Parameter Estimation and Decision Making

Wei Wang @ HKUST(GZ)

April 3, 2025

Outline

1 Parameter Estimation

Example

- Problem:
  - Observe whether the sky is cloudy or not cloudy on n successive days
  - Predict whether the sky will be cloudy on the $n+1^{\text{th}}$ day
- Step 1: Parameter estimation
  - Model the unknown as a random variable with a parameterized distribution with unknown parameter (Bayesian) or Model the unknown as a fixed but unknown constant (Frequentist).
  - Guess the unknown parameter/constant.
- Step 2: Decision making
  - Use guess about unknown parameter to find probability of event of interest
  - Decide based on the probability

Two Major Categories

Suppose you have $x_1, x_2, \cdots, x_R \sim_{\text{(i.i.d.)}} \mathcal{N}(\mu, \sigma^2)$
But you don't know μ (you do know σ^2)

- Maximum Likelihood (MLE): For which μ is $x_1, x_2, \cdots, x_R$ most likely?
- Maximum a Posterior (MAP): Which μ maximizes $p(\mu | x_1, x_2, \cdots, x_R, \sigma^2)$

Question

Which one do you prefer?

Question

Despite the intuitiveness of MAP, we'll spend 95% of our time on MLE. Why?

Frequentist Estimation Problem

- Problem: find "the true value" of a parameter based on data sample
- Estimator: function from **sample space** to **parameter space**
- Estimate: specific **point** in sample space.
- Loss: measure of error w.r.t. true value of parameter

Properties of Estimators

- (Asymptotic) Consistency
  - Whether true value is recovered for infinite sample size
- Bias
  - Expected deviation of estimate from true value
- Variance
- Mean squared error
  - Bias-variance trade-off

Properties of Maximum Likelihood Estimator

- Asymptotically Unbiased
- Consistent
- Smallest variance among unbiased estimators (aka. asympototic efficiency)

Bayesian Parameter Estimation

- Model parameter θ as a random variable
- Prior distribution $P(\theta)$

- Maximum a posteriori probability estimation problem
  - Find posterior distribution $P(\theta|D)$ of θ given observed data D

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}$$

- Likelihood $L(\theta) = P(D|\theta)$

Three Types of Point Estimation

- Frequentist:
  - **Maximum likelihood estimator**

$$\theta_{ML} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} P(D|\theta)$$

- Bayesian:
  - **Maximum a posterior estimator**

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(\theta)P(D|\theta)$$

  - **Bayesian estimator**

$$\theta_{Bayes} = E[\theta] = \int \theta P(\theta|D) d\theta$$

Question

Can you draw a figure to distinguish the three?

Outline

We will illustrate how to perform these point estimation using examples.

Example 1: Maximum Likelihood Estimator

- Given a sequence of coin tosses, guess probability of getting head H
- Model $X \sim_{\text{i.i.d.}} Ber(p)$
- Likelihood $L(p) = P(X_1, X_2, \ldots; p)$
- Log likelihood

$$\ell(p) \overset{\text{def}}{=} \log L(p) = \sum_i P(X_i; p) = n_H \log p + n_T \log(1-p)$$

  where n_H is #Heads and n_T is #Tails in N tosses
- Maximize $\ell(p)$ by setting $\frac{\partial \ell(p)}{\partial p} = 0$ and verify maximality.
- Maximum likelihood estimate

$$\hat{p}_{ML} = \arg\max_p \ell(p) = \frac{n_H}{N}$$

Example 1: MAP Estimator

Model p as random variable with a prior distribution

$$p \sim Beta(a, b); \quad f(p) \propto p^{a-1}(1-p)^{b-1} \qquad \text{(Conjugate prior)}$$

Formulate posterior distribution

$$p(p|D) \propto f(p) \sum_i P(X_i; p) = p^{a+n_H-1}(1-p)^{b+n_T-1}$$

- Because

$$\pi(p|x) \propto \binom{n}{x} p^x(1-p)^{n-x} \cdot \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}$$

$$\pi(p|x) \propto p^x(1-p)^{n-x} \cdot p^{a-1}(1-p)^{b-1}$$

$$\pi(p|x) \propto p^{x+a-1}(1-p)^{n-x+b-1} \qquad \text{(rearrange } p \text{ and } 1-p \text{ terms)}$$

Maximum a posteriori estimate

$$\hat{p}_{MAP} = \arg\max_p p(p|D) = \frac{n_H + a - 1}{N + a + b - 2}$$

Example 1: Bayes Estimator

Model p as random variable with a prior distribution

$$p \sim Beta(a, b); \quad f(p) \propto p^{a-1}(1-p)^{b-1} \qquad \text{(Conjugate prior)}$$

Formulate posterior distribution

$$\begin{aligned} p(p|D) \propto f(p) \sum_i P(X_i; p) &= p^{a+n_H-1}(1-p)^{b+n_T-1} \\ &= Beta(a+n_H, b+n_T) \end{aligned}$$

Bayes estimate

$$\hat{p}_B = \mathbb{E}[p \mid X_1, \ldots, X_n] = \frac{n_H + a}{N + a + b}$$

Example 1: Bayes Estimator

$$\begin{aligned}
\hat{p}_B &= E[p|X_1, \ldots, X_n] \\
&= \frac{n_H + a}{N + a + b} \\
&= \frac{a + b}{N + a + b} \cdot \frac{n_H}{a + b} + \frac{N}{N + a + b} \cdot \frac{n_H}{N} \\
&= \frac{a + b}{N + a + b} \cdot E[p] + \frac{N}{N + a + b} \cdot \hat{p}_{ML}
\end{aligned}$$

- Weighted average of prior mean and MLE
- Weight of MLE proportional to number of observations

Role of priors

- Uniform prior vs. Beta prior
- With uniform prior

$$f(p) \propto 1$$

$$p(p|D) \propto f(p) \sum_i P(X_i; p) = p^{n_H+1}(1-p)^{n_T+1}$$

$$\hat{p}_{MAP} = \arg\max_p p(p|D) = \frac{n_H + 1}{N + 2}$$

Example 2: MLE for univariate Gaussian

- Suppose you have $x_1, x_2, ...x_R \sim$ (i.i.d) $N(\mu, \sigma^2)$
- But you don't know μ (you do know σ^2)
- MLE: For which μ is $x_1, x_2, ...x_R$ most likely?

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, ...x_R | \mu, \sigma^2)$$

Example 2: Algebra Euphoria

$$\begin{aligned}\mu^{mle} &= \arg\max_{\mu} p(x_1, x_2, ...x_R|\mu, \sigma^2) \\ &= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i|\mu, \sigma^2) \qquad \text{(by i.i.d)}\end{aligned}$$

Example 2: Algebra Euphoria

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, ...x_R|\mu, \sigma^2)$$

$$= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i|\mu, \sigma^2) \qquad \text{(by i.i.d)}$$

$$= \arg\max_{\mu} \sum_{i=1}^{R} \log p(x_i|\mu, \sigma^2) \qquad \text{(monotonicity of log)}$$

Example 2: Algebra Euphoria

$$\begin{aligned}
\mu^{mle} &= \arg\max_{\mu} p(x_1, x_2, ... x_R | \mu, \sigma^2) \\
&= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i | \mu, \sigma^2) && \text{(by i.i.d)} \\
&= \arg\max_{\mu} \sum_{i=1}^{R} \log p(x_i | \mu, \sigma^2) && \text{(monotonicity of log)} \\
&= \arg\max_{\mu} \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{R} -\frac{(x_i - \mu)^2}{2\sigma^2} && \text{(plug in formula for Gaussian)}
\end{aligned}$$

Example 2: Algebra Euphoria

$$
\begin{aligned}
\mu^{mle} &= \arg\max_{\mu} p(x_1, x_2, ...x_R|\mu, \sigma^2) \\
&= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i|\mu, \sigma^2) && \text{(by i.i.d)} \\
&= \arg\max_{\mu} \sum_{i=1}^{R} \log p(x_i|\mu, \sigma^2) && \text{(monotonicity of log)} \\
&= \arg\max_{\mu} \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{R} -\frac{(x_i-\mu)^2}{2\sigma^2} && \text{(plug in formula for Gaussian)} \\
&= \arg\min_{\mu} \sum_{i=1}^{R} (x_i-\mu)^2 && \text{(after simplification)}
\end{aligned}
$$

Intermission: A General Scalar MLE strategy

Task: Find MLE θ assuming known form for $P(\text{Data} \mid \theta, \text{stuff})$

1. Write $\ell = \log P(\text{Data} \mid \theta, \text{stuff})$
2. Work out $\frac{\partial \ell}{\partial \theta}$
3. Set $\frac{\partial \ell}{\partial \theta} = 0$ for a maximum, creating an equation in terms of θ
4. Solve it*
5. Check that you've found a maximum rather than a minimum or saddle-point, and be careful if θ is constrained

*This is a perfect example of something that works perfectly in all textbook examples and usually involves surprising pain if you need it for something new.

Example 2: The MLE μ

$$
\begin{aligned}
\mu^{mle} &= \arg\max_{\mu} p(x_1, x_2, \ldots x_R | \mu, \sigma^2) \\
&= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2 \\
&= \mu \text{ s.t. } 0 = \frac{\partial \ell}{\partial \mu} = \ldots
\end{aligned}
$$

Example 2: The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, ...x_R|\mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$= \mu \text{ s.t. } 0 = \frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{R} (x_i - \mu)^2 = \sum_{i=1}^{R} 2(x_i - \mu)$$

Example 2: The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, ...x_R | \mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$= \mu \text{ s.t. } 0 = \frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{R} (x_i - \mu)^2 = \sum_{i=1}^{R} 2(x_i - \mu)$$

Thus, $\mu = \frac{1}{R} \sum_{i=1}^{R} x_i$.

Example 2: Lawks-a-lawdy!

$$\mu^{mle} = \frac{1}{R} \sum_{i=1}^{R} x_i$$

- The best estimate of the mean of a distribution is the mean of the sample!

1. Unsurprising, but with MLE justifications
2. Naive and MLE estimates of σ^2 will be different

Example 3: MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \dots x_R \sim_{\text{(i.i.d)}} \mathcal{N}(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu, \sigma^2)$ is $x_1, x_2, \dots x_R$ most likely?

$$\log p(x_1, x_2, \dots x_R | \mu, \sigma^2) = -R(\log \pi + \frac{1}{2} \log \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{R} (x_i - \mu)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{R} (x_i - \mu)^2$$

Example 3: MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R \sim_{\text{(i.i.d)}} \mathcal{N}(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu, \sigma^2)$ is $x_1, x_2, \ldots x_R$ most likely?

$$\log p(x_1, x_2, \ldots x_R | \mu, \sigma^2) = -R(\log \pi + \frac{1}{2} \log \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2} \sum_{i=1}^{R} (x_i - \mu)$$

$$0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{R} (x_i - \mu)^2$$

Example 3: MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R \sim_{\text{(i.i.d)}} \mathcal{N}(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu, \sigma^2)$ is $x_1, x_2, \ldots x_R$ most likely?

$$\log p(x_1, x_2, \ldots x_R | \mu, \sigma^2) = -R(\log \pi + \frac{1}{2} \log \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2} \sum_{i=1}^{R} (x_i - \mu) \Rightarrow \mu = \frac{1}{R} \sum_{i=1}^{R} x_i$$

$$0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{R} (x_i - \mu)^2 \Rightarrow \text{what?}$$

Example 3: MLE for univariate Gaussian

- Suppose you have $x_1, x_2, ... x_R \sim_{(\text{i.i.d})} \mathcal{N}(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu, \sigma^2)$ is $x_1, x_2, ... x_R$ most likely?

$$\log p(x_1, x_2, ... x_R | \mu, \sigma^2) = -R(\log \pi + \frac{1}{2} \log \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$\mu^{mle} = \frac{1}{R} \sum_{i=1}^{R} x_i$$

$$\sigma^2_{mle} = \frac{1}{R} \sum_{i=1}^{R} (x_i - \mu^{mle})^2$$

Unbiased Estimators

- An estimator of a parameter is **unbiased** if the expected value of the estimate is the **same** as the true value of the parameters.
- If $x_1, x_2, \ldots x_R \sim_{(\text{i.i.d})} \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbf{E}\left[\!\left[\mu^{mle}\right]\!\right] = \mathbf{E}\left[\!\left[\frac{1}{R}\sum_{i=1}^{R} x_i\right]\!\right] = \mu$$

- Hence, μ^{mle} is unbiased

Biased Estimators

- An estimator of a parameter is **biased** if the expected value of the estimate is **different from** the true value of the parameters.
- If $x_1, x_2, \ldots x_R \sim_{\text{(i.i.d)}} \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbf{E}\left[\!\left[\sigma^2_{mle}\right]\!\right] = \mathbf{E}\left[\!\left[\frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2\right]\!\right] = \mathbf{E}\left[\!\left[\frac{1}{R}\sum_{i=1}^{R}\left(x_i - \frac{1}{R}\sum_{j=1}^{R}x_j\right)^2\right]\!\right] \neq \sigma^2$$

- Hence, σ^2_{mle} is biased

MLE Variance Bias

- If $x_1, x_2, \ldots x_R \sim_{\text{(i.i.d)}} \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbf{E}\left[\!\left[\sigma^2_{mle}\right]\!\right] = \mathbf{E}\left[\!\left[\frac{1}{R}\sum_{i=1}^{R}\left(x_i - \frac{1}{R}\sum_{j=1}^{R}x_j\right)^2\right]\!\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

- Intuition check: consider the case of $R = 1$

Question

Why should our guts expect that σ^2_{mle} would be an underestimate of true σ^2?

Question

How could you prove

$\mathbf{E}\left[\!\left[\frac{1}{R}\sum_{i=1}^{R}\left(x_i - \frac{1}{R}\sum_{j=1}^{R}x_j\right)^2\right]\!\right] = \left(1 - \frac{1}{R}\right)\sigma^2$?

Unbiased estimate of Variance

- If $x_1, x_2, ...x_R \sim_{\text{(i.i.d)}} \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbf{E}\left[\!\left[\sigma^2_{mle}\right]\!\right] = \mathbf{E}\left[\!\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right]\!\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

So define $\sigma^2_{\text{unbiased}} = \frac{\sigma^2_{mle}}{\left(1 - \frac{1}{R}\right)}$

And $\mathbf{E}\left[\!\left[\sigma^2_{\text{unbiased}}\right]\!\right] = \sigma^2$

$$\sigma^2_{\text{unbiased}} = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

Unbiaseditude discussion

Question

Which one is better?

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

$$\sigma^2_{\text{unbiased}} = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

Unbiaseditude discussion

Question

Which one is better?

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

$$\sigma^2_{\text{unbiased}} = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

Answer:

- It depends on the task
- And doesn't make much difference once $R \to$ large

Don't get too excited about being unbiased

- Assume $x_1, x_2, ...x_R \sim_{(\text{i.i.d})} \mathcal{N}(\mu, \sigma^2)$
- Suppose we had these estimators for the mean

$$\mu^{\text{suboptimal}} = \frac{1}{R + 7\sqrt{R}} \sum_{i=1}^{R} x_i$$

$$\mu^{\text{crap}} = x_1$$

Questions

- Are either of these unbiased?
- Will either of them asymptote to the correct value as R gets large?
- Which is more useful?

Decision Theory

- Choose a specific point estimate under uncertainty
- Loss functions measure extent of error
- Choice of estimate depends on loss function

Loss Functions

- 0-1 loss

$$L(y,a) = I(y \neq a) = \begin{cases} 0 \text{ if } a = y \\ 1 \text{ if } a \neq y \end{cases}$$

  - Minimized by MAP estimate (posterior mode)
- l_2 loss

$$L(y,a) = (y-a)^2$$

  - Expected loss: $E[(y-a)^2|x]$ (Min mean squared error)
  - Minimized by Bayes estimate (posterior mean)
- l_1 loss

$$L(y,a) = |y-a|$$

  - Minimized by posterior median

Loss Functions

- Cross-entropy loss
  - Binary classificaiton: y is the prob of positive class

$$L(y, a) = y \log(a) + (1 - y) \log(1 - a)$$

  - Multi-class classificaiton: $y(a)$ is the prob distribution of all K classes, and k is the true class

$$L(y, a) = \log(a_k), k \text{ is the true class}$$

  - Equivalent to KL divergence

$$H(y, a) = H(y) + D_{KL}(y||a)$$

Predictive distribution

- Find the probability of the outcome of the $n+1^{\text{th}}$ experiment given outcomes of previous n experiments

$$P(A_{n+1}|A_1, ..., A_n)$$

- Frequentist
  - Construct point estimate of parameter $\hat{\theta}$ from n outcomes

$$P(A_{n+1}|A_1, ..., A_n) \cong P(A_{n+1}; \hat{\theta})$$

- Bayesian
  - Consider the entire posterior distribution of θ

$$P(A_{n+1}|A_1, ..., A_n) = \int P(A|\theta)P(\theta|A_1, ..., A_n)d\theta$$

Summary

- Parameter estimation problem
- Frequentist vs Bayesian
- MLE, MAP and Bayes estimators for Bernoulli trials
- Optimal estimators for different loss functions
- Prediction using estimated parameters