Reproducing Kernel Hilbert Spaces and Polynomial Smoothing Splines

Sam Watson

November 21, 2008

Introduction

Suppose we have some data \( \{x_i, y_i\} \in \mathcal{T} \times \mathbb{R} \) for \( i = 1, 2, \ldots, N \) and \( \mathcal{T} \) an arbitrary set, and we want to find a function \( f: \mathcal{T} \rightarrow \mathbb{R} \) that minimizes the penalized residual sum-of-squares

\[
\text{PRSS}(f) = \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f),
\]

where \( L \) is a loss function, \( \lambda \geq 0 \) is a parameter of the model to be fixed later by cross-validation, and \( J \) is a functional that imposes a smoothness criterion on \( f \). We will use reproducing kernel Hilbert spaces to develop a unifying framework for this problem and many other estimation and minimization problems in statistics.

Reproducing Kernel Hilbert Spaces

Recall that a Hilbert space \( \mathcal{H} \) is an inner-product space which is complete in the metric induced by its norm. For any Hilbert space of functions on a set \( \mathcal{T} \), we may define for each \( t \in \mathcal{T} \) the evaluation functional \( f \mapsto f(t) \). If every such evaluation functional is bounded, then we say that the Hilbert space is a reproducing kernel Hilbert space (RKHS). Observe that \( L^2 \) is not an RKHS, because the Dirac delta functions are not in \( L^2 \). For an RKHS, by the Riesz representation theorem we can find for each \( t \) an \( R_t \in \mathcal{H} \) so that

\[
f(t) = \langle f, R_t \rangle.
\]

Then we can define a function \( R: \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R} \) (the kernel) by \( R(s, t) = R_s(t) \). Such a function is symmetric because \( R_t(s) = \langle R_s, R_t \rangle = \langle R_t, R_s \rangle = R_s(t) \), and by direct application of (2) it has the reproducing property \( \langle R(\cdot, s), R(\cdot, t) \rangle = R(s, t) \). It is also nonnegative definite, meaning that for all \( t_1, t_2, \ldots, t_n \in \mathcal{T} \) and real numbers \( a_1, a_2, \ldots, a_n \),

\[
\sum_{i=1}^{n} a_i a_j R(t_i, t_j) \geq 0,
\]

\(^1\)The material presented here is standard; see References.
because the quantity on the left-hand side is equal to $\langle \sum_{i=1}^{n} R(\cdot, t_i) a_i, \sum_{i=1}^{n} R(\cdot, t_i) a_i \rangle = \|R(\cdot, t_i) a_i\|^2$. Moreover, each RKHS $\mathcal{H}$ has a unique kernel, because if $R$ and $R'$ are kernels, then $\|R(\cdot, y) - R'(\cdot, y)\|^2 = \langle R_y - R'_y, R_y - R'_y \rangle = \langle R_y - R'_y, R_y - R'_y \rangle = 0$ using the reproducing property. We now state an important theorem regarding the correspondence between reproducing kernel Hilbert spaces and their kernel functions $[1]$.

**Theorem 1.** To every RKHS there is a unique nonnegative definite kernel with the reproducing property, and conversely for any symmetric, nonnegative definite $R : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$, there is a unique RKHS $\mathcal{H}_R$ of functions on $\mathcal{T}$ whose kernel is $R$.

To obtain the RKHS for a kernel $R$, we first consider all finite linear combinations of the functions $\{R(\cdot, s) | s \in \mathcal{T}\}$, and define the inner product by $\langle R_s, R_t \rangle = R(s, t)$ along with linearity. We confirm that $R$ is a kernel for this space: first let $f = \sum a_i R(\cdot, t_i)$ where $i$ ranges over a finite index set and calculate $\langle \sum a_i R(\cdot, t_i), R(\cdot, s) \rangle = \sum a_i \langle R(\cdot, t_i), R(\cdot, s) \rangle = \sum a_i R(s, t_i) = f(t)$. A Cauchy sequence of such functions is guaranteed to converge to a limiting function pointwise, because $|f_n(t) - f_m(t)| = \langle f_n - f_m, R_t \rangle \leq \|f_n - f_m\|\|R_t\|$ by Cauchy-Schwarz. Therefore, we obtain a complete inner product space $\mathcal{H}_R$ if we include all linear combinations of the $R(\cdot, s)$ as well as limits of Cauchy sequences of these functions.

Let’s see generally how the RKHS approach can help with a problem like $[1]$. Suppose we can find a kernel $\tilde{R}$ whose RKHS $\mathcal{H}_{\tilde{R}}$ has a squared norm corresponding to the penalty functional $J$. Then by Hilbert space theory, for any $f \in \mathcal{H}_{\tilde{R}}$ we can write $f = \tilde{f} + f^\perp$, where $\tilde{f}$ is a linear combination of the $N$ functions $\{R(\cdot, x_i)\}_{i=1}^N$, and $\langle f^\perp, K(\cdot, x_i) \rangle = 0$ for all $i = 1, 2, \ldots, N$. The following theorem demonstrates the usefulness of this representation.

**Theorem 2.** For $f \in \mathcal{H}_R$ with $f = \tilde{f} + f^\perp$ and $\tilde{f} \in \text{span}(\{K(\cdot, x_i)\}_{i=1}^N)$,

$$
\sum_{i=1}^{N} L(y_i - \tilde{f}(x_i)) + \lambda \|f\|^2 \geq \sum_{i=1}^{N} L(y_i - \tilde{f}(x_i)) + \lambda \|\tilde{f}\|^2, \tag{3}
$$

with equality iff $f^\perp = 0$.

**Proof.** The summations on each side of (3) will be equal since they are only evaluated at the $N$ values $x_i$, and $f^\perp(x_i) = 0 \Rightarrow f(x_i) = \tilde{f}(x_i)$. Then observe that $\|f\|^2 = \|	ilde{f}\|^2 + \|f^\perp\|^2 \geq \|	ilde{f}\|^2$ with equality iff $f^\perp = 0$. \hfill \Box

This theorem shows that the solution for (3) is of the form

$$
f_\lambda = \sum_{i=1}^{N} c_i R(\cdot, x_i), \tag{4}
$$

and therefore reduces the infinite-dimension minimization problem $[1]$ to a finite-dimensional minimization problem. The sacrifice required was narrowing the space of candidate functions $f$ from the space of all functions on $\mathcal{T}$ to the space $\mathcal{H}_R$. It is therefore important to choose $R$ so that $\mathcal{H}_R$ contains a broad class of functions. Let us illustrate these ideas with an example.
Application: Polynomial Smoothing Splines

If the input data \( \{x_i\}_{i=1}^N \) are one-dimensional, then without loss of generality we may assume \( T = [0, 1] \). A common choice for smoothing a one-dimensional interpolating function is to penalize the integral of the square of the \( m \)th derivative of \( f \): \( J(f) = \int_0^1 (f^{(m)}(x))^2 \, dx \). Our goal is to construct an RKHS whose norm corresponds to this \( J \). Recall that Taylor’s Theorem with integral remainder term states that if \( f \) has \( m-1 \) absolutely continuous derivatives \([0, 1]\) and \( f^{(m)} \in L^2[0, 1] \), then

\[
f(t) = \sum_{i=1}^{m-1} \frac{t^i}{i!} f^{(i)}(0) + \int_0^t \frac{(t-x)^{m-1}}{(m-1)!} f^{(m)}(x) \, dx.
\]

(5)

It will be helpful to rewrite the integral as \( \int_0^1 \frac{(t-x)^{m-1}}{(m-1)!} f^{(m)}(x) \, dx \), where \( (x)_+ = x \) for \( x \geq 0 \) and 0 for negative \( x \). If we define \( W_m^0 \) to be the functions under the hypotheses of (5) whose first \( m-1 \) derivatives are zero at \( t = 0 \), then for \( f \in W_m^0 \), we have

\[
f(t) = \int_0^1 G_m(t,x) f^{(m)}(x) \, dx,
\]

(6)

where we have written \( G_m(t,x) \) for \( (t-x)^m / (m-1)! \). With the inner product

\[
\langle f, g \rangle = \int_0^1 f^{(m)}(x) g^{(m)}(x) \, dx
\]

and kernel \( R^1(s,t) = \int_0^1 G_m(r,x) G_m(s,x) \, dx \), the space \( W_m^0 \) forms an RKHS. To see that \( R^1 \) is a kernel, first notice that \( R^1_m(s,t) = G_m(s,t) \) from (6). Then \( \langle f, R^1(\cdot,t) \rangle = \int_0^1 f^{(m)}(x) G_m(s,x) \, dx = f(t) \). Now define the nullspace of the penalty functional: \( \mathcal{H}_0 = \text{span}(\{\phi_i\}_{i=1}^m) \), where \( \phi_i(t) = t^{i-1} / (i-1)! \). The kernel for \( \mathcal{H}_0 \) is \( R^0(s,t) = \sum_{j=1}^m \phi_i(s) \phi_j(t) \). It may easily be verified [1] that the space \( W_m \) of functions with \( m-1 \) absolutely continuous derivatives and \( m \) derivatives can be written as a direct sum \( \mathcal{H} = \mathcal{H}_0 \oplus W_m^0 \) with kernel \( R = R^1 + R^0 \). Also, \( J(f) \) corresponds to the squared norm of the projection \( Pf \) of \( f \) onto \( W_m^0 \), so (1) with \( J(f) = \int_0^1 (f^{(m)}(x))^2 \, dx \) becomes

\[
\sum_{i=1}^N L(y_i, f(x_i)) + \lambda \|P f\|^2
\]

(7)

for \( f \in \mathcal{H} \). This problem differs from (4) only by the appearance of the projection operator, and in fact by following the method of our proof of (3), it is easy to prove that the solution to (7) is the natural generalization of (4):

\[
f_{\lambda} = \sum_{i=1}^N c_i R^1(\cdot,x_i) + \sum_{j=1}^m d_j \phi_j.
\]

(8)

In other words, \( f_{\lambda} \) consists of an unpenalized component from \( \mathcal{H}_0 \) as well as a linear combination of the representations of evaluation at the \( N \) input data points. For the squared-error loss \( L(y_i - f(x_i)) = (y_i - f(x_i))^2 \), this solution actually corresponds to the natural polynomial spline. We will prove this in the case \( m = 2 \), but the general case is very similar.
Theorem 3. The function \( f_\lambda \in W_2 \) that minimizes the penalized residual sum of squares

\[
\text{PRSS}(f) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int_{0}^{1} (f''(x))^2 \, dx
\]  

is the natural cubic spline with nodes at \( x_1, \ldots, x_N \).

Proof. Calculate

\[
R(s,t) = \int_{0}^{1} G_m(r,x)G_m(s,x) \, dx = \int_{0}^{1} (r-x)_+(s-x)_+ \, dx = \begin{cases} \frac{1}{2}rs^2 - \frac{1}{6}s^3 & \text{if } s \leq r \\ \frac{1}{2}sr^2 - \frac{1}{6}r^3 & \text{if } r \leq s \end{cases}
\]

which is a cubic in \( r \) on \([0,s] \) and linear on \([s,1] \) with continuous first and second derivative. Then (8) differs from the natural cubic spline only in that the latter is required be linear on the interval \([0,x_1]\). We will show that this is also the case for \( f_\lambda \) (in what follows we will preserve arbitrary \( m \)). Define the \( N \times m \) matrix \((T)_{ij} = \phi_j(x_i)\) and the \( N \times N \) matrix \((S)_{ij} = R^1(x_i,x_j)\). Then in matrix notation, (9) becomes

\[
\frac{1}{N} \|y - (Sc +Td)\|^2 + \lambda c^TSc.
\]  

(10)

Taking derivatives to minimize (10), we find that

\[
c = M^{-1}(I - T(T^TM^{-1}T)^{-1}T^TM^{-1})y,
\]

where \( M = S + N\lambda I \). Therefore, \( T^Tc = 0 \), from which we obtain \( \sum_{i=1}^{N} c_i x_i^k = 0 \) for all \( k = 0,1,\ldots,m-1 \). Therefore, for \( t \in [0,x_1] \), we have

\[
Pf_\lambda (t) = \sum_{i=1}^{N} c_i R^1(t,x_i) = \int_{0}^{t} \frac{(t-x)^{m-1}}{(m-1)!} \sum_{i=1}^{N} (x_i-x)^{m-1} \, dx = 0,
\]

since no powers of the \( x_i \) greater than \( m-1 \) appear in the expansion of \( (x_i-x)^{m-1} \). Therefore, the only nonzero component of \( f_\lambda \) are the constant and linear terms from the projection of \( f_\lambda \) onto \( \mathcal{H}_0 \).

\[\Box\]

References

