STA732
Statistical Inference

Lecture 10: Bayes pros and cons

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

Recap from Lecture 09

1. Defined Bayes risk, Bayes estimator
2. Bayes estimators are usually biased and usually admissible

Whether to believe the parameter is random is rather a philosophical choice that we have to make. Think about the Bayesian model $X \sim p_\theta(x), \theta \sim \Lambda$.

- If X's are i.i.d., we get to see the empirical distribution after multiple draws, can test the goodness of fit
- But there is only one draw of θ, it is not even directly observed

Choosing prior is the biggest issue in Bayesian estimation!!!

Goal of Lecture 10

1. Conjugate priors
2. Where do priors come from?
3. Pros and cons

Chap. 4.1 in Lehmann and Casella

Reading materials for a fair assessment of pros and cons of Bayes

It is helpful to read multiple opinions

- Chap. 4.1 in Lehmann and Casella
- Chap. 4 in Statistical Decision Theory and Bayesian Analysis. Berger, 1985
- From Andrew Gelman,
    - Objections to Bayesian statistics, Gelman, 2008
    - Rejoinder, Gelman, 2008

Conjugate priors

Conjugate priors

Def. conjugate prior

If the posterior is from the same family as the prior, we say that the prior is conjugate to the likelihood.

Easy to build conjugate prior for exponential family

- Likelihood in s-parameter exponential family

$$X_i \mid \eta \overset{\text{i.i.d.}}{\sim} p_\eta(x) = \exp(\eta^\top T(x) - A(\eta))h(x), \quad \eta \in \Xi \subseteq \mathbb{R}^s$$

- Prior: define $s+1$-parameter exponential family

$$\lambda_{k\mu,k}(\eta) = \exp\left(k\mu^\top \eta - kA(\eta) - B(k\mu, k)\right)\right) \lambda_0(\eta)$$

with sufficient statistics $\begin{pmatrix} \eta \\ -A(\eta) \end{pmatrix} \in \mathbb{R}^{s+1}$, natural parameters $\begin{pmatrix} k\mu \\ k \end{pmatrix}$

Prove that the posterior is

$$\lambda(\eta \mid X_1, \dots, X_n) = \lambda_{k\mu+n\bar{T},k+n}(\eta)$$

where $\bar{T}(X) = \frac{1}{n}\sum_{i=1}^n T(X_i)$.

Interpretation of the posterior

Two ways to see the posterior

- Take prior $\lambda_{k\mu,k}$, observe average sufficient statistics $\bar{T}$ on sample size n
- Take prior λ_0, observe average sufficient statistics
  - μ on sample size k (pseudo-data)
  - $\bar{T}$ on sample size n

This gives one way to construct prior: from previous data experience

Examples of conjugate priors - Beta-Binomial

Likelihood

$$X_i \mid \theta \sim \text{Binomial}(n, \theta)$$

$$p_\theta(x) = \theta^x (1-\theta)^x \binom{n}{x}$$

Prior

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$\lambda(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Normal-Normal

Likelihood

$$X_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$$

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta - x)^2}{2\sigma^2}\right)$$

σ^2 is fixed

Prior

$$\theta \sim \mathcal{N}(\mu, \tau^2)$$

$$\lambda(x) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\theta - \mu)^2}{2\tau^2}\right)$$

Poisson-Gamma

Likelihood

$$
\begin{aligned}
X_i \mid \theta &\sim \text{Poisson}(\theta) \\
&= \frac{\theta^x e^{-\theta}}{x!}
\end{aligned}
$$

Prior

$$
\begin{aligned}
\theta &\sim \text{Gamma}(\alpha, \beta) \\
&= \frac{1}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \beta^{\alpha}
\end{aligned}
$$

Where do priors come from?

Ideas to construct priors

1. Prior data experience
2. Subjective prior
3. “Objective” prior
4. Convenience prior

1. Prior experience

Prior is estimated from previous experience, by assuming that we are encoutering similar problems

- can be estimated from previous data, leading to empirical Bayes
- can test the validity of prior since we have multiple draws

This is a relatively non-controversial way of choosing prior

2. Subjective prior

Prior that arises from purely subjective assessment

Pros

- can bring outside knowledge to modeling

Cons

- scientists find subjectivitiy offputting because validation becomes harder: what if two people come up with two different priors?
- in general, difficulty to mathematically formalize priors about the joint distribution of a high dimension parameter, like in $\mathbb{R}^{10}$

3. “Objective” prior

An objective prior

- expresses vague or general information about a variable (like it is positive or it is bounded)
- then applies the principle of indifference, which assigns equal probability to all possibilities

Example: Gaussian mean estimation with flat prior

Suppose $X_i \mid \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$. Determine the posterior mean under flat prior

flat prior can be obtained as a limit of $\Theta \sim \mathcal{N}(0, \tau^2)$, $\tau^2 \to \infty$

Jeffreys prior

An objective prior that requires additional invariance under a change of variable ("invariant" under reparametrization)

$$p(\theta) \propto \sqrt{\det(I(\theta))}$$

Jeffreys prior for Binomial(n, θ)

Jeffreys prior in Gaussian sequence model

$X \sim \mathcal{N}(\mu, \mathbb{I}_d), \mu \in \mathbb{R}^d$

- Determine Jeffreys prior
- Posterior mean for estimating μ
- Posterior mean for estimating $\|\mu\|_2^2$

4. Convenience prior

Prior chosen for mathematical and computational convenience

Examples

- Conjugate priors
- Choose to use Laplace prior instead of spike-and-slab lasso prior for sparse problems for computational convenience

Pros

Pros

1. Bayes estimator is defined straightforwardly
2. Bayes optimal is appealing
3. Detailed output

1. Straightforward definition of estimator

$$\delta_\Lambda(x) = \arg\min_\delta \int L(\theta, \delta(x)) \lambda(\theta \mid X = x) d\theta$$

Bayes estimators are usually easier to find than minimax estimators

Finding Bayes estimator can also be reduced to pure computation

1. Sample from the posterior $\lambda(\theta \mid X = x)$
2. Optimize over θ

Separation of modeling and computation is traditionally known as a big advantage!

- Can use complex models
- Can use loss L that we care about

2. Bayes optimal

- Bayes estimator is by definition Bayes optimal: there exists a prior such that the estimator is optimal in the average risk sense
- Bayes estimator is usually admissible

3. Detailed output

$$\delta_\Lambda(x) = \arg\min_{\delta} \int L(\theta, \delta(x))\lambda(\theta \mid X = x)d\theta$$

- Get the joint distribution over all parameters
- Being able to sample from the posterior, then one can generate the estimates of any $g(\theta)$

Cons

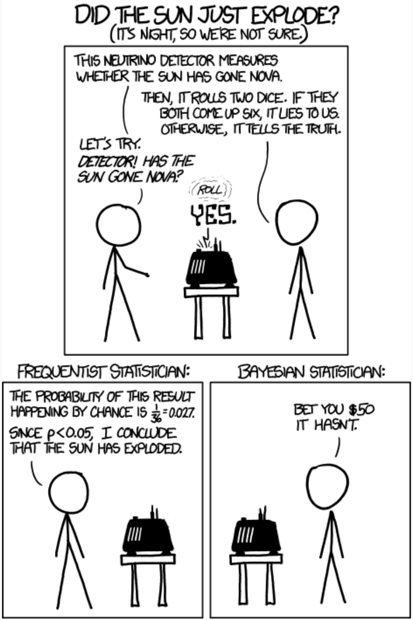
Cons

1. Difficulty in choosing prior Λ
2. Specifying model in full detail might be difficult

1. Difficulty in choosing prior Λ

- Hard to choose Λ, especially in high dimension
- A frequentist will always doubt whether the prior Λ gets any mass near the true θ

How to offend a Bayesian



also see the discussion here

2. Specifying model in full detail might be difficult

The flipside of having detailed output is the requirement to specify the prior

Example

Nonparametric estimation of $g(P) = \mathbb{E}_P X_i, X_i \overset{\text{i.i.d.}}{\sim} P$.

- $\bar{X}$ is UMVU, and a natural choice of estimator
- How to find a Bayes estimator? Must specify a prior over all distributions on $\mathbb{R}$ first!

What is next?

- Empirical Bayes, also James-Stein estimator
- Hierarchical Bayes

Thank you