S&DS 242/542: Theory of Statistics

Lecture 17: Bayesian Inference I

Bayesian inference

Given data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ from a parametric model, we've discussed two approaches for estimating the parameter θ:

- Method-of-moments
- Maximum likelihood

We developed tools to study the sampling distributions of these estimators for large n and to construct confidence intervals for θ.

All of our discussion has treated θ as an unknown but *non-random* parameter, which describes the distribution of our data. This is called the **frequentist** paradigm of statistical inference.

In this lecture, we introduce the **Bayesian** paradigm, where θ is also modeled as random. Bayesian inference starts with our prior belief about θ, and updates this belief based on observed data.

Prior and posterior distributions

Conditional and marginal distributions

If X and Y have joint PDF or PMF $f_{X,Y}(x, y)$, then the **marginal distribution** of X is given by the PDF or PMF

$$f_X(x) = \int_{\mathcal{Y}} f_{X,Y}(x, y)dy, \qquad f_X(x) = \sum_{y \in \mathcal{Y}} f_{X,Y}(x, y)$$

where $\mathcal{Y}$ is the domain of possible values for $\mathcal{Y}$. This describes the probability distribution of X alone, marginalizing over all possible values of Y.

The **conditional distribution** of Y given $X = x$ is then defined by the PDF or PMF

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

This describes the distribution of Y after observing $X = x$.

Prior distribution and likelihood

In Bayesian inference, the unknown parameter Θ is modeled as a random variable, with a probability distribution

$$\Theta \sim f_\Theta(\theta)$$

called its **prior distribution**. This represents our prior belief about the value of Θ, before observing any data.

The parametric model or likelihood function describing the distribution of our data $\mathbf{X} = (X_1, \ldots, X_n)$ is then interpreted as a conditional distribution of $\mathbf{X}$ given Θ, which we now write as

$$\mathbf{X} \sim f_{\mathbf{X}|\Theta}(\mathbf{x} \mid \theta)$$

(In previous lectures, we wrote this simply as $f(\mathbf{x} \mid \theta)$.)

Marginal distribution of the data

This defines a joint probability distribution over both the parameter Θ and the observed data $\mathbf{X}$, with joint PDF/PMF

$$f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) = f_{\mathbf{X}|\Theta}(\mathbf{x} \mid \theta) f_{\Theta}(\theta)$$

Under this joint distribution, the marginal distribution of Θ is simply the prior $f_{\Theta}(\theta)$. The marginal distribution of the data $\mathbf{X}$ is[1]

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) d\theta = \int f_{\mathbf{X}|\Theta}(\mathbf{x} \mid \theta) f_{\Theta}(\theta) d\theta$$

This represents the probability distribution for our observed data $\mathbf{X}$, if we assume Θ is also random and marginalize over its randomness.

[1] If Θ is discrete, then this integral should be replaced by a sum.

Posterior distribution of the parameter

Bayesian inference is based upon the conditional distribution of the parameter Θ given the data $\mathbf{X}$. This conditional distribution is

$$f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x}) = \frac{f_{\mathbf{X},\Theta}(\mathbf{x}, \theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}|\Theta}(\mathbf{x} \mid \theta) f_{\Theta}(\theta)}{f_{\mathbf{X}}(\mathbf{x})}$$

This is called the **posterior distribution** of Θ: It represents our knowledge about the parameter Θ after observing $\mathbf{X}$.

We often summarize the preceding equation simply as

$$f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x}) \propto f_{\mathbf{X}|\Theta}(\mathbf{x} \mid \theta) f_{\Theta}(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

where the symbol $\propto$ hides the proportionality factor $1/f_{\mathbf{X}}(\mathbf{x})$ which does not depend on θ.

Example: Uniform prior for a Bernoulli proportion

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$. Suppose p is unknown, and we model it as random with prior $P \sim \text{Uniform}(0, 1)$. To derive the posterior:

$$f_P(p) = 1 \quad \text{for } p \in (0,1)$$

$$f_{X_i|P}(x_i|p) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases} = p^{x_i}(1-p)^{1-x_i}$$

$$\Rightarrow f_{\underline{X},P}(x_1, \ldots, x_n, p) = \underbrace{\left(\prod_{i=1}^{n} f_{X_i|P}(x_i|p)\right)}_{\text{likelihood}} \cdot \underbrace{f_P(p)}_{\text{prior}}$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n - \sum_{i=1}^{n} x_i}$$

$$= p^{S}(1-p)^{n-S} \quad \text{where } S = X_1 + \cdots + X_n$$

Example: Uniform prior for a Bernoulli proportion

So the marginal distribution of $\underline{X} = (X_1, \ldots, X_n)$ is

$$f_{\underline{X}}(x_1, \ldots, x_n) = \int_0^1 f_{\underline{X},P}(x_1, \ldots, x_n, p)\, dp$$

$$= \int_0^1 p^s (1-p)^{n-s}\, dp = B(s+1, n-s+1)$$

where $B(\alpha, \beta)$ is the Beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\, dp$$

$$\Rightarrow f_{P|\underline{X}}(p \mid x_1, \ldots, x_n) = \frac{f_{\underline{X},P}(x_1, \ldots, x_n, p)}{f_{\underline{X}}(x)} = \frac{1}{B(s+1, n-s+1)} p^s (1-p)^{n-s}$$

So the posterior distribution is $P \sim \text{Beta}(s+1, n-s+1)$, $s = x_1 + \cdots + x_n$.

Example: Uniform prior for a Bernoulli proportion

We computed explicitly the marginal distribution $f_{\mathbf{X}}(\mathbf{x})$, but this was not actually necessary to derive the posterior distribution. Indeed, omitting factors not depending on p,

$$f_{P|\mathbf{X}}(p \mid \mathbf{x}) \propto f_{\mathbf{X}|P}(\mathbf{x} \mid p) f_P(p) \propto p^s (1-p)^{n-s}$$

This means the PDF of the posterior distribution of P has the form

$$f_{P|\mathbf{X}}(p \mid \mathbf{x}) = \frac{1}{C(\mathbf{x})} p^s (1-p)^{n-s}$$

where the proportionality constant $C(\mathbf{x})$ must be the unique value that makes this PDF integrate to 1 over $p \in (0,1)$. So $C(\mathbf{x})$ is the normalizing constant for the Beta$(s+1, n-s+1)$ distribution, and the posterior distribution must be Beta$(s+1, n-s+1)$.

Example: Beta prior for a Bernoulli proportion

Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p). We model p as random with prior $P \sim \text{Beta}(\alpha, \beta)$. To derive the posterior distribution:

$$f_P(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{over } p \in (0,1)$$

$$f_{P|\underline{X}}(p \mid x_1, \ldots, x_n) \propto f_{\underline{X}|P}(x_1, \ldots, x_n \mid p) \cdot f_P(p)$$

$$\propto p^s (1-p)^{n-s} \cdot p^{\alpha-1}(1-p)^{\beta-1}$$

where $s = x_1 + \cdots + x_n$

$$\propto \underbrace{p^{s+\alpha-1}(1-p)^{n-s+\beta-1}}_{\text{proportional to PDF of Beta}(s+\alpha,\, n-s+\beta)}$$

$\Rightarrow$ Posterior distribution is $P \sim \text{Beta}(s+\alpha,\, n-s+\beta)$, $s = x_1 + \cdots + x_n$

Example: Gamma prior in the Poisson model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$. We model λ as random with prior $\Lambda \sim \text{Gamma}(\alpha, \beta)$. To derive the posterior distribution:

$$f_\Lambda(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad \text{for } \lambda > 0$$

$$f_{X_i|\Lambda}(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$\Rightarrow f_{\Lambda|\underline{X}}(\lambda|x_1,\ldots,x_n) \propto f_{\underline{X}|\Lambda}(x_1,\ldots,x_n|\lambda) \times f_\Lambda(\lambda)$$

$$\propto \left(\prod_{i=1}^n e^{-\lambda}\lambda^{x_i}\right) \times \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \lambda^{S+\alpha-1} e^{-(n+\beta)\lambda} \quad \text{where } S = X_1 + \cdots + X_n$$

So the posterior distribution is $\Lambda \sim \text{Gamma}(S+\alpha, n+\beta)$.

Conjugate priors

A family of priors $f_\Theta(\theta)$ are called **conjugate priors** for a parametric model $f_{\mathbf{X}|\Theta}(\mathbf{x} \mid \theta)$ if the resulting posteriors $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$ have the same parametric form as the priors. So

- Beta(α, β) are conjugate priors for the model Bernoulli(p)
- Gamma(α, β) are conjugate priors for the model Poisson(λ)

Conjugate priors exist for many simple parametric models, and are mathematically/computationally convenient because they lead to posteriors that are common distributions and easy to compute.

This convenience does not necessarily imply desirable statistical properties — for example, conjugate priors are typically light-tailed distributions that bias posterior inferences towards the prior mean. Non-conjugate (in particular heavier-tailed) priors are also often used to yield posterior inferences that are more robust to the choice of the prior, even if this requires more intensive computation.

Example: Normal prior for a normal mean

Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, \frac{1}{\xi})$, where for convenience we parametrize the model by the *precision* $\xi = \frac{1}{\sigma^2}$ instead of the variance σ^2. Suppose ξ is known, θ is unknown with prior $\Theta \sim \mathcal{N}(\mu_{\text{prior}}, \frac{1}{\xi_{\text{prior}}})$.

$$f_{\Theta}(\theta) = \sqrt{\frac{\xi_{prior}}{2\pi}}\, e^{-\frac{\xi_{prior}}{2}(\theta-\mu_{prior})^2}$$

$$f_{X_i|\Theta}(x_i|\theta) = \sqrt{\frac{\xi}{2\pi}}\, e^{-\frac{\xi}{2}(x_i-\theta)^2}$$

$$\Rightarrow f_{\Theta|\underline{X}}(\theta|x_1,\ldots,x_n) \propto f_{\underline{X}|\Theta}(x_1,\ldots,x_n|\theta) \cdot f_{\Theta}(\theta)$$

$$\propto \left(\prod_{i=1}^{n} e^{-\frac{\xi}{2}(x_i-\theta)^2}\right) \times e^{-\frac{\xi_{prior}}{2}(\theta-\mu_{prior})^2}$$

$$\propto \exp\left(-\sum_{i=1}^{n}\frac{\xi}{2}(x_i-\theta)^2 - \frac{\xi_{prior}}{2}(\theta-\mu_{prior})^2\right)$$

Example: Normal prior for a normal mean

$$\propto \exp\left(-\sum_{i=1}^{n}\left(\frac{\xi}{2}\theta^2 - \xi X_i \theta\right) - \left(\frac{\xi_{prior}}{2}\theta^2 - \xi_{prior}\mu_{prior}\theta\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(n\xi + \xi_{prior}\right)\theta^2 + \left(\xi\sum_{i=1}^{n} X_i + \xi_{prior}\mu_{prior}\right)\theta\right)$$

Let $\xi_{post} = n\xi + \xi_{prior}$

$$\mu_{post} = \frac{\xi\sum_{i=1}^{n} X_i + \xi_{prior}\mu_{prior}}{n\xi + \xi_{prior}}$$

$$\Rightarrow f_{\Theta|\underline{X}}(\theta|x_1,\ldots,x_n) \propto \exp\left(-\frac{\xi_{post}}{2}\theta^2 + \xi_{post}\mu_{post}\theta\right)$$

$$\propto \exp\left(-\frac{\xi_{post}}{2}(\theta-\mu_{post})^2\right)$$

$\Rightarrow$ Posterior is $\Theta \sim N\left(\mu_{post}, \frac{1}{\xi_{post}}\right)$

Example: Normal prior for a normal variance

Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, \frac{1}{\xi})$. Suppose θ is known, ξ is unknown with prior $\Xi \sim \text{Gamma}(\alpha, \beta)$.

$$f_{\Xi}(\xi) = \frac{\beta^\alpha}{\Gamma(\alpha)} \xi^{\alpha-1} e^{-\beta\xi}, \quad \xi > 0$$

$$f_{X_i|\Xi}(x_i|\xi) = \sqrt{\frac{\xi}{2\pi}}\, e^{-\frac{\xi}{2}(x_i-\theta)^2}$$

$$\Rightarrow f_{\Xi|\underline{X}}(\xi|x_1,\ldots,x_n) \propto f_{\underline{X}|\Xi}(x_1,\ldots,x_n|\xi) \times f_{\Xi}(\xi)$$

$$\propto \left(\prod_{i=1}^{n} \sqrt{\xi}\, e^{-\frac{\xi}{2}(x_i-\theta)^2}\right) \times \xi^{\alpha-1} e^{-\beta\xi}$$

$$\propto \xi^{\alpha+\frac{n}{2}-1} e^{-\xi\left(\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2+\beta\right)}$$

$$\Rightarrow \text{Posterior is } \Xi \sim \text{Gamma}\left(\alpha+\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2+\beta\right)$$

Point estimates and credible intervals

Bayesian point estimates and credible intervals

In many applications, we may want a single estimate $\hat{\theta}$ and an interval that quantifies our uncertainty about θ, instead of a full posterior distribution.

The **posterior mean** and **posterior mode** are the mean and mode of the posterior distribution of Θ, and either may be used as a Bayesian estimate $\hat{\theta}$ for θ.

A **Bayesian credible interval** with coverage $1 - \alpha$ is an interval $I(\mathbf{X})$ that contains Θ with posterior probability $1 - \alpha$:

$$\mathbb{P}[\,\Theta \in I(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\,] = 1 - \alpha$$

where the probability is over Θ, instead of $\mathbf{X}$ as in the frequentist confidence interval. A common choice is the interval from the lower-$\alpha/2$ to upper-$\alpha/2$ point of the posterior distribution of Θ.

Example: Beta prior for a Bernoulli proportion

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$, with prior $P \sim \text{Beta}(\alpha, \beta)$. Recall that the posterior distribution is

$$P \sim \text{Beta}(S + \alpha,\ n - S + \beta), \qquad S = X_1 + \ldots + X_n$$

The posterior mean is

$$\hat{p} = \frac{S + \alpha}{n + \alpha + \beta}$$

This is different from the MLE/MoM-estimator $\bar{X} = \frac{S}{n}$, and is biased towards the prior mean.

One interpretation is that $\hat{p}$ is the sample mean as if we had observed — *a priori* — α heads and β tails before seeing our data. This gives an interpretation also for the Beta(α, β) prior.

Example: Beta prior for a Bernoulli proportion

An alternative interpretation is that $\hat{p}$ is a weighted average of the sample mean and prior mean,

$$\hat{p} = \frac{S+\alpha}{n+\alpha+\beta} = \underbrace{\frac{n}{n+\alpha+\beta}}_{\text{sample weight}} \cdot \underbrace{\frac{S}{n}}_{\text{sample mean}} + \underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{\text{prior weight}} \cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}}$$

For any fixed prior Beta(α, β), as n increases, the sample weight increases to 1 and prior weight decreases to 0.

The interval from the lower-0.05 point to the upper-0.05 point of the Beta$(S+\alpha, n-S+\beta)$ posterior distribution forms a 90% Bayesian credible interval for p.