Foundations of machine learning

Gaussian process priors, reproducing kernel Hilbert spaces, and Splines

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2024

Outline

- 6 equivalent representations of the posterior mean in the Normal-Normal model.
- Gaussian process priors for regression functions.
- Reproducing Kernel Hilbert Spaces and splines.
- Applications from my own work, to
    1. Optimal treatment assignment in experiments.
    2. Optimal insurance and taxation.

Takeaways for this part of class

- In a Normal means model with Normal prior, there are a number of equivalent ways to think about regularization.
- Posterior mean, penalized least squares, shrinkage, etc.
- We can extend from estimation of means to estimation of functions using Gaussian process priors.
- Gaussian process priors yield the same function estimates as penalized least squares regressions.
- Theoretical tool: Reproducing kernel Hilbert spaces.
- Special case: Spline regression.

Normal posterior means – equivalent representations
Setup

- $\theta \in \mathbb{R}^k$
- $\boldsymbol{X}|\theta \sim N(\theta, I_k)$
- Loss

$$L(\widehat{\theta}, \theta) = \sum_i (\widehat{\theta}_i - \theta_i)^2$$

- Prior

$$\theta \sim N(0, C)$$

6 equivalent representations of the posterior mean

1. Minimizer of weighted average risk
2. Minimizer of posterior expected loss
3. Posterior expectation
4. Posterior best linear predictor
5. Penalized least squares estimator
6. Shrinkage estimator

1) Minimizer of weighted average risk

- Minimize weighted average risk (= Bayes risk),
- averaging loss $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$ over both
  1. the sampling distribution $f_{\boldsymbol{X}|\theta}$, and
  2. weighting values of θ using the decision weights (prior) π_θ.
- Formally,

$$\widehat{\theta}(\cdot) = \underset{t(\cdot)}{\text{argmin}} \int E_\theta[L(t(\boldsymbol{X}), \theta)]d\pi(\theta).$$

2) Minimizer of posterior expected loss

- Minimize posterior expected loss,
- averaging loss $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$ over
  1. just the posterior distribution $\pi_{\theta|\boldsymbol{X}}$.
- Formally,

$$\widehat{\theta}(x) = \underset{t}{\operatorname{argmin}} \int L(t, \theta) d\pi_{\theta|\boldsymbol{X}}(\theta|x).$$

3 and 4) Posterior expectation and posterior best linear predictor

- Note that

$$\begin{pmatrix} X \\ \theta \end{pmatrix} \sim N\left(0, \begin{pmatrix} C+I & C \\ C & C \end{pmatrix}\right).$$

- Posterior expectation:

$$\widehat{\theta} = E[\theta|\boldsymbol{X}].$$

- Posterior best linear predictor:

$$\widehat{\theta} = E^*[\theta|\boldsymbol{X}] = C \cdot (C+I)^{-1} \cdot \boldsymbol{X}.$$

5) Penalization

- Minimize
  1. the sum of squared residuals,
  2. plus a quadratic penalty term.
- Formally,

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

- where

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

6) Shrinkage

- Diagonalize $\boldsymbol{C}$: Find
  1. orthonormal matrix $\boldsymbol{U}$ of eigenvectors, and
  2. diagonal matrix $\boldsymbol{D}$ of eigenvalues, so that
$$C = UDU'.$$

- Change of coordinates, using $\boldsymbol{U}$:
$$\tilde{\boldsymbol{X}} = U'\boldsymbol{X}$$
$$\tilde{\theta} = U'\theta.$$

- Componentwise shrinkage in the new coordinates:
$$\widehat{\tilde{\theta}}_i = \frac{d_i}{d_i + 1}\tilde{X}_i. \tag{1}$$

Practice problem

Show that these 6 objects are all equivalent to each other.

Solution (sketch)

1. Minimizer of weighted average risk = minimizer of posterior expected loss: See decision slides.
2. Minimizer of posterior expected loss = posterior expectation:
   - First order condition for quadratic loss function,
   - pull derivative inside,
   - and switch order of integration.
3. Posterior expectation = posterior best linear predictor:
   - $\boldsymbol{X}$ and $\boldsymbol{\theta}$ are jointly Normal,
   - conditional expectations for multivariate Normals are linear.
4. Posterior expectation $\Rightarrow$ penalized least squares:
   - Posterior is symmetric unimodal $\Rightarrow$ posterior mean is posterior mode.
   - Posterior mode = maximizer of posterior log-likelihood = maximizer of joint log likelihood,
   - since denominator $f_{\boldsymbol{X}}$ does not depend on $\boldsymbol{\theta}$.

Solution (sketch) continued

5. Penalized least squares $\Rightarrow$ posterior expectation:
   - Any penalty of the form
     $$t'At$$
     for A symmetric positive definite
   - corresponds to the log of a Normal prior
     $$\theta \sim N\left(0, A^{-1}\right).$$
6. Componentwise shrinkage $=$ posterior best linear predictor:
   - Change of coordinates turns $\widehat{\theta} = C \cdot (C+I)^{-1} \cdot \boldsymbol{X}$ into
     $$\widehat{\tilde{\theta}} = D \cdot (D+I)^{-1} \cdot \boldsymbol{X}.$$
   - Diagonality implies
     $$D \cdot (D+I)^{-1} = \operatorname{diag}\left(\frac{d_i}{d_i+1}\right).$$

Gaussian processes for machine learning
Machine Learning $\Leftrightarrow$ metrics dictionary

machine learning	metrics
supervised learning	regression
features	regressors
weights	coefficients
bias	intercept

Gaussian prior for linear regression

- Normal linear regression model:
- Suppose we observe $\boldsymbol{n}$ i.i.d. draws of (Y_i, X_i), where Y_i is real valued and $\boldsymbol{X}_i$ is a $\boldsymbol{k}$ vector.
- $Y_i = X_i \cdot \beta + \varepsilon_i$
- $\varepsilon_i | \boldsymbol{X}, \beta \sim N(0, \sigma^2)$
- $\beta | \boldsymbol{X} \sim N(0, \Omega)$ (prior)
- Note: will leave conditioning on $\boldsymbol{X}$ implicit in following slides.

Practice problem ("weight space view")

- Find the posterior expectation of β
- Hints:
  1. The posterior expectation is the maximum a posteriori.
  2. The $\log$ likelihood takes a penalized least squares form.
- Find the posterior expectation of $\boldsymbol{x} \cdot \beta$ for some (non-random) point $\boldsymbol{x}$.

Solution

- Joint log likelihood of Y, β:

$$\log(f_{\mathbf{Y}\beta}) = \log(f_{\mathbf{Y}|\beta}) + \log(f_\beta)$$
$$= const. - \frac{1}{2\sigma^2}\sum_i (Y_i - X_i\beta)^2 - \frac{1}{2}\beta'\Omega^{-1}\beta.$$

- First order condition for maximum a posteriori:

$$0 = \frac{\partial f_{\mathbf{Y}\beta}}{\partial \beta} = \frac{1}{\sigma^2}\sum_i (Y_i - X_i\beta)\cdot X_i - \beta'\Omega^{-1}.$$

$$\Rightarrow \quad \widehat{\beta} = \left(\sum_i X_i'X_i + \sigma^2\Omega^{-1}\right)^{-1} \cdot \sum X_i'Y_i.$$

- Thus

$$E[x\cdot\beta|\mathbf{Y}] = x\cdot\widehat{\beta} = x\cdot\left(\mathbf{X}'\mathbf{X} + \sigma^2\Omega^{-1}\right)^{-1}\cdot\mathbf{X}'\mathbf{Y}.$$

- Previous derivation required inverting $k \times k$ matrix.
- Can instead do prediction inverting an $n \times n$ matrix.
- n might be smaller than k if there are many "features."
- This will lead to a "function space view" of prediction.

Practice problem ("kernel trick")

- Find the posterior expectation of

$$f(x) = E[Y|X = x] = x \cdot \beta.$$

- Wait, didn't we just do that?
- Hints:
  1. Start by figuring out the variance / covariance matrix of $(\boldsymbol{x} \cdot \beta, \boldsymbol{Y})$.
  2. Then deduce the best linear predictor of $\boldsymbol{x} \cdot \beta$ given $\boldsymbol{Y}$.

Solution

- The joint distribution of $(x \cdot \beta, \boldsymbol{Y})$ is given by

$$\begin{pmatrix} x \cdot \beta \\ \boldsymbol{Y} \end{pmatrix} \sim N\left(0, \begin{pmatrix} x\Omega x' & x\Omega \boldsymbol{X}' \\ \boldsymbol{X}\Omega x' & \boldsymbol{X}\Omega \boldsymbol{X}' + \sigma^2 I_n \end{pmatrix}\right)$$

- Denote $\boldsymbol{C} = X\Omega X'$ and $\boldsymbol{c}(x) = x\Omega X'$.
- Then

$$E[x \cdot \beta | \boldsymbol{Y}] = \boldsymbol{c}(x) \cdot \left(\boldsymbol{C} + \sigma^2 I_n\right)^{-1} \cdot \boldsymbol{Y}.$$

- Contrast with previous representation:

$$E[x \cdot \beta | \boldsymbol{Y}] = x \cdot \left(\boldsymbol{X}'\boldsymbol{X} + \sigma^2 \Omega^{-1}\right)^{-1} \cdot \boldsymbol{X}'\boldsymbol{Y}.$$

General GP regression

- Suppose we observe $\boldsymbol{n}$ i.i.d. draws of (Y_i, X_i), where Y_i is real valued and X_i is a k vector.
- $Y_i = f(X_i) + \varepsilon_i$
- $\varepsilon_i | \boldsymbol{X}, f(\cdot) \sim N(0, \sigma^2)$
- Prior: f is distributed according to a Gaussian process,

  $$f | \boldsymbol{X} \sim GP(0, C),$$

  where C is a covariance kernel,

  $$\text{Cov}(f(x), f(x') | \boldsymbol{X}) = C(x, x').$$

- We will again leave conditioning on $\boldsymbol{X}$ implicit in following slides.

Practice problem

- Find the posterior expectation of $f(x)$.
- Hints:
  1. Start by figuring out the variance / covariance matrix of $(f(x), \boldsymbol{Y})$.
  2. Then deduce the best linear predictor of $f(x)$ given $\boldsymbol{Y}$.

Solution

- The joint distribution of $(f(x), \boldsymbol{Y})$ is given by

$$\begin{pmatrix} f(x) \\ \boldsymbol{Y} \end{pmatrix} \sim N\left(0, \begin{pmatrix} C(x,x) & c(x) \\ c(x)' & C + \sigma^2 I_n \end{pmatrix}\right),$$

  where
  - $c(x)$ is the n vector with entries $C(x, X_i)$,
  - and C is the $n \times n$ matrix with entries $C_{i,j} = C(X_i, X_j)$.

- Then, as before,

$$E[f(x)|\boldsymbol{Y}] = c(x) \cdot \left(C + \sigma^2 I_n\right)^{-1} \cdot \boldsymbol{Y}.$$

- Read: $\widehat{f}(\cdot) = E[f(\cdot)|\boldsymbol{Y}]$
  - is a linear combination of the functions $C(\cdot, X_i)$
  - with weights $(C + \sigma^2 I_n)^{-1} \cdot \boldsymbol{Y}$.

Hyperparameters and marginal likelihood

- Usually, covariance kernel $C(\cdot,\cdot)$ depends on on hyperparameters η.
- Example: squared exponential kernel with $\eta = (l, \tau^2)$ (length-scale l, variance τ^2).

$$C(x,x') = \tau^2 \cdot \exp\left(-\frac{1}{2l}\|x - x'\|^2\right)$$

- Following the empirical Bayes paradigm, we can estimate η by maximizing the marginal $\log$ likelihood:

$$\widehat{\eta} = \underset{\eta}{\operatorname{argmax}} \; -\frac{1}{2}|\det(C_\eta + \sigma^2 I)| - \frac{1}{2}\boldsymbol{Y}'(C_\eta + \sigma^2 I)^{-1}\boldsymbol{Y}$$

- Alternatively, we could choose η using cross-validation or Stein's unbiased risk estimate.

Splines and Reproducing Kernel Hilbert Spaces

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

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

- Leading case: Splines, e.g.,

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

- 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

$$\widehat{f} = \operatorname*{argmin}_{f} \sum_{i} (Y_i - f(X_i))^2 + \lambda \|f\|_C^2.$$

Practice problem

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

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

  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

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

- Given $\boldsymbol{a}$, 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.*