REPRODUCING KERNEL HILBERT SPACES

Overview

A reproducing kernel Hilbert space (RKHS) is a space of real-valued functions defined in terms of a positive definite kernel. By optimizing over a RKHS, we can derive many ML algorithms – some previously seen and some new.

A Rough Definition

Everything that follows can be made rigorous, but I will only present some rough ideas to enable some intuition.

Let $k: X \times X \to \mathbb{R}$ where X is the input space.
We previously asserted that the following are equivalent:

(a) k is symmetric and positive definite

(b) $\exists$ an inner product space $\mathcal{H}$ and a feature map

$\Phi: \mathcal{X} \to \mathcal{H}$ such that $k(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$.

The proof of (b) $\Rightarrow$ (a) was easy. To prove (a) $\Rightarrow$ (b), first define

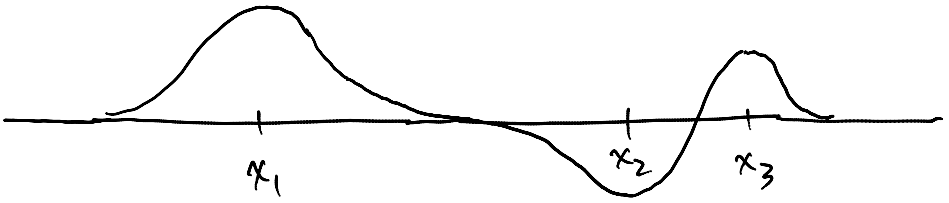
$$\mathcal{H}_0 = \left\{ \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \;\middle|\; m \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$

↑ set of natural numbers

where $k(\cdot, x')$ is the function

$$x \longmapsto k(x,x').$$

Thus, $\mathcal{H}_0$ is a space of functions. For example, if k is the Gaussian kernel, then



depicts an element of $\mathcal{H}_0$.

Define an inner product on $\mathcal{H}_0$ by

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

it can be shown that this is a valid inner product.

Now, consider the feature map $\Phi: \mathcal{X} \to \mathcal{H}_0$ given by

$$\Phi(x) = k(\cdot, x).$$

By definition of the inner product, observe

$$\langle \Phi(x), \Phi(x') \rangle = \langle k(\cdot, x), k(\cdot, x') \rangle$$
$$= k(x, x').$$

This establishes (b). The feature map above is known as the canonical feature map. It works for any PD kernel.

The reproducing property states that for any $f \in \mathcal{H}_0$, $x \in \mathcal{X}$

$$f(x) = \langle f, k(\cdot, x) \rangle.$$

To see this, let $f = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$. Then

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

For technical reasons that we won't cover, we need to enlarge $\mathcal{H}_0$ slightly. Define $\mathcal{H}$ to be the completion of $\mathcal{H}_0$, so that $\mathcal{H}$ is a Hilbert space (the name

for a complete inner product space). Completion essentially adds to $\mathcal{H}_0$ those functions that are not in $\mathcal{H}_0$, but that can be approximated arbitrarily accurately by elements of $\mathcal{H}_0$.

It can be shown that the reproducing property still holds for all $f \in \mathcal{H}$. $\mathcal{H}$ is known as the reproducing kernel Hilbert space associated with the PD kernel k. k is called the reproducing kernel of $\mathcal{H}$.

Remark The results and derivations below actually hold if we take $\mathcal{H} = \mathcal{H}_0$. Completeness of $\mathcal{H}$ is needed for more theoretical considerations.

The Representer Theorem

Previously, we have derived kernel methods by optimizing over a class of linear models and then kernelizing. As an alternative, we can optimize over the RKHS directly. Even though an RKHS may be infinite dimensional, by the following result, optimization problems of a certain type reduce to finite-dimensional problems.

reduce to finite-dimensional problems.

Theorem Let $\mathcal{H}$ be an RKHS consisting of functions defined on $\mathcal{X}$. Consider an optimization problem of the form

$$\min_{f \in \mathcal{H}} \; J(f)$$

where

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

for some $x_1, \ldots, x_n \in \mathcal{X}$, and where Λ is nondecreasing and $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$. Then there exists a minimizer of the form

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

Furthermore, if Λ is strictly increasing, then every minimizer has the given form.

Remark The notation $L(f(x_1), \ldots, f(x_n))$ indicates that this term does not depend on values of f outside of $\{x_1, \ldots, x_n\}$.

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

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

Since S is finite dimensional, the projection theorem holds and

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

where $S^{\perp} = \{ g \in \mathcal{H} \mid \langle f, g \rangle = 0 \quad \forall f \in S \}$.

Thus, for each $f \in \mathcal{H}$, we may write $f = f_{\parallel} + f_{\perp}$ where $f_{\parallel} \in S$, $f_{\perp} \in S^{\perp}$ are unique. Since $\forall i = 1, \ldots, n$

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

reproducing property

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

Let us write $L(f)$ for $L(f(x_1), \ldots, f(x_n))$. Then

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

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

Therefore, if f is a minimizer, then so is $f_{\parallel}$. Since

$f_{\|} \in S$, it has the desired form.

If Λ is strictly increasing, then $J(f_{\|}) < J(f)$ unless $f_{\perp} = 0$. Thus, every minimizer has the stated form. ∎

Kernel Ridge Regression

Let's apply the framework with

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

$$\Lambda(t) = \lambda t^2.$$

By the representer theorem, the solution of

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

has the form

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

Therefore it suffices to solve

$$\min_{\alpha} \quad \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \alpha_j k(x_i, x_j)\right)^2 + \lambda \left\| \sum_j \alpha_j k(\cdot, x_j) \right\|_{\mathcal{H}}^2$$

Now

$$\left\| \sum_j \alpha_j k(\cdot, x_j) \right\|^2 = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

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

$- \alpha \Lambda \alpha$

where

$$K = \left[k(x_i, x_j) \right]_{ij} \qquad (n \times n)$$

is the kernel matrix. Also,

$$\sum_{i=1}^{n} \left(y_i - \sum_j \alpha_j k(x_i, x_j) \right)^2 = \| y - K\alpha \|_2^2$$

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

So we need to minimize

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

Equating the gradient to zero we have

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

$$= 2K\left[(K + n\lambda I)\alpha - y \right]$$

which is solved by

$$\alpha = (K + n\lambda I)^{-1} y .$$

Therefore, we have re-derived kernel ridge regression without offset.

Support Vector Machines

Now consider

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

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

Again by the representer theorem, the minimizer has the form

$$f = \sum_{i=1}^{n} r_i \, k(\cdot, x_i), \qquad r_i \in \mathbb{R}.$$

Denoting $r = [r_1 \ \cdots \ r_n]^T$, and substituting $C = \frac{1}{\lambda}$, the optimization problem reduces to

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

Introducing slack variables $\xi = [\xi_1 \ \cdots \ \xi_n]^T$, this is equivalent to

$$\min_{r, \xi} \ \frac{1}{2} r^T K r + \frac{C}{n} \sum \xi_i$$

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

$$\xi_i \geq 0 \qquad \forall i.$$

This is a convex, differentiable optimization problem with affine constraints, so strong duality holds. The problem is also differentiable, and so the KKT conditions hold.

The Lagrangian is

$$L(r, \xi, \alpha, \beta) = \frac{1}{2} r^T K r + \frac{C}{n} \sum \xi_i$$

$$- \sum_{i=1}^{n} \alpha_i \left[y_i \left(\sum_j r_j k(x_i, x_j) \right) - 1 + \xi_i \right] - \sum_i \beta_i \xi_i .$$

Let us write

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

where $\odot$ denotes element-wise product. By the KKT conditions,

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

This is solved by

$$r = \alpha \odot y$$

$$\Updownarrow$$

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

Also,

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

Converting to the dual we have

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

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

This is the dual for the SVM without offset. We also recover the SVM decision function:

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

Kernel Logistic Regression

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

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

One-Class SVM

One-class classification refers to binary classification where there are training examples for only one of the classes.

Let $x_1, \ldots, x_n$ denote the data from the observed class.

We will consider a classifier of the form

$$x \longmapsto \text{sign}\left(f(x) - 1\right)$$

where $f \in \mathcal{H}$ and $+1$ is the observed class. To select f, we would like to trade off between two criteria:

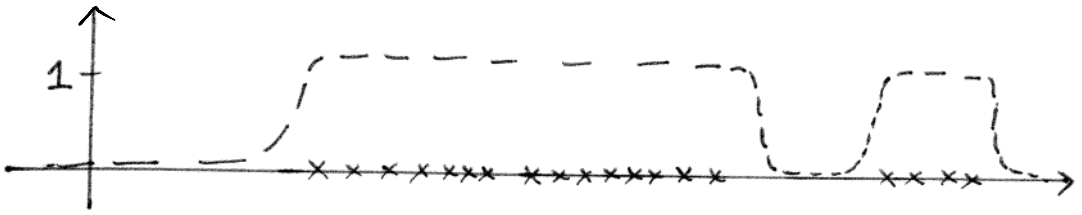
- Correctly classify training data
- Small $\|f\|$ to avoid overfitting.

Consider

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_i \phi\left(f(x_i)\right) + \frac{1}{2}\|f\|^2$$

where $\phi(t) = \max(0, 1-t)$ is the hinge loss. Notice

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



Using the representer theorem and KKT conditions as before, one can show that the solution is

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

where $\alpha = [\alpha_1 \cdots \alpha_n]^T$ solves

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

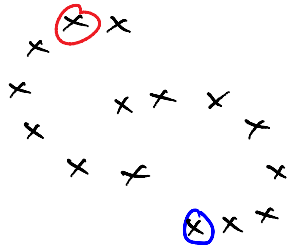
$$\text{s.t.} \quad 0 \leq \alpha_i \leq \frac{1}{\lambda n}.$$

Note] The OC-SVM is often parametrized differently. See Lee and Scott, "The one-class support vector machine solution path," ICASSP 2007.

Semi-Supervised Learning

Semi-supervised learning refers to a supervised learning problem where, in addition to labeled examples, there are also unlabeled examples. The goal is to leverage the unlabeled examples to improve the performance of a method that uses only the labeled data.

For example, consider a classification problem:



⊗ = labeled blue
⊗ = labeled red
× = unlabeled

Assuming the clusters are the classes, the unlabeled data help us learn the clusters which allows us to generalize well.

Let $(x_1, y_1), \ldots, (x_m, y_m), x_{m+1}, \ldots, x_{m+n}$ denote the training data. For regression, consider

$$\min_{f \in \mathcal{H}} \quad \lambda \|f\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_i (y_i - f(x_i))^2 + \gamma \cdot \underbrace{\frac{1}{2} \sum_{i,j} w_{ij} (f(x_i) - f(x_j))^2}_{= \underline{f}^T L \underline{f}}$$

where $\underline{f} = [f(x_{m+1}) \cdots f(x_{m+n})]^T$

where $W = [w_{ij}]_{i,j=1}^n$ is an $n \times n$ weighted adjacency matrix based on the unlabeled data. Intuitively, this term encourages f to take similar values at similar points. The solution can be derived in a manner similar to kernel ridge regression.