Maximum Likelihood Estimator

Review

Consider a sample of n iid random variables $X_1, X_2, \dots, X_n$.

Maximum Likelihood Estimator (MLE)

What parameter θ **maximizes the likelihood** of our observed data $(X_1, X_2, \dots, X_n)$?

$$L(\theta) = f(X_1, X_2, \dots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} \underbrace{f(X_1, X_2, \dots, X_n|\theta)}_{\text{likelihood of data}}$$

Observations:

- MLE determines θ value that maximizes the probability of observing the sample.
- If we're estimating θ, couldn't we just maximize the probability of θ?

Today: **Bayesian estimation** using the Bayesian definition of probability!

Maximum A Posteriori (MAP) Estimator

Not Review! New!

Consider a sample of n iid random variables $X_1, X_2, \dots, X_n$.

Maximum Likelihood Estimator (MLE)

What parameter θ **maximizes the likelihood** of our observed data $(X_1, X_2, \dots, X_n)$?

$$L(\theta) = f(X_1, X_2, \dots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} \underbrace{f(X_1, X_2, \dots, X_n|\theta)}_{\text{likelihood of data}}$$

Maximum a Posteriori (MAP) Estimator

Given the sample data $(X_1, X_2, \dots, X_n)$, what is the **most probable parameter** θ?

$$\theta_{MAP} = \arg\max_{\theta} \underbrace{f(\theta|X_1, X_2, \dots, X_n)}_{\text{posterior distribution of } \theta}$$

Maximum A Posteriori (MAP) Estimator

Consider a sample of n iid random variables $X_1, X_2, \ldots, X_n$.

<u>def</u> The **Maximum a Posteriori (MAP) Estimator** of θ is the value of θ that maximizes the **posterior** distribution of θ.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

Intuition with Bayes' Theorem:

$L(\theta)$, probability of data given parameter θ

notice that both the prior and the posterior are focusing on θ!

After seeing data, posterior belief of θ

likelihood prior

posterior

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})}$$

Before seeing data, prior belief of θ

Solving for θ_{MAP}

- Observe data: $X_1, X_2, \ldots, X_n$, all iid
- Let likelihood be same as MLE: $f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\,\theta)$
- Let the prior distribution of θ be $g(\theta)$.

posterior distribution likelihood function priori distribution

$$\theta_{MAP} = \arg\max_{\theta} \underbrace{f(\theta|X_1, X_2, \ldots, X_n)}_{\text{data}} = \arg\max_{\theta} \frac{f(X_1, X_2, \ldots, X_n|\theta)g(\theta)}{h(X_1, X_2, \ldots, X_n)} \quad \text{(Bayes' Theorem)}$$

data distribution

$$= \arg\max_{\theta} \frac{g(\theta)\prod_{i=1}^{n} f(X_i|\,\theta)}{\cancel{h(X_1, X_2, \ldots, X_n)}} \quad \text{(independence)}$$

θ와 관련없음

$$= \arg\max_{\theta} g(\theta)\prod_{i=1}^{n} f(X_i|\,\theta) \quad (1/h(X_1, X_2, \ldots, X_n) \text{ is a positive constant w.r.t. } \theta)$$

log 변환

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\,\theta)\right)$$

θ_{MAP}: Interpretation 1

- Observe data: $X_1, X_2, \ldots, X_n$, all iid
- Let likelihood be same as MLE: $f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\,\theta)$
- Let the prior distribution of θ be $g(\theta)$.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n) = \arg\max_{\theta} \frac{f(X_1, X_2, \ldots, X_n|\theta)g(\theta)}{h(X_1, X_2, \ldots, X_n)} \quad \text{(Bayes' Theorem)}$$

$$= \arg\max_{\theta} \frac{g(\theta)\prod_{i=1}^{n} f(X_i|\,\theta)}{h(X_1, X_2, \ldots, X_n)} \quad \text{(independence)}$$

$$= \arg\max_{\theta} g(\theta) \prod_{i=1}^{n} f(X_i|\,\theta) \quad (1/h(X_1, X_2, \ldots, X_n) \text{ is a positive constant w.r.t. } \theta)$$

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\,\theta) \right)$$

θ_{MAP} maximizes
log prior + log-likelihood

θ_{MAP}: Interpretation 2

- Observe data: $X_1, X_2, \ldots, X_n$, all iid
- Let likelihood be same as MLE: $f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\,\theta)$
- Let the prior distribution of θ be $g(\theta)$.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n) = \text{ar}\ldots \quad \text{(Bayes' Theorem)}$$

The mode of the
posterior distribution of θ

$$= \arg\max_{\theta} \frac{g(\theta)\prod_{i=1}^{n} f(X_i|\,\theta)}{h(X_1, X_2, \ldots, X_n)} \quad \text{(independence)}$$

$$= \arg\max_{\theta} g(\theta) \prod_{i=1}^{n} f(X_i|\,\theta) \quad (1/h(X_1, X_2, \ldots, X_n) \text{ is a positive constant w.r.t. } \theta)$$

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\,\theta) \right)$$

θ_{MAP} maximizes
log prior + log-likelihood

Mode: A statistic of a random variable

The mode of a random variable X is defined as:

(X discrete, PMF $p(x)$) $\arg\max_x p(x)$

$\arg\max_x f(x)$ (X continuous, PDF $f(x)$)

- Intuitively: The value of X that is "most likely".
- Note that some distributions may not have a unique mode (e.g., Uniform distribution, or Bernoulli(0.5))

$$\theta_{MAP} = \arg\max_\theta f(\theta | X_1, X_2, \ldots, X_n)$$

θ_{MAP} is the most likely θ given the data $X_1, X_2, \ldots, X_n$.

Back to our happy Laplace

Consider our previous 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

Recall θ_{MLE}: $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$ ⚠️
$p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

What are your Laplace estimates for each roll outcome?

$$p_i = \frac{X_i + 1}{n + m}$$

$X_3 = 0 \Rightarrow \frac{0+1}{12+6} = \frac{1}{18}$

$p_1 = 4/18, p_2 = 3/18, p_3 = 1/18,$ ✅
$p_4 = 4/18, p_5 = 2/18, p_6 = 4/18$

Laplace smoothing:
- Easy to implement/remember
- **Avoids parameter estimation of 0**

often very important

Conjugate distributions

MAP estimator:

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

The mode of the posterior distribution of θ

"ideal" model for prior and posterior

Distribution parameter	Conjugate distribution
Bernoulli p	Beta
Binomial p	Beta
Multinomial p_i	Dirichlet
Poisson λ	Gamma
Exponential λ	Gamma
Normal μ	Normal
Normal σ^2	Inverse Gamma

generalization of Beta