Bayesian Statistics

Wenjie Tu

Spring Semester 2022

Basics in Bayesian statistics

- Likelihood: $f(x \mid \theta)$
- Prior: $\pi(\theta)$
- Posterior: $\pi(\theta \mid x) = \dfrac{\pi(\theta) f(x \mid \theta)}{f(x)} \propto \pi(\theta) f(x \mid \theta)$
- Prior predictive density: $f(x) = \int f(x \mid \theta)\pi(\theta) d\theta$
- Posterior predictive density: $f(y \mid x) = \int f(y \mid x, \theta)\pi(\theta \mid x) d\theta$

Bayesian point estimates

- Posterior mean:
$$\mathbb{E}((\theta - T)^2 \mid x) = \mathbb{E}(\theta^2 \mid x) - 2\mathbb{E}(\theta \mid x)T + T^2$$
This is minimized for $T = T(X) = \mathbb{E}(\theta \mid x)$
- Posterior median:
$$\mathbb{E}(|\theta - T| \mid x) = \int_{-\infty}^{T} (T - \theta)\pi(\theta \mid x) d\theta + \int_{T}^{\infty} (\theta - T)\pi(\theta \mid x) d\theta$$
Using the Leibniz integral rule it follows that
$$\frac{\partial}{\partial T}\mathbb{E}(|\theta - T| \mid x) = \int_{-\infty}^{T} \pi(\theta \mid x) d\theta - \int_{T}^{\infty} \pi(\theta \mid x) d\theta$$
This equals zero if $T = T(X) = \text{median}\pi(\theta \mid x)$
- Posterior mode:
$$\mathbb{E}(1_{[-\varepsilon,\varepsilon]^c}(T - \theta) \mid x) = 1 - \int_{T-\varepsilon}^{T+\varepsilon} \pi(\theta \mid x) d\theta$$
For small ε, we have
$$\int_{T-\varepsilon}^{T+\varepsilon} \pi(\theta \mid x) d\theta \approx 2\varepsilon\pi(\theta \mid x)$$
This is maximized, i.e., $\mathbb{E}(1_{[-\varepsilon,\varepsilon]^c}(T - \theta) \mid x)$ is minimized, for $T = T(X) = \text{mode}\pi(\theta \mid x)$

Bayesian decision theory

- Posterior risk:

$$\rho(T(x),\pi) = \mathbb{E}(L(T(X),\theta) \mid x) = \int_\Theta L(T(x),\theta)\pi(\theta \mid x)d\theta$$

- Frequentist risk:

$$R(T,\theta) = \mathbb{E}_\theta(L(T(X),\theta),\theta) = \int_{\mathbf{X}} L(T(x),\theta)f(x \mid \theta)dx$$

- Bayes factor:

$$B_{01}(x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)} = \frac{\pi(\theta_0 \mid x)\pi(\theta_1)}{\pi(\theta_1 \mid x)\pi(\theta_0)} = \frac{\frac{\pi(\theta_0|x)}{\pi(\theta_1|x)}}{\frac{\pi(\theta_0}{\pi(\theta_1)}} = \frac{\text{Posterior odds}}{\text{Prior odds}}$$

Bayesian asymptotics

- Frequentist asymptotocs:

$$\widehat{\theta}_n \overset{\text{approx}}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{n}I(\theta_n)^{-1}\right)$$

$$2\left(\log L_n(\widehat{\theta}_n) - \log L_n(\theta_n)\right) \xrightarrow{\text{d}} \chi^2_p$$

- Bayesian asymptotics:

$$\theta \mid (x_1, \cdots, x_n) \overset{\text{approx}}{\sim} \mathcal{N}\left(\widehat{\theta}_n, \frac{1}{n}I(\widehat{\theta}_n)^{-1}\right)$$

Likelihood principle

- Repeat the trial a fixed number n times and observe the random number X of trials where the event occurred:

$$P(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$$

  - Binomial distribution
  - Reject H_0 if $x > c_1$

- Repeat the experiment a random number of N times until the event has occurred a fixed number of times x:

$$P(N = n) = \binom{n-1}{x-1} p^x(1-p)^{n-x}$$

  - Negative binomial distribution
  - Reject H_0 if $n < c_2$

Priors

- Conjugate priors:

$$\pi_\xi(\theta)f(x \mid \theta) \propto \pi_{\xi'}(\theta)$$

- Improper priors:

$$\int_\theta \pi(\theta) = \infty$$

- Jeffreys prior:

$$\pi(\theta) \propto \det(I(\theta))^{1/2}$$

- Reference prior:

$$
\begin{aligned}
I(X,\theta) &= \int_X f(x) \int_\Theta \pi(\theta \mid x) \log \frac{\pi(\theta \mid x)}{\pi(\theta)} d\theta dx \\
&= \int_X f(x) \int_\Theta \frac{\pi(x,\theta)}{f(x)} \log \frac{\pi(x,\theta)}{\pi(\theta) f(x)} d\theta dx \\
&= \int_{X\times\Theta} \pi(x,\theta) \log \frac{\pi(x,\theta)}{\pi(\theta) f(x)} dx d\theta \\
&= KL(\pi(x,\theta), \pi(\theta) f(x))
\end{aligned}
$$

Hierarchical Bayes models

- Hierarchical Bayes models:

$$\pi(\xi)\pi(\theta \mid \xi) f(x \mid \theta)$$

- Marginal posterior: approach 1
  - Compute the marginal prior:

$$\pi(\theta) = \int \pi(\theta \mid \xi)\pi(\xi) d\xi$$

  - Then use Bayes formula:

$$\pi(\theta \mid x) \propto \pi(\theta) f(x \mid \theta)$$

- Marginal posterior: approach 2
  - Law of total probability (computationally easier):

$$\pi(\theta \mid x) = \int \pi(\theta \mid x,\xi)\pi(\xi \mid x) d\xi \propto \int \pi(\theta \mid x,\xi)\pi(\xi) f(x \mid \xi) d\xi$$

    * $\pi(\xi \mid x) = \int \pi(\theta, \xi \mid x) d\theta = \int \pi(\theta \mid x)\pi(\theta \mid \xi) d\theta$
    * $\pi(\xi \mid x) = \dfrac{\pi(\theta,\xi \mid x)}{\pi(\theta \mid x,\xi)}$
    * $\int f(x \mid \xi) = \int \pi(x,\theta \mid \xi) d\theta = \int f(x \mid \theta)\pi(\theta \mid \xi) d\theta$

Empirical Bayes method

- Marginal posterior can be computed as

$$\pi(\theta \mid x) \propto \int \pi(\theta \mid x,\xi) f(x \mid \xi)\pi(\xi) d\xi$$

- Instead of approximating this integral, the empirical Bayes method uses

$$\pi(\theta \mid x) \approx \pi(\theta \mid x, \hat{\xi}(x)) = \frac{f(x \mid \theta)\pi(\theta \mid \hat{\xi}(x))}{f(x \mid \hat{\xi}(x))}$$

  where

$$\hat{\xi}(x) = \arg\max_\xi f(x \mid \xi) = \arg\max_\xi \int f(x \mid \theta)\pi(\theta \mid \xi) d\theta$$

Bayesian linear regression

- Model:

$$y = \alpha \mathbf{1} + X_\gamma \beta_\gamma + \varepsilon$$

- g-prior of Zellner:

$$\beta_\gamma \mid \sigma^2 \sim \mathcal{N}\left(\beta_\gamma^0, g\sigma^2 (X_\gamma^T X_\gamma^{-1})\right)$$

  - β_γ^0 is the prior mean. Often $\beta_\gamma^0 = 0$

Laplace approximation

- Laplace approximations are used to approximate integrals of the form

$$\int h(\theta)q(\theta)d\theta$$

  where

  - q is a possibly unnormalized smooth density which is concentrated around its mode $\theta_0 = \arg\max \log q(\theta)$
  - h is an arbitrary smooth function

- Expanding $\log q(\theta)$ into a second-order Taylor series at θ_0:

$$\log q(\theta) \approx \log q(\theta_0) - \frac{1}{2}(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)$$

$$q(\theta) \approx q(\theta_0) \exp\left(-\frac{1}{2}(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)\right)$$

  where $J(\theta)$ is the negative Hessian:

$$J(\theta)_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log q(\theta)$$

- Expanding $h(\theta)$ into a first-order Taylor series at θ_0:

$$h(\theta) \approx h(\theta_0) + \frac{\partial h}{\partial \theta}(\theta_0)^T(\theta - \theta_0)$$

- Laplace approximation is given by

$$\begin{aligned}
\int h(\theta)q(\theta)d\theta &\approx \int \left(h(\theta_0) + \frac{\partial h}{\partial \theta}(\theta_0)^T(\theta - \theta_0)\right) q(\theta_0) \exp\left(-\frac{1}{2}(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)\right) d\theta \\
&= h(\theta_0)q(\theta_0) \int \exp\left(-\frac{1}{2}(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)\right) d\theta \\
&= \int h(\theta)q(\theta)d\theta \approx h(\theta_0)q(\theta_0)(\det J(\theta_0))^{-1/2}(2\pi)^{p/2}
\end{aligned}$$

  - $\theta_0 = \arg\max \log q(\theta)$
  - $J(\theta)$ is the negative Hessian

$$J(\theta)_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log q(\theta)$$

Importance sampling

- Monte Carlo sampling:

$$\mathbb{E}(f(x)) = \int f(x)p(x)dx \approx \frac{1}{n}\sum_i f(x_i)$$

- Importance sampling:

$$\mathbb{E}(f(x)) = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx \frac{1}{n}\sum_i f(x_i)\frac{p(x_i)}{q(x_i)}$$

Hamiltonian Monte Carlo

- Hamiltonian $H(x,u)$:

$$H(x,u) = -\log \pi(x) + \sum_{i=1}^{p} \frac{u_i^2}{2m_i}$$

$$\tilde{\pi}(x,u) \propto \exp\left(-H(x,u)\right)$$

- Physical interpretation:
  - x is the position
  - u is the momentum
  - $-\log \pi(x)$ is the potential energy
  - $\sum_{i=1}^{p} \frac{u_i^2}{2m_i}$ is the kinetic energy
  - $H(x,u)$ is the total energy in the system