r_j k(x_i, x_j) \right) = (\alpha \odot y)^T K r$$

$\odot$ ← element-wise product

By the KKT conditions,

$$0 = \frac{\partial L}{\partial r} = Kr - K(\alpha \odot y)$$

$$\iff r = \alpha \odot y \qquad \text{[if } K \text{ invertible]}$$

$$\iff r_i = \alpha_i y_i \quad \forall i$$

and

$$0 = \frac{\partial L}{\partial \xi_i} = \frac{C}{n} - \alpha_i - \beta_i$$

Converting to the dual we have

$$\max_{\alpha} \; -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_i \alpha_i$$

$$\text{s.t.} \quad 0 \le \alpha_i \le \frac{C}{n}$$

(without offset)

This is the SVM dual. Furthermore, since $r_i = \alpha_i y_i$, we have

$$f^*(x) = \sum_{i=1}^{n} \alpha_i y_i k(x, x_i).$$

Thus, we recover the SVM decision function.

Other Instances

- Kernel logistic regression

$$L(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i f(x_i))\right)$$

$$\Lambda(\|f\|) = \frac{\lambda}{2}\|f\|^2$$

Exercise Derive a Newton-type algorithm to find the optimal α_i.

- Cubic smoothing spline

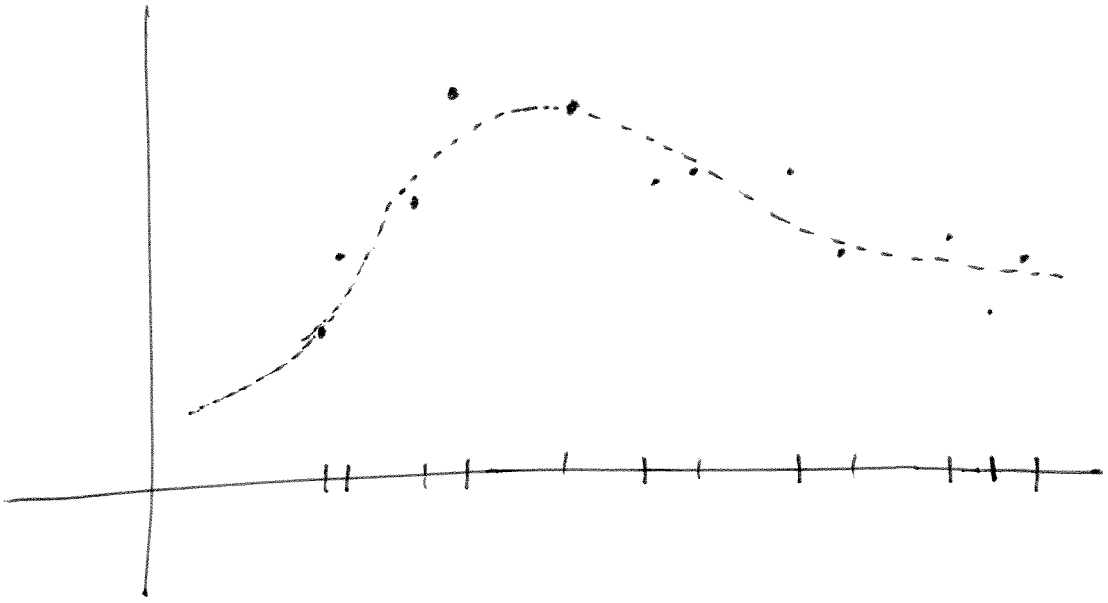
$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b [f''(x)]^2 dx$$

where $\mathcal{H} = \left\{ f : [a, b] \longrightarrow \mathbb{R} \mid \int_a^b [f''(x)]^2 dx < \infty \right\}$.

It can be shown, using a more general version of the representer theorem, that the solution is a cubic spline with knots at the x_i.

cubic spline $\Leftrightarrow$ piecewise cubic polynomial.

Thus the optimization problem reduces to a finite dimensional one, and the optimal cubic spline results from solving a system of linear equations.



Reference] Nancy Heckman, "The theory and application of penalized methods, or Reproducing Kernel Hilbert Spaces made easy," 2011.

- Robust Kernel Density Estimation
  → Kim and Scott, "On the robustness of kernel density M-estimators," ICML 2011

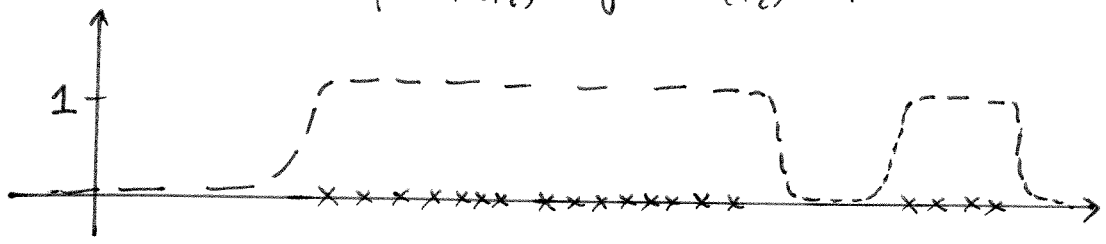
- one-class SVM for novelty detection

given training data from one class: $x_1, \ldots, x_n$

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \phi(f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \quad , \quad \lambda > 0$$

where $\phi(t) = \max(0, 1-t)$, so

$$\phi(f(x_i)) = \begin{cases} 0 & \text{if } f(x_i) \geq 1 \\ 1 - f(x_i) & \text{if } f(x_i) < 1 \end{cases}.$$



Final classifier: $\text{sign}(f(x) - \gamma)$

Using representer theorem and KKT conditions (as in the case of the SVM above), one can show the solution is $f(x) = \sum \alpha_i k(x_i, x_j)$, where $\alpha = (\alpha_1, \ldots, \alpha_n)$ solves

$$\max_{\alpha} \; -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) + \sum \alpha_i$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq \frac{1}{\lambda n}$$

Note: The OC-SVM is often parameterized in a different way: see Lee and Scott, "The one-class support vector machine solution path," ICASSP 2007.

- Multitask Learning

  → Evgeniou & Pontil, "Learning multiple tasks with kernel methods," JMLR, 2005.

- Transfer Learning

  → Blanchard, Lee, & Scott, "Generalizing from several related classification tasks to a new unlabeled sample," NIPS 2011.

Summary

- infinite dimensional optimization problem

  ↓ (representer theorem)

  finite dimensional optimization problem

- recover existing methods
- derive new ones