MLE Sheet 8

Aarohi Verma

June 2025

Exercise 2: Bayesian Linear Regression and Basis Function Expansion

Task 1

We are given the following setup for Bayesian Linear Regression:

Prior

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}_p)$$

Likelihood

$$\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

Hint: Likelihood as a Gaussian in $\boldsymbol{\beta}$

By viewing the likelihood as a function of $\boldsymbol{\beta}$ (up to proportionality), it is proportional to:

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}) \propto \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\beta}_{\text{MLE}}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right)$$

The posterior is proportional to the product of the prior and the likelihood:

$$p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, \tau^2 \mathbf{I}_p) \times \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\beta}_{\text{MLE}}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right)$$

Product of Two Gaussians

Recall the result: if

$$\mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \times \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \propto \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$$

then:

$$\boldsymbol{\Sigma}_{\text{post}} = \left(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Sigma}_{\text{post}} \left(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2\right)$$

Let:

$$\boldsymbol{\mu}_1 = \mathbf{0}, \quad \boldsymbol{\Sigma}_1 = \tau^2 \mathbf{I}_p$$
$$\boldsymbol{\mu}_2 = \boldsymbol{\beta}_{\text{MLE}}, \quad \boldsymbol{\Sigma}_2 = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

Then:

Posterior Covariance:

$$\boldsymbol{\Sigma}_{\text{post}} = \left(\frac{1}{\tau^2} \mathbf{I}_p + \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} \right)^{-1}$$

Posterior Mean:

$$\begin{aligned} \boldsymbol{\mu}_{\text{post}} &= \boldsymbol{\Sigma}_{\text{post}} \left(\frac{1}{\tau^2} \mathbf{0} + \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}_{\text{MLE}} \right) \\ &= \boldsymbol{\Sigma}_{\text{post}} \cdot \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

Final Posterior Distribution

The posterior distribution of $\boldsymbol{\beta}$ given data is:

$$p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$$

with:

$$\boxed{\boldsymbol{\Sigma}_{\text{post}} = \left(\frac{1}{\tau^2} \mathbf{I}_p + \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} \right)^{-1}} \quad \text{and} \quad \boxed{\boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Sigma}_{\text{post}} \cdot \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{y}}$$

Task 2

In a fully Bayesian model, we want to make predictions by integrating over the posterior distribution of the parameters. For a new input $\mathbf{x}_\star$, the predictive distribution is given by:

$$p(y_\star \mid \mathbf{x}_\star, \mathbf{X}, \mathbf{y}) = \int p(y_\star \mid \mathbf{x}_\star, \boldsymbol{\beta})\, p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y})\, d\boldsymbol{\beta}$$

This integral accounts for:

- **Aleatoric uncertainty** — the inherent noise in observations.
- **Epistemic uncertainty** — uncertainty about the model parameters.

Why is the Predictive Distribution Gaussian?

The likelihood is Gaussian:

$$y_\star \mid \mathbf{x}_\star, \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{x}_\star^\top \boldsymbol{\beta}, \sigma^2)$$

The posterior distribution over the parameters is also Gaussian:

$$\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$$

Since the predictive distribution involves a linear transformation of a Gaussian variable (plus independent Gaussian noise), the resulting marginal distribution is also Gaussian.

Predictive Mean

The predictive mean is the expectation of $y_\star$ under the posterior over $\boldsymbol{\beta}$:

$$m(\mathbf{x}_\star) = E_{p(\boldsymbol{\beta}|\mathbf{y})}\left[\mathbf{x}_\star^\top \boldsymbol{\beta}\right] = \mathbf{x}_\star^\top \boldsymbol{\mu}_{\text{post}}$$

Predictive Variance

We use the law of total variance:

$$v(\mathbf{x}_\star) = E_{\boldsymbol{\beta}}\left[\text{Var}(y_\star \mid \mathbf{x}_\star, \boldsymbol{\beta})\right] + \text{Var}_{\boldsymbol{\beta}}\left[E(y_\star \mid \mathbf{x}_\star, \boldsymbol{\beta})\right]$$

Compute the two terms:

$$\begin{aligned} \text{Var}(y_\star \mid \mathbf{x}_\star, \boldsymbol{\beta}) &= \sigma^2 \\ \text{Var}_{\boldsymbol{\beta}}(\mathbf{x}_\star^\top \boldsymbol{\beta}) &= \mathbf{x}_\star^\top \boldsymbol{\Sigma}_{\text{post}} \mathbf{x}_\star \end{aligned}$$

Therefore, the predictive variance is:

$$v(\mathbf{x}_\star) = \sigma^2 + \mathbf{x}_\star^\top \boldsymbol{\Sigma}_{\text{post}} \mathbf{x}_\star$$

Final Predictive Distribution

The predictive distribution is:

$$\boxed{p(y_\star \mid \mathbf{x}_\star, \mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\mathbf{x}_\star^\top \boldsymbol{\mu}_{\text{post}},\ \sigma^2 + \mathbf{x}_\star^\top \boldsymbol{\Sigma}_{\text{post}} \mathbf{x}_\star\right)}$$

Posterior Parameters (from Previous Derivation)

$$\begin{aligned} \boldsymbol{\Sigma}_{\text{post}} &= \left(\frac{1}{\tau^2}\mathbf{I}_p + \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{X}\right)^{-1} \\ \boldsymbol{\mu}_{\text{post}} &= \boldsymbol{\Sigma}_{\text{post}} \cdot \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{y} \end{aligned}$$

Task 3

We extend Bayesian linear regression by applying a nonlinear basis function expansion to the input data using a feature map:

$$\boldsymbol{\phi} : R^n \to R^D, \quad x \mapsto \boldsymbol{\phi}(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_D(x) \end{bmatrix}$$

Our model becomes:

$$y = \boldsymbol{\phi}(x)^\top \mathbf{w} + \varepsilon, \quad \mathbf{w} \sim \mathcal{N}(0, \tau^2 \mathbf{I}_D), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

(a) Why must the model remain linear in parameters?

Even though the model is nonlinear in the input x, it is crucial that the model remains **linear in the parameters w** so that:

- The **conjugacy** between the Gaussian prior and Gaussian likelihood is preserved.
- We retain closed-form expressions for the posterior and predictive distributions.
- The model remains computationally tractable, and Bayesian inference remains analytically solvable.

This strategy allows us to model nonlinear relationships while still using linear algebra tools.

(b) Bayesian Model in Expanded Feature Space

Let the transformed feature matrix be:

$$\Phi = \begin{bmatrix} \boldsymbol{\phi}(x_1)^\top \\ \boldsymbol{\phi}(x_2)^\top \\ \vdots \\ \boldsymbol{\phi}(x_n)^\top \end{bmatrix} \in R^{n \times D}$$

Likelihood:

$$p(\mathbf{y} \mid \Phi, \mathbf{w}) = \mathcal{N}(\Phi \mathbf{w}, \sigma^2 \mathbf{I}_n)$$

Prior:

$$\mathbf{w} \sim \mathcal{N}(0, \tau^2 \mathbf{I}_D)$$

Posterior:

$$\mathbf{w} \mid \Phi, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$$

where

$$\boldsymbol{\Sigma}_{\text{post}} = \left(\frac{1}{\tau^2} \mathbf{I}_D + \frac{1}{\sigma^2} \Phi^\top \Phi \right)^{-1}, \quad \boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Sigma}_{\text{post}} \cdot \frac{1}{\sigma^2} \Phi^\top \mathbf{y}$$

Predictive Mean for New Input $x_\star$:

$$m(x_\star) = E[y_\star \mid x_\star, \mathbf{y}] = \boldsymbol{\phi}(x_\star)^\top \boldsymbol{\mu}_{\text{post}}$$

(c) Dependence on Inner Products

Using the posterior mean:

$$m(x_\star) = \boldsymbol{\phi}(x_\star)^\top \boldsymbol{\mu}_{\text{post}} = \frac{1}{\sigma^2} \boldsymbol{\phi}(x_\star)^\top \boldsymbol{\Sigma}_{\text{post}} \Phi^\top \mathbf{y}$$

Substitute $\boldsymbol{\Sigma}_{\text{post}}$:

$$m(x_\star) = \frac{1}{\sigma^2} \boldsymbol{\phi}(x_\star)^\top \left(\frac{1}{\tau^2} \mathbf{I}_D + \frac{1}{\sigma^2} \Phi^\top \Phi \right)^{-1} \Phi^\top \mathbf{y}$$

This expression only involves **inner products** of the form $\boldsymbol{\phi}(x_i)^\top \boldsymbol{\phi}(x_j)$ inside $\Phi^\top \Phi$, showing that the prediction depends only on those inner products.

(d) Applying the Kernel Trick

Define the **kernel matrix** $\mathbf{K} = \Phi \Phi^\top \in R^{n \times n}$, with entries:

$$K_{ij} = k(x_i, x_j) = \boldsymbol{\phi}(x_i)^\top \boldsymbol{\phi}(x_j)$$

Define the kernel vector:

$$\mathbf{k}_\star = \begin{bmatrix} k(x_1, x_\star) \\ k(x_2, x_\star) \\ \vdots \\ k(x_n, x_\star) \end{bmatrix} = \Phi \boldsymbol{\phi}(x_\star)$$

It can be shown that:

$$\frac{1}{\sigma^2} \boldsymbol{\phi}(x_\star)^\top \boldsymbol{\Sigma}_{\text{post}} \Phi^\top = \mathbf{k}_\star^\top \left(\mathbf{K} + \frac{\sigma^2}{\tau^2} \mathbf{I}_n \right)^{-1}$$

Therefore, the predictive mean becomes:

$$\boxed{m(x_\star) = \mathbf{k}_\star^\top \left(\mathbf{K} + \frac{\sigma^2}{\tau^2} \mathbf{I}_n \right)^{-1} \mathbf{y}}$$

This expression uses only the kernel $k(x, x')$, eliminating the need to compute $\boldsymbol{\phi}(x)$ explicitly.

Predictive Variance (Sketch): The predictive variance also becomes:

$$v(x_\star) = k(x_\star, x_\star) - \mathbf{k}_\star^\top \left(\mathbf{K} + \frac{\sigma^2}{\tau^2} \mathbf{I}_n \right)^{-1} \mathbf{k}_\star$$

(e) From Kernel Model to Gaussian Processes

The final conceptual step is to interpret the kernel function $k(x, x')$ as a **covariance function** of a stochastic process. That is:

$$f(x) \sim \mathcal{GP}(0, k(x, x'))$$

Key observations:

- Instead of placing a prior on parameters $\mathbf{w}$, we place a prior on functions $f(x)$.
- The kernel function defines the covariance between function values at different inputs.
- As the number of basis functions $D \to \infty$, the Bayesian linear model with basis expansion converges to a **Gaussian Process**.
- This makes GPs a fully non-parametric Bayesian model, where inference and prediction are performed entirely using the kernel function.