Statistics for Astronomers: Lecture 18, 2021.01.06

Prof. Sundar Srinivasan

IRyA/UNAM

Review: Bayesian inference

Prior selection: choose a prior as long as it isn't a delta function.
Prior-dominated vs. evidence/data-dominated posterior.
Bayesian point estimates – maximum a posteriori (MAP) estimate.
Bayesian interval estimates – credible intervals; the highest posterior density interval.
Informative and non-informative priors. Improper priors.

References

Bayesian Data Analysis, Third Edition – A. Gelman, et al.

Prior and posterior predictive distributions

Prior predictive distribution: Before the experiment, given the prior probability distribution $\pi(\theta)$ of the unknown parameter, what is the probability distribution of expected data values?

Discrete: $P(x) = \sum_{i=1}^{N} P(x|\theta_i)\ \pi(\theta_i)$ continuous: $p(x) = \int p(x|\theta)\ \pi(\theta)\ d\theta.$

Prior and posterior predictive distributions

Prior predictive distribution: Before the experiment, given the prior probability distribution $\pi(\theta)$ of the unknown parameter, what is the probability distribution of expected data values?

Discrete: $P(x) = \sum_{i=1}^{N} P(x|\theta_i)\ \pi(\theta_i)$ continuous: $p(x) = \int p(x|\theta)\ \pi(\theta)\ d\theta$.

Posterior probability distribution: Given the data, what is the probability distribution of the unknown parameter?

Discrete: $P(\theta_i|x) = \dfrac{P(x|\theta_i)\ \pi(\theta_i)}{\sum_{i=1}^{N} P(x|\theta_i)\ \pi(\theta_i)}$ Continuous: $p(\theta|x) = \dfrac{p(x|\theta)\ \pi(\theta)}{\int p(x|\theta)\ \pi(\theta)\ d\theta}$.

Prior and posterior predictive distributions

Prior predictive distribution: Before the experiment, given the prior probability distribution $\pi(\theta)$ of the unknown parameter, what is the probability distribution of expected data values?

Discrete: $P(x) = \sum_{i=1}^{N} P(x|\theta_i)\ \pi(\theta_i)$ continuous: $p(x) = \int p(x|\theta)\ \pi(\theta)\ d\theta$.

Posterior probability distribution: Given the data, what is the probability distribution of the unknown parameter?

Discrete: $P(\theta_i|x) = \dfrac{P(x|\theta_i)\ \pi(\theta_i)}{\sum_{i=1}^{N} P(x|\theta_i)\ \pi(\theta_i)}$ Continuous: $p(\theta|x) = \dfrac{p(x|\theta)\ \pi(\theta)}{\int p(x|\theta)\ \pi(\theta)\ d\theta}$.

Posterior predictive distribution: Given the posterior probability distribution, what is the probability distribution of data values in a future experiment?

Discrete: $P(\widetilde{x}|x) = \sum_{i=1}^{N} P(\widetilde{x}|\theta_i, x)\ P(\theta_i|x) = \sum_{i=1}^{N} P(\widetilde{x}|\theta_i)\ P(\theta_i|x)$ (given θ, $\widetilde{x} \perp x$).

Continuous: $p(\widetilde{x}|x) = \int p(\widetilde{x}|\theta, x)\ p(\theta|x)\ d\theta = \int p(\widetilde{x}|\theta)\ p(\theta|x)\ d\theta$

Prior and posterior predictive distributions: example

A fair coin is tossed once. If the outcome is H, a red light is turned on. If not, the coin is tossed again. If the outcome is H, the red light turns on. If not, a blue light turns on.
Given that a red light turned on, (1) what is the posterior distribution for outcomes of the first toss? (2) what are the prior and posterior predictive distributions for the observations?

Prior and posterior predictive distributions: example

A fair coin is tossed once. If the outcome is H, a red light is turned on. If not, the coin is tossed again. If the outcome is H, the red light turns on. If not, a blue light turns on.
Given that a red light turned on, (1) what is the posterior distribution for outcomes of the first toss? (2) what are the prior and posterior predictive distributions for the observations?

Unknown "parameter": outcome of first flip, $t \in \{\mathrm{H}, \mathrm{T}\}$.

Prior distribution for t: $P(t = \mathrm{H}) = P(t = \mathrm{T}) = 1/2$.

Prior and posterior predictive distributions: example

A fair coin is tossed once. If the outcome is H, a red light is turned on. If not, the coin is tossed again. If the outcome is H, the red light turns on. If not, a blue light turns on.
Given that a red light turned on, (1) what is the posterior distribution for outcomes of the first toss? (2) what are the prior and posterior predictive distributions for the observations?

Unknown "parameter": outcome of first flip, $t \in \{\mathrm{H}, \mathrm{T}\}$.

Prior distribution for t: $P(t = \mathrm{H}) = P(t = \mathrm{T}) = 1/2$.

Prior predictive distribution for data:

$P(\mathrm{red}) = P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H}) + P(\mathrm{red}|t = \mathrm{T})P(t = \mathrm{T}) = 1 \cdot 1/2 + 1/2 \cdot 1/2 = 3/4$; $P(\mathrm{blue}) = 1/4$.

Prior and posterior predictive distributions: example

A fair coin is tossed once. If the outcome is H, a red light is turned on. If not, the coin is tossed again. If the outcome is H, the red light turns on. If not, a blue light turns on.
Given that a red light turned on, (1) what is the posterior distribution for outcomes of the first toss? (2) what are the prior and posterior predictive distributions for the observations?

Unknown "parameter": outcome of first flip, $t \in \{\mathrm{H}, \mathrm{T}\}$.

Prior distribution for t: $P(t = \mathrm{H}) = P(t = \mathrm{T}) = 1/2$.

Prior predictive distribution for data:

$P(\mathrm{red}) = P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H}) + P(\mathrm{red}|t = \mathrm{T})P(t = \mathrm{T}) = 1 \cdot 1/2 + 1/2 \cdot 1/2 = 3/4; P(\mathrm{blue}) = 1/4.$

Likelihood that red light turns on: $P(\mathrm{red}|t = \mathrm{H}) = 1, P(\mathrm{red}|t = \mathrm{T}) = 1/2.$

Posterior distribution for t:

$$P(t = \mathrm{H}|\mathrm{red}) = \frac{P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H})}{P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H}) + P(\mathrm{red}|t = \mathrm{T})P(t = \mathrm{T})} = \frac{1 \cdot 1/2}{1 \cdot 1/2 + 1/2 \cdot 1/2} = 2/3.$$

$P(t = \mathrm{T}|\mathrm{red}) = 1/3.$

Prior and posterior predictive distributions: example

A fair coin is tossed once. If the outcome is H, a red light is turned on. If not, the coin is tossed again. If the outcome is H, the red light turns on. If not, a blue light turns on.
Given that a red light turned on, (1) what is the posterior distribution for outcomes of the first toss? (2) what are the prior and posterior predictive distributions for the observations?

Unknown "parameter": outcome of first flip, $t \in \{\mathrm{H}, \mathrm{T}\}$.

Prior distribution for t: $P(t = \mathrm{H}) = P(t = \mathrm{T}) = 1/2$.

Prior predictive distribution for data:

$P(\mathrm{red}) = P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H}) + P(\mathrm{red}|t = \mathrm{T})P(t = \mathrm{T}) = 1 \cdot 1/2 + 1/2 \cdot 1/2 = 3/4$; $P(\mathrm{blue}) = 1/4$.

Likelihood that red light turns on: $P(\mathrm{red}|t = \mathrm{H}) = 1, P(\mathrm{red}|t = \mathrm{T}) = 1/2$.

Posterior distribution for t:

$$P(t = \mathrm{H}|\mathrm{red}) = \frac{P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H})}{P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H}) + P(\mathrm{red}|t = \mathrm{T})P(t = \mathrm{T})} = \frac{1 \cdot 1/2}{1 \cdot 1/2 + 1/2 \cdot 1/2} = 2/3.$$

$P(t = \mathrm{T}|\mathrm{red}) = 1/3$.

Posterior predictive distribution for future data:

$P(\mathrm{red}|\mathrm{red}) = P(\mathrm{red}|t = \mathrm{H})P(t = \mathrm{H}|\mathrm{red}) + P(\mathrm{red}|t = \mathrm{T})P(t = \mathrm{T}|\mathrm{red}) = 1 \cdot 2/3 + 1/2 \cdot 1/3 = 5/6$.

$P(\mathrm{blue}|\mathrm{red}) = 1/6$.

The Jeffreys Prior

Recall: The MLE of a parameter θ is $\hat{\theta}_{\rm MLE}$ such that $\dfrac{\partial \ln \mathscr{L}}{\partial\theta} = 0$ at $\theta = \hat{\theta}_{\rm MLE}$.

Cramér-Rao Bound: $\mathrm{Var}[\hat{\theta}_{\rm MLE}] \geq \mathcal{I}(\theta)^{-1}$, where $\mathcal{I}(\theta)$ is the Fisher Information.

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial\theta}\right)^2\right] = \text{(under some conditions)} = -\mathbb{E}\left[\frac{\partial^2 \ln \mathscr{L}}{\partial\theta^2}\right]$$

The Jeffreys Prior

Recall: The MLE of a parameter θ is $\hat{\theta}_{\rm MLE}$ such that $\dfrac{\partial \ln \mathscr{L}}{\partial \theta} = 0$ at $\theta = \hat{\theta}_{\rm MLE}$.

Cramér-Rao Bound: $\mathrm{Var}[\hat{\theta}_{\rm MLE}] \geq \mathcal{I}(\theta)^{-1}$, where $\mathcal{I}(\theta)$ is the Fisher Information.

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right] = \text{(under some conditions)} = -\mathbb{E}\left[\frac{\partial^2 \ln \mathscr{L}}{\partial \theta^2}\right]$$

Jeffreys Prior: a non-informative prior that is also invariant over transformation of the parameter.

$\pi_J(\theta) \propto \sqrt{\mathcal{I}(\theta)}$. For multidimensional case, $\pi_J(\vec{\theta}) \propto \sqrt{\mathrm{Det}\ \boldsymbol{\mathcal{I}}(\vec{\theta})}$.

The Jeffreys Prior

Recall: The MLE of a parameter θ is $\hat{\theta}_{\rm MLE}$ such that $\dfrac{\partial \ln \mathscr{L}}{\partial \theta} = 0$ at $\theta = \hat{\theta}_{\rm MLE}$.

Cramér-Rao Bound: $\mathrm{Var}[\hat{\theta}_{\rm MLE}] \geq \mathcal{I}(\theta)^{-1}$, where $\mathcal{I}(\theta)$ is the Fisher Information.

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right] = \text{(under some conditions)} = -\mathbb{E}\left[\frac{\partial^2 \ln \mathscr{L}}{\partial \theta^2}\right]$$

Jeffreys Prior: a non-informative prior that is also invariant over transformation of the parameter.

$$\pi_J(\theta) \propto \sqrt{\mathcal{I}(\theta)}.\ \text{For multidimensional case, } \pi_J(\vec{\theta}) \propto \sqrt{\mathrm{Det}\ \boldsymbol{\mathcal{I}}(\vec{\theta})}.$$

Invariance: If ψ is a function of θ (*e.g.*, $\theta = P(\mathrm{Head})$ and $\psi = P(\mathrm{Tail}) = 1 - \theta$),

$$\pi_J(\psi) = \pi_J(\theta)\left|\frac{d\theta}{d\psi}\right| \propto \sqrt{\mathcal{I}(\theta)\left(\frac{d\theta}{d\psi}\right)^2} = \sqrt{\mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right]\left(\frac{d\theta}{d\psi}\right)^2} = \sqrt{\mathbb{E}\left[\left(\frac{d\theta}{d\psi}\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right]}$$

$$= \sqrt{\mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \psi}\right)^2\right]} = \sqrt{\mathcal{I}(\psi)}.$$

Jeffreys prior example: coin toss (Bernoulli trial)

Let $P(\text{success}) = \theta$. We perform one coin toss and obtain a value $X = x$.

Jeffreys prior example: coin toss (Bernoulli trial)

Let $P(\text{success}) = \theta$. We perform one coin toss and obtain a value $X = x$.

Likelihood associated with this observation: $\mathscr{L}(\theta) \propto \theta^x (1-\theta)^{1-x} = \mathrm{Beta}(x+1, 2-x)$

Jeffreys prior example: coin toss (Bernoulli trial)

Let $P(\text{success}) = \theta$. We perform one coin toss and obtain a value $X = x$.

Likelihood associated with this observation: $\mathscr{L}(\theta) \propto \theta^x (1-\theta)^{1-x} = \mathrm{Beta}(x+1, 2-x)$

$$\implies \ln \mathscr{L} = x \ln \theta + (1-x) \ln (1-\theta) \implies \frac{\partial \ln \mathscr{L}}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

Recall: for Bernoulli distribution, $\mathbb{E}[X] = \theta, \mathrm{Var}[X] \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \theta(1-\theta)$.

Jeffreys prior example: coin toss (Bernoulli trial)

Let $P(\text{success}) = \theta$. We perform one coin toss and obtain a value $X = x$.

Likelihood associated with this observation: $\mathscr{L}(\theta) \propto \theta^x (1-\theta)^{1-x} = \text{Beta}(x+1, 2-x)$

$$\implies \ln \mathscr{L} = x \ln \theta + (1-x) \ln (1-\theta) \implies \frac{\partial \ln \mathscr{L}}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

Recall: for Bernoulli distribution, $\mathbb{E}[X] = \theta, \text{Var}[X] \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \theta(1-\theta)$.

$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right] = \frac{1}{\theta(1-\theta)} \implies$ prior: $\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \text{Beta}(a, b)$ with $a = b = 1/2$.

prior mean: $\frac{a}{a+b} = \frac{1/2}{1/2 + 1/2} = 0.5$ as expected.

Jeffreys prior example: coin toss (Bernoulli trial)

Let $P(\mathrm{success}) = \theta$. We perform one coin toss and obtain a value $X = x$.

Likelihood associated with this observation: $\mathscr{L}(\theta) \propto \theta^x (1-\theta)^{1-x} = \mathrm{Beta}(x+1, 2-x)$

$$\implies \ln \mathscr{L} = x \ln \theta + (1-x) \ln (1-\theta) \implies \frac{\partial \ln \mathscr{L}}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

Recall: for Bernoulli distribution, $\mathbb{E}[X] = \theta, \mathrm{Var}[X] \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \theta(1-\theta)$.

$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right] = \frac{1}{\theta(1-\theta)} \implies$ prior: $\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \mathrm{Beta}(a, b)$ with $a = b = 1/2$.

$$\text{prior mean: } \frac{a}{a+b} = \frac{1/2}{1/2+1/2} = 0.5 \text{ as expected.}$$

Posterior: $p(\theta|\mathrm{data}) \propto \mathscr{L}(\theta)\, \pi_J(\theta) = \mathrm{Beta}(x+1, 2-x) \times \mathrm{Beta}\left(\frac{1}{2}, \frac{1}{2}\right) = \mathrm{Beta}\left(x + \frac{1}{2}, \frac{3}{2} - x\right)$.

Posterior mean: $\frac{1}{2}\left(x + \frac{1}{2}\right) = \frac{1}{2}(\mathrm{sample\ mean} + \mathrm{prior\ mean})$. Effective sample size: 2.

Jeffreys prior example: coin toss (Bernoulli trial)

Let $P(\text{success}) = \theta$. We perform one coin toss and obtain a value $X = x$.

Likelihood associated with this observation: $\mathscr{L}(\theta) \propto \theta^x (1-\theta)^{1-x} = \text{Beta}(x+1, 2-x)$

$$\implies \ln \mathscr{L} = x \ln \theta + (1-x) \ln(1-\theta) \implies \frac{\partial \ln \mathscr{L}}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

Recall: for Bernoulli distribution, $\mathbb{E}[X] = \theta, \text{Var}[X] \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \theta(1-\theta)$.

$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial \theta}\right)^2\right] = \frac{1}{\theta(1-\theta)} \implies$ prior: $\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \text{Beta}(a, b)$ with $a = b = 1/2$.

prior mean: $\frac{a}{a+b} = \frac{1/2}{1/2 + 1/2} = 0.5$ as expected.

Posterior: $p(\theta|\text{data}) \propto \mathscr{L}(\theta)\, \pi_J(\theta) = \text{Beta}(x+1, 2-x) \times \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right) = \text{Beta}\left(x + \frac{1}{2}, \frac{3}{2} - x\right)$.

Posterior mean: $\frac{1}{2}\left(x + \frac{1}{2}\right) = \frac{1}{2}(\text{sample mean} + \text{prior mean})$. Effective sample size: 2.

Note that the posterior and prior are both Beta distributions. In such a case, we say that the Beta distribution is the conjugate prior to a Bernoulli likelihood. The Beta distribution is also conjugate to binomial likelihoods (cf. Lecture 17).

Jeffreys prior for a Poisson distribution

Poisson problem with unknown rate parameter λ and observation $X = x$ (say).

Recall: $P(X = x) = \dfrac{\lambda^x \; e^{-\lambda}}{x!}; \mathbb{E}[X] = \lambda; \mathrm{Var}[X] = \lambda.$

Likelihood: $\mathscr{L}(\lambda) = \dfrac{\lambda^x \; e^{-\lambda}}{x!}$

Jeffreys prior: $\dfrac{\partial \ln \mathscr{L}(\lambda)}{\partial \lambda} =?$ $\quad \mathcal{I}(\lambda) = \mathbb{E}\left[\left(\dfrac{\partial \ln \mathscr{L}(\lambda)}{\partial \lambda}\right)^2\right] =?$ $\quad \pi_J(\lambda) = \sqrt{\mathcal{I}(\lambda)} =?$

Prior predictive distribution?

Posterior?

Posterior predictive distribution?

Jeffreys priors for a univariate normal distribution

Homework.

More on priors

See ▸ Jaynes (1968) for a good discussion of the applicability of this procedure to problems in fundamental physics.