Foundations of machine learning

Bonus Sides: Reproducing Kernel Hilbert Spaces and Splines

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2025

Splines and Reproducing Kernel Hilbert Spaces

- Penalized least squares: For some (semi-)norm $\|f\|$,

$$\widehat{f} = \underset{f}{\operatorname{argmin}} \sum_i (Y_i - f(X_i))^2 + \lambda \|f\|^2.$$

- Leading case: Splines, e.g.,

$$\widehat{f} = \underset{f}{\operatorname{argmin}} \sum_i (Y_i - f(X_i))^2 + \lambda \int f''(x)^2 dx.$$

- Can we think of penalized regressions in terms of a prior?

- If so, what is the prior distribution?

The finite dimensional case

- Consider the finite dimensional analog to penalized regression:

$$\widehat{\theta} = \underset{t}{\operatorname{argmin}} \sum_{i=1}^{n} (X_i - t_i)^2 + \|t\|_C^2,$$

where

$$\|t\|_C^2 = t' C^{-1} t.$$

- We saw before that this is the posterior mean when
  - $X|\theta \sim N(\theta, I_k)$,
  - $\theta \sim N(0, C)$.

The reproducing property

- The norm $\|t\|_C$ corresponds to the inner product
$$\langle t, s\rangle_C = t' C^{-1} s.$$
- Let $C_i = (C_{i1}, \ldots, C_{ik})'$.
- Then, for any vector y,
$$\langle C_i, y\rangle_C = y_i.$$

Practice problem

Verify this.

Reproducing kernel Hilbert spaces

- Now consider a general Hilbert space of functions equipped with an inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \|$,
- such that for all x there exists an M_x such that for all f

$$f(x) \leq M_x \cdot \|f\|.$$

- Read: "Function evaluation is continuous with respect to the norm $\| \cdot \|$."
- Hilbert spaces with this property are called reproducing kernel Hilbert spaces (RKHS).
- Note that L^2 spaces are not RKHS in general!

The reproducing kernel

- Riesz representation theorem:
  For every continuous linear functional L on a Hilbert space $\mathscr{H}$,
  there exists a $g_L \in \mathscr{H}$ such that for all $f \in \mathscr{H}$

$$L(f) = \langle g_L, f \rangle.$$

- Applied to function evaluation on RKHS:

$$f(x) = \langle C_x, f \rangle$$

- Define the reproducing kernel:

$$C(x_1, x_2) = \langle C_{x_1}, C_{x_2} \rangle.$$

- By construction:

$$C(x_1, x_2) = C_{x_1}(x_2) = C_{x_2}(x_1)$$

Practice problem

- Show that $C(\cdot,\cdot)$ is positive semi-definite, i.e.,
  for any $(x_1,\ldots,x_k)$ and $(a_1,\ldots,a_k)$

$$\sum_{i,j} a_i a_j C(x_i,x_j) \geq 0.$$

- Given a positive definite kernel $C(\cdot,\cdot)$,
  construct a corresponding Hilbert space.

Solution

- Positive definiteness:

$$\sum_{i,j} a_i a_j C(x_i, x_j) = \sum_{i,j} a_i a_j \langle C_{x_i}, C_{x_j} \rangle$$
$$= \left\langle \sum_i a_i C_{x_i}, \sum_j a_j C_{x_j} \right\rangle = \left\| \sum_i a_i C_{x_i} \right\|^2 \geq 0.$$

- Construction of Hilbert space: Take linear combinations of the functions $C(x, \cdot)$ (and their limits) with inner product

$$\left\langle \sum_i a_i C(x_i, \cdot), \sum_j b_j C(y_j, \cdot) \right\rangle_C = \sum_{i,j} a_i a_j C(x_i, y_j).$$

- Kolmogorov consistency theorem:
  For a positive definite kernel $C(\cdot,\cdot)$
  we can always define a corresponding prior

$$f \sim GP(0, C).$$

- Recap:
  - For each regression penalty,
  - when function evaluation is continuous w.r.t. the penalty norm
  - there exists a corresponding prior.
- Next:
  - The solution to the penalized regression problem
  - is the posterior mean for this prior.

Solution to penalized regression

- Let f be the solution to the penalized regression

$$\widehat{f} = \underset{f}{\operatorname{argmin}} \sum_i (Y_i - f(X_i))^2 + \lambda \|f\|_C^2.$$

Practice problem

- Show that the solution to the penalized regression has the form

$$\widehat{f}(x) = c(x) \cdot (C + n\lambda I)^{-1} \cdot Y,$$

where $C_{ij} = C(X_i, X_j)$ and $c(x) = (C(X_1, x), \ldots, C(X_n, x))$.

- Hints
  - Write $\widehat{f}(\cdot) = \sum a_i \cdot C(X_i, \cdot) + \rho(\cdot)$,
  - where ρ is orthogonal to $C(X_i, \cdot)$ for all i.
  - Show that $\rho = 0$.
  - Solve the resulting least squares problem in $a_1, \ldots, a_n$.

Solution

- Using the reproducing property, the objective can be written as

$$
\begin{aligned}
&\sum_i (Y_i - f(X_i))^2 + \lambda \|f\|_C^2 \\
=&\sum_i \left(Y_i - \langle C(X_i,\cdot), f\rangle\right)^2 + \lambda \|f\|_C^2 \\
=&\sum_i \left(Y_i - \left\langle C(X_i,\cdot), \sum_j a_j \cdot C(X_j,\cdot) + \rho \right\rangle\right)^2 + \lambda \left\| \sum_i a_i \cdot C(X_i,\cdot) + \rho \right\|_C^2 \\
=&\sum_i \left(Y_i - \sum_j a_j \cdot C(X_i, X_j)\right)^2 + \lambda \left(\sum_{i,j} a_i a_j C(x_i, x_j) + \|\rho\|_C^2\right) \\
=&\|Y - C\cdot a\|^2 + \lambda \left(a' C a + \|\rho\|_C^2\right)
\end{aligned}
$$

- Given a, this is minimized by setting $\rho = 0$.

- Now solve the quadratic program using first order conditions.

Splines

- Now what about the spline penalty

$$\int f''(x)^2 dx?$$

- Is function evaluation continuous for this norm?
- Yes, if we restrict to functions such that $f(0) = f'(0) = 0$.
- The penalty is a semi-norm that equals 0 for all linear functions.
- It corresponds to the GP prior with

$$C(x_1, x_2) = \frac{x_1 x_2^2}{2} - \frac{x_2^3}{6}$$

for $x_2 \leq x_1$.

- This is in fact the covariance of integrated Brownian motion!

Practice problem

Verify that C is indeed the reproducing kernel for the inner product

$$\langle f, g \rangle = \int_0^1 f''(x) g''(x) dx.$$

- Takeaway: Spline regression is equivalent to the limit of a posterior mean where the prior is such that

$$f(x) = A_0 + A_1 \cdot x + g$$

where

$$g \sim GP(0, C)$$

and

$$A \sim N(0, v \cdot I)$$

as $v \to \infty$.

Solution

- Have to show: $\langle C_x, g \rangle = g(x)$
- Plug in definition of C_x
- Last 2 steps: use integration by parts, use $g(0) = g'(0) = 0$
- This yields:

$$\begin{aligned}
\langle C_x, g \rangle &= \int C_x''(y) g''(y) dy \\
&= \int_0^x \left(\frac{xy^2}{2} - \frac{y^3}{6} \right)'' g''(y) dy + \int_x^1 \left(\frac{yx^2}{2} - \frac{x^3}{6} \right)'' g''(y) dy \\
&= \int_0^x (x - y) \, g''(y) dy \\
&= x \cdot (g'(x) - g'(0)) + \int_0^x g'(y) dy - (y g'(y)) \Big|_{y=0}^{x} \\
&= g(x).
\end{aligned}$$

References

- Gaussian process priors:
  Williams, C. and Rasmussen, C. (2006). Gaussian processes for machine learning. *MIT Press, chapter 2.*

- Splines and Reproducing Kernel Hilbert Spaces
  Wahba, G. (1990). Spline models for observational data, *volume 59. Society for Industrial Mathematics, chapter 1.*