Week 5 Tutorial Tasks

Sarat Moka

With a single predictor but a high-degree polynomial basis, the coding task illustrates

1. how a Gaussian prior prevents over-fit, and
2. how Bayesian credible bands differ from OLS confidence bands.

Theoretical Task: OLS vs Bayesian Linear Regression

We use a single predictor x but fit a degree-d polynomial basis

$$\phi(x) = [1,\, x,\, x^2, \ldots, x^d]^\top.$$

Denote $X \in \mathbb{R}^{n\times(d+1)}$ with rows $\phi(x_i)^\top$, and observe noisy targets $y_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Assume that $d \le n$ and x_i's are distinct.

Task 1. Derive the ordinary least squares estimator $\hat{\beta}_{\text{OLS}}$ that minimizes the mean squared error $\|y - X\beta\|_2^2$.

Solution. Set gradient to zero:

$$\frac{\partial}{\partial\beta}\|y - X\beta\|_2^2 = 2\,X^\top(X\beta - y) \overset{!}{=} 0 \quad\Longrightarrow\quad X^\top X\,\beta = X^\top y.$$

Since $X^\top X$ is invertible,

$$\boxed{\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1}\,X^\top y.}$$

Task 2. Introduce a Gaussian prior $\beta \sim \mathcal{N}(0, \tau^2 I)$, where I denotes the identity matrix of appropriate dimension. Derive the posterior distribution $p(\beta \mid X, y)$.

Solution. The likelihood is $\mathcal{N}(X\beta, \sigma^2 I)$ and the prior on β is $\mathcal{N}(0, \tau^2 I)$. Then up to normalisation,

$$p(\beta \mid X, y) \propto \exp\Bigl(-\tfrac{1}{2\sigma^2}\|y - X\beta\|^2 - \tfrac{1}{2\tau^2}\|\beta\|^2\Bigr).$$

Collect the quadratic form in β:

$$-\tfrac{1}{2}\,\beta^\top\Bigl(\tfrac{1}{\sigma^2}X^\top X + \tfrac{1}{\tau^2}I\Bigr)\beta \;+\; \beta^\top\Bigl(\tfrac{1}{\sigma^2}X^\top y\Bigr) + \text{const}.$$

Completing the square yields a Gaussian posterior with

$$\Sigma_{\text{post}} = \Bigl(\tfrac{1}{\sigma^2}X^\top X + \tfrac{1}{\tau^2}I\Bigr)^{-1}, \qquad \mu_{\text{post}} = \Sigma_{\text{post}}\,\frac{1}{\sigma^2}X^\top y,$$

so

$$\boxed{p(\beta \mid X, y) = \mathcal{N}\bigl(\mu_{\text{post}},\, \Sigma_{\text{post}}\bigr).}$$

Task 3. Derive the pointwise 95% *confidence interval* under OLS and the 95% *credible interval* under the Bayesian posterior for the predicted value $\hat{f}(x_*) = \phi(x_*)^\top\beta$.

Solution.

- *OLS confidence interval.* The OLS estimate at x_* is $\hat{f}_{\text{OLS}}(x_*) = \phi(x_*)^\top\hat{\beta}_{\text{OLS}}$. Its sampling variance is

  $$\mathbb{V}ar\bigl(\hat{f}_{\text{OLS}}(x_*)\bigr) = \sigma^2\,\phi(x_*)^\top(X^\top X)^{-1}\phi(x_*).$$

  A 95% confidence interval is

  $$\hat{f}_{\text{OLS}}(x_*) \;\pm\; 1.96\cdot\sigma\cdot\sqrt{\phi(x_*)^\top(X^\top X)^{-1}\phi(x_*)}.$$

- *Bayesian credible interval.* When you ask "What will I observe next?" you need to account for two sources of uncertainty:

    1. Parameter uncertainty: the posterior spread in β;
    2. Observation noise: the fact that even if β were known exactly, each new ε still has variance σ^2;

    Thus the prediction for a new input x_* is:

    $$f(x_*) = \phi(x_*)^\top \beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

    Using the law of total expectation:

    $$\begin{aligned}\mathbb{E}[f(x_*) \mid X, y] &= \mathbb{E}\big[\phi(x_*)^\top \beta + \epsilon \mid X, y\big] \\ &= \phi(x_*)^\top \mathbb{E}[\beta \mid X, y] + \mathbb{E}[\epsilon] \\ &= \phi(x_*)^\top \mu_{\text{post}} + 0.\end{aligned}$$

    Thus predictive mean is

    $$\boxed{\mathbb{E}[f(x_*) \mid X, y] = \phi(x_*)^\top \mu_{\text{post}}}$$

    By the law of total variance:

    $$\begin{aligned}\mathbb{V}ar[f(x_*) \mid X, y] &= \mathbb{V}ar\big[\phi(x_*)^\top \beta + \epsilon \mid X, y\big] \\ &= \mathbb{V}ar\big[\phi(x_*)^\top \beta \mid X, y\big] + \mathbb{V}ar[\epsilon] \\ &= \phi(x_*)^\top \mathbb{V}ar(\beta \mid X, y)\, \phi(x_*) + \sigma^2 \\ &= \phi(x_*)^\top \Sigma_{\text{post}}\, \phi(x_*) + \sigma^2.\end{aligned}$$

    Thus predictive variance is

    $$\boxed{\mathbb{V}ar[f(x_*) \mid X, y] = \phi(x_*)^\top \Sigma_{\text{post}} \phi(x_*) + \sigma^2}$$

    Since $f(x_*)$ given (X, y) is Gaussian, its $100(1-\alpha)\%$ credible interval is:

    $$\mu \pm z_{1-\alpha/2} \sqrt{\mathbb{V}ar},$$

    where $z_{0.975} \approx 1.96$. By substituting the above values, the 95% credible interval to be

    $$\boxed{\phi(x_*)^\top \mu_{\text{post}} \;\pm\; 1.96 \sqrt{\phi(x_*)^\top \Sigma_{\text{post}} \phi(x_*) + \sigma^2}}$$

Coding Task

Consider above polynomial regression when the underlying model is

$$y = \sin(2\pi\, x) + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 0.3)$. Taking 30 linearly spaced x_i's on $[0, 1]$, compute the above confidence intervals and credible intervals for $d = 20$.

Soluation: Check the jupyter-notebook solution of this tutorial.