Parametric Statistics

Bayes Estimators, MLE estimation

Sofia Triantafillou

sof.triantafillou@gmail.com

University of Crete
Department of Mathematics and Applied Mathematics

November 1, 2023

Lecture Summary

Recap

- Statistical Inferences draws conclusions about unknown parameters using data.
- Two schools: Bayesian and Frequentist.
- Pick a prior distribution.
- Compute the likelihood.
- Use Bayes' theorem to compute the posterior distribution:

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

- Perform Sensitivity Analysis.
- Summarize the posterior distribution.

Another Example of Bayesian estimation - Normal distribution

- Let $X_1, \ldots, X_n$ be a random sample from $N\left(\theta, \sigma^2\right)$ where σ^2 is known
- Let the prior distribution of θ be $N\left(\mu_0, \nu_0^2\right)$ where μ_0 and ν_0^2 are known.
- Show that the posterior distribution $p(\theta \mid \mathbf{x})$ is $N\left(\mu_1, \nu_1^2\right)$ where

$$\mu_1 = \frac{\sigma^2\mu_0 + n\nu_0^2\overline{\mathbf{x}}_n}{\sigma^2 + n\nu_0^2} \quad \text{and} \quad \nu_1^2 = \frac{\sigma^2\nu_0^2}{\sigma^2 + n\nu_0^2}$$

The posterior mean is a linear combination of the prior mean μ_0 and the observed sample mean.

Conjugate priors

Likelihood	**Prior**	**Posterior**
Bernoulli(p)	Beta(α, β)	Beta$(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i)$
Bin(N, p)	Beta(α, β)	Beta$(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i)$
Pois(λ)	Gamma(α, β)	Gamma$(\alpha + \sum_{i=1}^{n} x_i, \beta + n)$
Expo(λ)	Gamma(α, β)	Gamma$(\alpha + n, \beta + \sum_{i=1}^{n} x_i)$
$\mathcal{N}(\theta, \sigma^2)$, known σ^2	$\mathcal{N}(\mu_0, \nu_0)$	$\mathcal{N}(\frac{\sigma^2 \mu_0 + n\nu_0 \bar{x}_n}{\sigma^2 + n\nu_0}, \frac{\sigma^2 \nu_0^2}{\sigma^2 + n\nu_0^2})$

Improper priors

- Improper Prior: A "pdf" $p(\theta)$ where $\int p(\theta) d\theta = \infty$
- Used to try to put more emphasis on data and down play the prior
- Used when there is little or no prior information about θ.
- Not clear that an improper prior is necessarily "non-informative".
- Danger: We always need to check that the posterior pdf is proper! (Integrates to 1)

Improper prior for Normal Distribution

- $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

- $\xi(\mu) = 1$

Improper prior for Normal Distribution

- $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

- $\xi(\mu) = 1$
- $f(\mu | x_1, \ldots, x_n) \sim \mathcal{N}(\overline{X}_n, 1/n)$

Point Estimator

- Often people wish to estimate the unknown parameter θ with a single number.

Suppose our observable data $X_1, \ldots, X_n$ is i.i.d. $f(x \mid \theta), \theta \in \Omega \subset \mathbb{R}$.

Estimator

A real valued function $\delta\left(X_1, \ldots, X_n\right)$ is an **estimator** of θ.

Estimate

Once you observe $x_1, \ldots, x_n$, $\hat{\theta} : \delta\left(x_1, \ldots, x_n\right)$, i.e. estimator evaluated at the observed values is the **estimate** for θ

- An estimator is a statistic and a random variable.
- An estimate is a number.

Loss Function

Loss function:

A real valued function $L(\theta, a)$ where $\theta \in \Omega$ and $a \in \mathbb{R}$.

$L(\theta, a) =$ what we loose by using a as an estimate when θ is the true value of the parameter.

Example Loss Functions

- Squared error loss function: $L(\theta, a) = (\theta - a)^2$
- Absolute error loss function: $L(\theta, a) = |\theta - a|$
- Zero-one loss: $L(\theta, a) = 0$, if $\theta = a$, 1, otherwise.

Expected Loss

$E[L(\theta, a)] = \int_\Omega L(\theta, a)\xi(\theta)d\theta$

Bayes Estimator

Idea

Choose an estimator $\delta(\mathbf{X})$ so that we minimize the expected loss.

Bayes Estimator/Estimate.

Let $L(\theta, a)$ be a loss function. For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, let $\delta^*(\boldsymbol{x})$ be a value of a such that $E[L(\theta, a) \mid \boldsymbol{x}]$ is minimized. Then δ^* is called a Bayes estimator of θ. Once $\boldsymbol{X} = \boldsymbol{x}$ is observed, $\delta^*(\boldsymbol{x})$ is called a Bayes estimate of θ.
Another way to describe a Bayes estimator δ^* is to note that, for each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, the value $\bar{\delta}^*(\boldsymbol{x})$ is chosen so that

$$E\left[L\left(\theta, \delta^*(\boldsymbol{x})\right) \mid \boldsymbol{x}\right] = \min_{\text{All } a} E[L(\theta, a) \mid \boldsymbol{x}].$$

Bayes Estimators

Bayes Estimator for Squared Error Loss

Let θ be a real-valued parameter. Suppose that the squared error loss function is used and that the posterior mean of $\theta, E(\theta \mid \boldsymbol{X})$, is finite. Then, a Bayes estimator of θ is $\delta^*(\boldsymbol{X}) = E(\theta \mid \boldsymbol{X})$.

Bayes Estimator for Absolute Error Loss

When the absolute error loss function is used, a Bayes estimator of a real valued parameter is $\delta^*(\boldsymbol{X})$ equal to a median of the posterior distribution of θ.

Consistency

Consistent estimators

A sequence of estimators that converges in probability to the unknown value of the parameter being estimated, as $n \to \infty$, is called a consistent sequence of estimators.

Example

Consider the Bernoulli Distribution with true unknown parameter θ. The Bayes Estimator for Squared Error Loss is the mean of the posterior,

$$\delta^*(\boldsymbol{X}) = \frac{\alpha + \sum_{i=1}^{n} X_i}{\alpha + \beta + n}$$

Consistency

Consistent estimators

A sequence of estimators that converges in probability to the unknown value of the parameter being estimated, as $n \to \infty$, is called a consistent sequence of estimators.

Example

Consider the Bernoulli Distribution with true unknown parameter θ. The Bayes Estimator for Squared Error Loss is the mean of the posterior,

$$\delta^*(\boldsymbol{X}) = \frac{\alpha + \sum_{i=1}^{n} X_i}{\alpha + \beta + n} \xrightarrow{p} \theta$$

Under fairly general conditions and for a wide range of loss functions, the Bayes estimator is consistent.

Recap

- Bayesian estimation computes the posterior distribution for parameter(s) θ.
- Steps to Bayesian Estimation: Define prior, compute likelihood, compute posterior.
- You can then select a single point as the estimate, e.g., the posterior mean/mode/median.
- The process (function) of finding a point estimate is called an estimator.
- The value of the estimator for a given set of observations is the estimate.
- Bayes estimators minimize a loss function for every possible set of data.

Likelihood

- What if you are a frequentist, and do not want to use prior distributions?
- When the joint pf $f_n(\mathbf{x} \mid \theta)$ is regarded as a function of θ for given observations $x_1, \ldots, x_n$ it is called the likelihood function.

Maximum Likelihood Estimator/Estimate.

(MLE): For each possible observed vector $\boldsymbol{x}$, let $\delta(x) \in \Omega$ denote a value of $\theta \in \Omega$ for which the likelihood function $f_n(\boldsymbol{x} \mid \theta)$ is a maximum, and let $\hat{\theta} = \delta(\boldsymbol{X})$ be the estimator of θ defined in this way. The estimator $\hat{\theta}$ is called a maximum likelihood estimator of θ. After $\boldsymbol{X} = \boldsymbol{x}$ is observed, the value $\delta(\boldsymbol{x})$ is called a maximum likelihood estimate of θ.

Maximum Likelihood Estimator

- Given $\mathbf{X} = \mathbf{x}$, the maximum likelihood estimate (MLE) will be a function of $\mathbf{x}$. Notation: $\hat{\theta} = \delta(\mathbf{X})$
- Potentially confusing notation: Sometimes $\hat{\theta}$ is used for both the estimator and the estimate.
- Note: The MLE is required to be in the parameter space Ω.
- Often it is easier to maximize the log-likelihood $L(\theta) = \log f_n(\mathbf{x} \mid \theta)$

Example

Assume $X_i \sim Expo(\lambda)$, and we observe $x_1 = 1.5, x_2 = 2.1, x_3 = 3$

MLE

- We pick the parameter that makes the observed data most likely.
- But: The likelihood is not a pdf/pf: If the likelihood of θ_1 is larger than the likelihood of θ_1, i.e. $f_n(\mathbf{x} \mid \theta_2) > f_n(\mathbf{x} \mid \theta_1)$ it does NOT mean that θ_2 is more likely.
- Remember: θ is not random here.

Examples

- Let $X \sim \text{Bernoulli}(\theta)$. Find the maximum likelihood estimator of θ. Say we observe $\sum x_i = 3$, what is the maximum likelihood estimate of θ?
- Let $X_1, \ldots, X_n$ be i.i.d. $N\left(\mu, \sigma^2\right)$.
- Find the MLE of μ when σ^2 is known.

Recap

Steps to MLE estimation:

- Find the likelihood function.
- Find the log likelihood function.
- Take the derivative to find the global optimum $\hat{\theta}$
- Use the second derivative to check that $\hat{\theta}$ is a maximizer.