RKHS and Mercel Kernels

1 Hilbert Spaces

A Hilbert space is a complete inner product space. We will see that a reproducing kernel Hilbert space (RKHS) is a Hilbert space with extra structure that makes it very useful for statistics and machine learning.

An example of a Hilbert space is

$$L_2[0,1] = \left\{ f : [0,1] \to \mathbb{R} : \int f^2 < \infty \right\}$$

endowed with the inner product

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

The corresponding norm is

$$||f|| = \sqrt{\langle f, f \rangle} = \sqrt{\int f^2(x)dx}.$$

We write $f_n \to f$ to mean that $||f_n - f|| \to 0$ as $n \to \infty$.

2 Evaluation Functional

The evaluation functional δ_x assigns a real number to each function. It is defined by $\delta_x f = f(x)$. In general, the evaluation functional is not continuous. This means we can have $f_n \to f$ but $\delta_x f_n$ does not converge to $\delta_x f$. For example, let $f(x) = 0$ and $f_n(x) = \sqrt{n} I(x < 1/n^2)$. Then $||f_n - f|| = 1/\sqrt{n} \to 0$. But $\delta_0 f_n = \sqrt{n}$ which does not converge to $\delta_0 f = 0$. Intuitively, this is because Hilbert spaces can contain very unsmooth functions. We shall see that RKHS are Hilbert spaces where the evaluation functional is continous. Intuitively, this means that the functions in the space are well-behaved.

What has this got to do with kernels? Hang on; we're getting there.

3 Motivating Example: Nonparametric Regression

We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we want to estimate $m(x) = \mathbb{E}(Y|X = x)$. The approach we used earlier was based on **smoothing kernels**:

$$\widehat{m}(x) = \frac{\sum_i Y_i \; K\left(\frac{||x-X_i||}{h}\right)}{\sum_i K\left(\frac{||x-X_i||}{h}\right)}.$$

Another approach is regularization: choose m to minimize

$$\sum_i (Y_i - m(X_i))^2 + \lambda J(m)$$

for some penalty J. This is equivalent to: choose $m \in \mathcal{M}$ to minimize $\sum_i (Y_i - m(X_i))^2$ where $\mathcal{M} = \{m : \; J(m) \leq L\}$ for some $L > 0$.

We would like to construct $\mathcal{M}$ so that it contains smooth functions. We shall see that a good choice is to use a RKHS.

4 Mercer Kernels

A RKHS is defined by a **Mercer kernel**. A Mercer kernel $K(x, y)$ is a function of two variables that is symmetric and positive definite. This means that, for any function f,

$$\int \int K(x, y) f(x) f(y) dx \, dy \geq 0.$$

(This is like the definition of a positive definite matrix: $x^T A x \geq 0$ for each x.)

Our main example is the Gaussian kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{\sigma^2}}.$$

Given a kernel K, let $K_x(\cdot)$ be the function ontained by fixing the first coordinate. That is, $K_x(y) = K(x, y)$. For the Gaussian kernel, K_x is a Normal, centered at x. We can create functions by taking linear combinations of the kernel:

$$f(x) = \sum_{j=1}^{k} \alpha_j K_{x_j}(x).$$

Let $\mathcal{H}_0$ denote all such functions:

$$\mathcal{H}_0 = \left\{ f : \sum_{j=1}^{k} \alpha_j K_{x_j}(x) \right\}.$$

Given two such functions $f(x) = \sum_{j=1}^{k} \alpha_j K_{x_j}(x)$ and $g(x) = \sum_{j=1}^{m} \beta_j K_{y_j}(x)$ we define an inner product

$$\langle f, g \rangle = \langle f, g \rangle_K = \sum_i \sum_j \alpha_i \beta_j K(x_i, y_j).$$

In general, f (and g) might be representable in more than one way. You can check that $\langle f, g \rangle_K$ is independent of how f (or g) is represented. The inner product defines a norm:

$$||f||_K = \sqrt{\langle f, f, \rangle} = \sqrt{\sum_j \sum_k \alpha_j \alpha_k K(x_j, x_k)} = \sqrt{\alpha^T \mathbb{K} \alpha}$$

where $\alpha = (\alpha_1, \ldots, \alpha_k)^T$ and $\mathbb{K}$ is the $k \times k$ matrix with $\mathbb{K}_{jk} = K(x_j, x_k)$.

5 The Reproducing Property

Let $f(x) = \sum_i K_{x_i}(x)$. Note the following crucual property:

$$\langle f, K_x \rangle = \sum_i \alpha_i K(x_i, x) = f(x).$$

This follows from the definition of $\langle f, g \rangle$ where we take $g = K_x$. This implies that

$$\langle K_x, K_y \rangle = K(x, y).$$

This is called the reproducing property. It also implies that K_x is the **representer** of the evaluation functional.

The completion of $\mathcal{H}_0$ with respect to $|| \cdot ||_K$ is denoted by $\mathcal{H}_K$ and is called the RKHS generated by K.

To verify that this is a well-defined Hilbert space, you should check that the following properties hold:

$$\begin{aligned} \langle f, g \rangle &= \langle g, f \rangle \\ \langle cf + dg, h \rangle &= c\langle f, h \rangle + c\langle g, h \rangle \\ \langle f, f \rangle = 0 \quad &\text{iff} \quad f = 0. \end{aligned}$$

The last one is not obvious so let us verify it here. It is easy to see that $f = 0$ impies that $\langle f, f\rangle = 0$. Now we must show that $\langle f, f\rangle = 0$ implies that $f(x) = 0$. So suppose that $\langle f, f\rangle = 0$. Pick any x. Then

$$\begin{aligned} 0 &\leq f^2(x) = \langle f, K_x\rangle^2 = \langle f, K_x\rangle \, \langle f, K_x\rangle \\ &\leq ||f||^2 \; ||K_x||^2 = \langle f, f\rangle^2 \; ||K_x||^2 = 0 \end{aligned}$$

where we used Cauchy-Schwartz. So $0 \leq f^2(x) \leq 0$ which means that $f(x) = 0$.

Returning to the evaluation functional, suppose that $f_n \to f$. Then

$$\delta_x f_n = \langle f_n, K_x\rangle \to \langle f, K_x\rangle = f(x) = \delta_x f$$

so the evaluation functional is continuous. **In fact, a Hilbert space is a RKHS if and only if the evaluation functionals are continuous.**

6 Examples

Example 1 *Let $\mathcal{H}$ be all functions f on $\mathbb{R}$ such that the support of the Fourier transform of f is contained in $[-a, a]$. Then*

$$K(x, y) = \frac{\sin(a(y - x))}{a(y - x)}$$

and

$$\langle f, g\rangle = \int fg.$$

Example 2 *Let $\mathcal{H}$ be all functions f on $(0, 1)$ such that*

$$\int_0^1 (f^2(x) + (f'(x))^2)x^2 dx < \infty.$$

Then

$$K(x, y) = (xy)^{-1} \left(e^{-x}\sinh(y)I(0 < x \leq y) + e^{-y}\sinh(x)I(0 < y \leq x)\right)$$

and

$$||f||^2 = \int_0^1 (f^2(x) + (f'(x))^2)x^2 dx.$$

Example 3 *The Sobolev space of order m is (roughly speaking) the set of functions f such that $\int (f^{(m)})^2 < \infty$. For $m = 1$ and $\mathcal{X} = [0, 1]$ the kernel is*

$$K(x, y) = \begin{cases} 1 + xy + \frac{xy^2}{2} - \frac{y^3}{6} & 0 \leq y \leq x \leq 1 \\ 1 + xy + \frac{yx^2}{2} - \frac{x^3}{6} & 0 \leq x \leq y \leq 1 \end{cases}$$

and

$$||f||_K^2 = f^2(0) + f'(0)^2 + \int_0^1 (f''(x))^2 dx.$$

7 Spectral Representation

Suppose that $\sup_{x,y} K(x,y) < \infty$. Define eigenvalues λ_j and orthonormal eigenfunctions ψ_j by

$$\int K(x,y)\psi_j(y)dy = \lambda_j \psi_j(x).$$

Then $\sum_j \lambda_j < \infty$ and $\sup_x |\psi_j(x)| < \infty$. Also,

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\psi_j(y).$$

Define the **feature map** Φ by

$$\Phi(x) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots).$$

We can expand f either in terms of K or in terms of the basis $\psi_1, \psi_2, \ldots$:

$$f(x) = \sum_i \alpha_i K(x_i, x) = \sum_{j=1}^{\infty} \beta_j \psi_j(x).$$

Furthermore, if $f(x) = \sum_j a_j \psi_j(x)$ and $g(x) = \sum_j b_j \psi_j(x)$, then

$$\langle f, g \rangle = \sum_{j=1}^{\infty} \frac{a_j b_j}{\lambda_j}.$$

Roughly speaking, when $||f||_K$ is small, then f is smooth.

8 Representer Theorem

Let ℓ be a loss function depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$ and on $f(X_1), \ldots, f(X_n)$. Let $\widehat{f}$ minimize

$$\ell + g(||f||_K^2)$$

where g is any monotone increasing function. Then $\widehat{f}$ has the form

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

for some $\alpha_1, \ldots, \alpha_n$.

9 RKHS Regression

Define $\widehat{m}$ to minimize

$$R = \sum_i (Y_i - m(X_i))^2 + \lambda ||m||_K^2.$$

By the representer theorem, $\widehat{m}(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$. Plug this into R and we get

$$R = ||Y - \mathbb{K}\alpha||^2 + \lambda \alpha^T \mathbb{K} \alpha$$

where $\mathbb{K}_{jk} = K(X_j, X_k)$ is the Gram matrix. The minimizer over α is

$$\widehat{\alpha} = (\mathbb{K} + \lambda I)^{-1} Y$$

and $\widehat{m}(x) = \sum_j \widehat{\alpha}_j K(X_i, x)$. The fitted values are

$$\widehat{Y} = \mathbb{K}\widehat{\alpha} = \mathbb{K}(\mathbb{K} + \lambda I)^{-1} Y = LY.$$

So this is a linear smoother.

We can use cross-validation to choose λ. **Compare this with smoothing kernel regression.**

10 Logistic Regression

Let

$$m(x) = \mathbb{P}(Y = 1 | X = x) = \frac{e^{f(x)}}{1 + e^{f(x)}}.$$

We can estimate m by minimizing

$$-\text{loglikelihood} + \lambda ||f||_K^2.$$

Then $\widehat{f} = \sum_j K(x_j, x)$ and α may be found by numerical optimization; see the chapter. In this case, smoothing kernels are much easier.

11 Support Vector Machines

Suppose $Y_i \in \{-1, +1\}$. Recall the the linear SVM minimizes the penalized hinge loss:

$$J = \sum_i [1 - Y_i(\beta_0 + \beta^T X_i)]_+ + \frac{\lambda}{2} ||\beta||_2^2.$$

The dual is to maximize

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle$$

subject to $0 \leq \alpha_i \leq C$.

The RKHS version is to minimize

$$J = \sum_i [1 - Y_i f(X_i)]_+ + \frac{\lambda}{2} ||f||_K^2.$$

The dual is the same except that $\langle X_i, X_j \rangle$ is replaced with $K(X_i, X_j)$. Thisis called the kernel trick.

12 The Kernel Trick

This is a fairly general trick. In many algorithms you can replace $\langle x_i, x_j \rangle$ with $K(x_i, x_j)$ and get a nonlinear version of the algorithm. This is equivalent to replacing x with $\Phi(x)$ and replacing $\langle x_i, x_j \rangle$ with $\langle \Phi(x_i), \Phi(x_j) \rangle$. However, $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ and $K(x_i, x_j)$ is much easier to compute.

In summary, by replacing $\langle x_i, x_j \rangle$ with $K(x_i, x_j)$ we turn a linear procedure into a nonlinear procedure without adding much computation.

13 Hidden Tuning Parameters

There are hidden tuning parameters in the RKHS. Consider the Gaussian kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{\sigma^2}}.$$

For nonparametric regression we minimize $\sum_i (Y_i - m(X_i))^2$ subject to $||m||_K \leq L$. We control the bias variance tradeoff by doing cross-validation over L. But what about σ?

This parameter seems to get mostly ignored. Suppose we have a uniform distribution on a circle. The eigenfunctions of $K(x, y)$ are the sines and cosines. The eigenvalues λ_k die off like $(1/\sigma)^{2k}$. So σ affects the bias-variance tradeoff since it weights things towards lower order Fourier functions. In principle we can compensate for this by varying L. But clearly there is some intercation between L and σ. The practical effect is not well understood.

Now consider the polynomial kernel $K(x, y) = (1 + \langle x, y \rangle)^d$. This kernel has the same eigenfunctions but the eignvalues decay at a polynomial rate depending on d. So there is an interaction between L, d and, the choice of kernel itself.

14 Two Sample Test

Gretton, Borgwardt, Rasch, Scholkopf and Smola (GBRSS 2008) show how to use kernels for two sample testing. Suppose that

$$X_1, \ldots, X_m \sim P \qquad Y_1, \ldots, Y_n \sim Q.$$

We want to test the null hypothesis $H_0 : P = Q$.

Let $\mathcal{F} = \{f : \ ||f||_K \leq 1\}$. Define

$$M = \sup_{f \in \mathcal{F}} \Bigg| \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] \Bigg|.$$

Under weak regulaarty conditions on K, it can be shown that $M = 0$ if and only if $P = Q$. Thus we can test H_0 by estimating M.

Define

$$\widehat{M} = \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f(X_i) - \frac{1}{n} \sum_{i=1}^{m} f(Y_i) \right|.$$

Some calcculations show that

$$\widehat{M}^2 = \frac{1}{m^2} \sum_{j,k} K(X_j, X_k) - \frac{2}{mn} \sum_{j,k} K(X_j, Y_k) + \frac{1}{n^2} \sum_{j,k} K(Y_j, Y_k).$$

We reject H_0 if $\widehat{M} > t$. **We can determine t exactly using a permutation test.**

Using McDiarmmid's inequality and a Rademacher bound, GBRSS shows that

$$\mathbb{P}\left(|\widehat{M} - M| > 2\left(\sqrt{\frac{C}{m}} + \sqrt{\frac{C}{n}} \right) + \epsilon \right) \leq \exp\left(-\frac{\epsilon^2 mn}{C(m+n)} \right).$$

There is a connection with smoothing kernels. Let

$$\widehat{f}_X(u) = \frac{1}{m} \sum_{i=1}^{n} \kappa(X_i - u)$$

and similarly for $\widehat{f}_Y$. Then

$$\int |\widehat{f}_X(u) - \widehat{f}_Y(u)|^2 du = \widehat{M}^2$$

where $\widehat{M}$ is based on the kernel $K(x, y) = \int \kappa(x - z)\kappa(y - z)dz$. So they are really the same!

In practice, one would use the Gaussian kernel $K_\sigma(x,y) = e^{-\frac{||x-y||^2}{\sigma^2}}$. Call the resulting statistic $\widehat{M}_\sigma$. For hypothesis testing, there is no need to choose a bandwidth σ. Just define

$$\widehat{M} = \sup_\sigma \widehat{M}_\sigma.$$

Again, the critical value can be obtained using permutation methods. This is needed since the distribution of $\widehat{M}$ under H_0 is very complex and involved unknown quantities. (See Rosenbaum (2005, *Biometrika*) for a cool, two-sample test with an exact, known, distribution free null distribution.)