Maximum Likelihood Estimation

Machine Learning for Research Students

Lecturer Name

Department of Machine Learning

May 23, 2025

Outline

The Role of Parameter Estimation

Key Ideas

- Model data-generating mechanism via $p(x \mid \theta)$
- Parameters θ capture location, scale, dependencies
- Accurate estimates enable reliable prediction and inference

Historical Evolution of MLE

Timeline

- 1898–1901: Pearson's Method of Moments
- 1922: Fisher's formalization of the likelihood principle
- 1930s–1940s: Development of asymptotic theory

Key Asymptotic Properties

Properties

- Consistency: $\hat{\theta}_{\text{MLE}} \to \theta_0$ as $n \to \infty$
- Asymptotic normality:
  $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$
- Efficiency: achieves Cramér–Rao lower bound asymptotically

Intuitive Picture of Likelihood

Conceptual View

- Likelihood $L(\theta)$ measures plausibility of θ given data
- Maximizing $L(\theta)$ aligns model to observations
- Negative log-likelihood $-\ell(\theta)$ as loss function

Illustration: Coin Toss Example

Setup

- Observations: $X_i \in \{\mathrm{H}, \mathrm{T}\}$, $i = 1, \ldots, n$
- Parameter $p = P(X_i = \mathrm{H})$
- Likelihood: $L(p) = p^k(1-p)^{n-k}$, where k heads
- MLE: $\hat{p} = k/n$

Data and Model Setup

Assumptions

- IID samples $X_1, \ldots, X_n \sim p(x \mid \theta)$
- Parameter space $\Theta \subseteq \mathbb{R}^d$
- Goal: infer true parameter θ_0

Defining the Likelihood Function

Equation

$$L(\theta; X_{1:n}) = \prod_{i=1}^{n} p(X_i \mid \theta).$$

Remarks

- Data fixed, viewed as function of θ
- Product form can underflow for large n

Log-Likelihood Transformation

Equation

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p(X_i \mid \theta).$$

Benefits

- Converts product to sum for numerical stability
- Same maximizer as $L(\theta)$

Score Function and Its Role

Equation

$$s(\theta) = \nabla_\theta \ell(\theta) = \sum_{i=1}^{n} \nabla_\theta \log p(X_i \mid \theta).$$

Interpretation

- Sensitivity of log-likelihood to parameter changes
- Critical point: $s(\hat{\theta}) = 0$

First- and Second-Order Conditions

Conditions for Maximum

- First order: $s(\hat{\theta}) = 0$
- Second order: $\nabla^2 \ell(\hat{\theta}) \prec 0$ (negative definite)
- Ensures local maximum of $\ell(\theta)$

Regularity Conditions

Required Assumptions

- Support of $p(x \mid \theta)$ independent of θ
- Differentiability under the integral sign
- Finite positive-definite Fisher information

Definition of the MLE

Equation

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} L(\theta).$$

Solution Approach

Solve $s(\theta) = 0$ and verify concavity or use numerical methods.

Example I: Bernoulli Model

Model

$X_i \sim \text{Bernoulli}(p)$, $p \in (0, 1)$.

Equation

$$\ell(p) = \sum_{i=1}^{n} \left[X_i \log p + (1 - X_i) \log(1 - p) \right].$$

MLE

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Example II: Gaussian Model

Model

$X_i \sim \mathcal{N}(\mu, \sigma^2)$, both parameters unknown.

Equation

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

MLE

$$\hat{\mu} = \bar{X}, \quad \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Example III: Exponential Family

Equation

$$f(x;\theta) = h(x)\exp\big\{\eta(\theta)^\top T(x) - A(\theta)\big\}.$$

Score Equation

$$\nabla A(\theta) = \frac{1}{n}\sum_{i=1}^{n} T(X_i).$$

Fisher Information Matrix

Equation

$$I(\theta) = -\mathbb{E}\big[\nabla^2 \ell(\theta)\big] = \mathbb{E}\big[s(\theta)s(\theta)^\top\big].$$

Zero Mean and Variance of the Score

Results

- $\mathbb{E}[s(\theta)] = 0$
- $\mathrm{Var}[s(\theta)] = I(\theta)$

Cramér–Rao Lower Bound

Equation

$$\mathrm{Var}(\hat{\theta}) \succeq \frac{1}{n} I(\theta)^{-1}.$$

Implication

MLE attains this bound asymptotically under regularity.

Consistency of the MLE

Sketch

- Law of large numbers: $\ell(\theta)/n \to \mathbb{E}[\log p(X \mid \theta)]$
- Unique maximizer at true θ_0
- Thus $\hat{\theta}_{\mathrm{MLE}} \to \theta_0$

Asymptotic Normality

Equation

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\big(0, I(\theta_0)^{-1}\big).$$

Sketch of Normality Proof

Outline

- Taylor expand $s(\hat{\theta})$ around θ_0
- Apply CLT to $s(\theta_0)$ and LLN to $-\nabla^2 \ell(\theta)$
- Solve for $\sqrt{n}(\hat{\theta} - \theta_0)$

Invariance of MLE

Property

For a one-to-one transformation g, $\widehat{g(\theta)} = g(\hat{\theta}_{\mathrm{MLE}})$.

Newton–Raphson Algorithm

Equation

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla^2 \ell(\theta^{(t)})\right]^{-1} \nabla \ell(\theta^{(t)}).$$

Fisher Scoring and EM

Methods

- Fisher scoring: replace Hessian by expected information $I(\theta)$
- EM algorithm: handle latent-variable models via E- and M-steps

Example: Logistic Regression

Equation

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \ln \sigma(x_i^\top \beta) + (1 - y_i) \ln(1 - \sigma(x_i^\top \beta)) \right].$$

Note

No closed-form solution; use iterative solvers (e.g., Newton–Raphson).

When MLE Struggles

Challenges

- Non-identifiable parameterizations
- Boundary estimates (e.g., $p = 0$ or 1)
- Model misspecification: bias and inconsistency

Summary of Key Takeaways

Recap

- MLE: principled estimation via likelihood maximization
- Asymptotic properties: consistency, normality, efficiency
- Practical computation: Newton–Raphson, Fisher scoring, EM

Further Reading

References

- Casella & Berger, *Statistical Inference*
- van der Vaart, *Asymptotic Statistics*
- Murphy, *Machine Learning: A Probabilistic Perspective*