Foundations of machine learning

Gaussian process priors

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2025

Outline

- 6 equivalent representations of the posterior mean in the Normal-Normal model.
- Gaussian process priors for regression functions.
- Bonus slides: 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.

Normal posterior means – equivalent representations
Setup

- $\theta \in \mathbb{R}^k$
- $X|\theta \sim N(\theta, I_k)$
- 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 $f_{X|\theta}$, and
  2. weighting values of θ using the decision weights (prior) π_θ.
- Formally,

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

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 $\pi_{\theta|X}$.
- Formally,

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

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:
$$\widehat{\theta} = E[\theta|X].$$
- Posterior best linear predictor:
$$\widehat{\theta} = E^*[\theta|X] = C \cdot (C+I)^{-1} \cdot X.$$

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 C: Find
  1. orthonormal matrix U of eigenvectors, and
  2. diagonal matrix D of eigenvalues, so that
  $$C = UDU'.$$
- Change of coordinates, using U:
  $$\tilde{X} = U'X$$
  $$\tilde{\theta} = U'\theta.$$
- 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:
   - X and θ 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_X does not depend on θ.

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 $\widehat{\theta} = C \cdot (C+I)^{-1} \cdot X$ into
   $$\widehat{\tilde{\theta}} = D \cdot (D+I)^{-1} \cdot X.$$
   - Diagonality implies
   $$D \cdot (D+I)^{-1} = \operatorname{diag}\left(\frac{d_i}{d_i+1}\right).$$

Normal posterior means – equivalent representations

Gaussian process regression

References

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 n i.i.d. draws of (Y_i, X_i), where Y_i is real valued and X_i is a k vector.
- $Y_i = X_i \cdot \beta + \varepsilon_i$
- $\varepsilon_i | X, \beta \sim N(0, \sigma^2)$
- $\beta | X \sim N(0, \Omega)$ (prior)
- Note: will leave conditioning on 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 $x \cdot \beta$ for some (non-random) point x.

Solution

- Joint log likelihood of Y, β:

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

- First order condition for maximum a posteriori:

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

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

- Thus

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

- 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 $(x \cdot \beta, Y)$.
  2. Then deduce the best linear predictor of $x \cdot \beta$ given Y.

Solution

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

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

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

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

- Contrast with previous representation:

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

General GP regression

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

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

  where C is a covariance kernel,

  $$\mathrm{Cov}(f(x), f(x') | X) = C(x, x').$$
- We will again leave conditioning on 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 $(f(x), Y)$.
  2. Then deduce the best linear predictor of $f(x)$ given Y.

Solution

- The joint distribution of $(f(x), Y)$ is given by

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

  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,

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

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

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:

$$\widehat{\eta} = \operatorname*{argmax}_{\eta} -\tfrac{1}{2}|\det(C_\eta + \sigma^2 I)| - \tfrac{1}{2} Y'(C_\eta + \sigma^2 I)^{-1} Y$$

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

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.*