Mathematics for Machine Learning

MAD-B3-2526-S2-MAT0611

Maximum Likelihood Estimation

Agenda

1. Probability Foundations
2. Random Variables & Distributions
3. Parameter Estimation Overview
4. **Maximum Likelihood Estimation**
5. MLE Methodology
6. Examples & Applications
7. Hands-on Exercises

Probability Recap

Sample Space & Events

- **Sample space** Ω: set of all possible outcomes
- **Event** $A \subseteq \Omega$: subset of outcomes
- **Probability measure** $P : \mathcal{F} \to [0, 1]$
  - $P(\Omega) = 1$
  - $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

Probability Operations

Key Concepts

- **Intersection** (both events occur): $P(A \cap B)$
- **Union** (either event occurs): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Complement** (event does not occur): $P(A^c) = 1 - P(A)$

Independence: Events A and B are independent if:

$$P(A \cap B) = P(A) \cdot P(B)$$

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Law of Total Probability:

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

Random Variables

Definition: A function $X : \Omega \to \mathbb{R}$ that assigns a real number to each outcome

Types

- **Discrete**: countable values (e.g., binary outcomes, dice rolls)
- **Continuous**: uncountable values (e.g., height, temperature)

Distributions

Definition: A probability distribution describes how probabilities are assigned to values of a random variable

Key Idea

- **Distribution** = complete description of random variable's behavior
- Specified by PDF/PMF (or equivalently, CDF)
- Captures all probabilistic information about the variable

Parametric families: Distributions characterized by parameters

- E.g., $N(\mu, \sigma^2)$ - two parameters define entire family
- Our goal: estimate these parameters from data

Probability Functions

Definition: Functions that describe the distribution of a random variable

Probability Mass Function (PMF)

For discrete random variables:

$$p_X(x) = P(X = x)$$

Probability Density Function (PDF)

For continuous random variables:

$f_X(x)$ where $P(a \leq X \leq b) = \int_a^b f_X(x)dx$

Cumulative Distribution Function (CDF)

$$F_X(x) = P(X \leq x)$$

Expected Value:

- Discrete: $E[X] = \sum_x x \cdot p_X(x)$
- Continuous: $E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$

Variance:

$$\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

Common Discrete Distributions

- **Bernoulli**
  $X \sim \mathrm{Ber}(p)$
  $$P(X = k) = p^k(1-p)^{1-k}, \quad k \in \{0, 1\}$$
- **Binomial**
  $X \sim \mathrm{Bin}(n, p)$
  $$P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$
- **Poisson**
  $X \sim \mathrm{Poi}(\lambda)$
  $$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

Common Discrete Distributions (contd.)

Geometric

$X \sim \mathrm{Geo}(p)$

If counts number of failures before first success:

$$P(X = k) = (1 - p)^k p, \quad k = 0, 1, 2, \ldots$$

If counts number of trials until first success:

$$P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

Common Continuous Distributions

- **Normal**

  $X \sim N(\mu, \sigma^2)$

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

- **Exponential**

  $X \sim \mathrm{Exp}(\lambda)$

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

- **Gamma**

  $X \sim \mathrm{Gamma}(\alpha, \beta)$

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x \geq 0$$

Common Continuous Distributions (contd.)

- **Uniform**
  $X \sim U(a, b)$

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

- **Beta**
  $X \sim \mathrm{Beta}(\alpha, \beta)$

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 \leq x \leq 1$$

- If $\alpha = \beta = 1$, the distribution is Uniform(0,1).

Random Samples

Random Variables: $X_1, X_2, \ldots, X_n$

- Theoretical quantities (before observation)
- Each X_i is a function from $\Omega \to \mathbb{R}$

Observations/Realizations: $x_1, x_2, \ldots, x_n$

- Actual data values obtained (after observation)
- Specific numbers: e.g., $x_1 = 3.2, x_2 = 5.1, \ldots$

Simple Random Sample: $X_1, \ldots, X_n$ are i.i.d.

- **Independent**: knowing X_i tells us nothing about X_j
- **Identically distributed**: all have same distribution $f(x; \theta)$

Joint Distribution of Sample

For i.i.d. random variables $X_1, \ldots, X_n \sim f(x; \theta)$:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Key properties:

- Product structure comes from independence
- Each factor is identical (same f, same θ)

Parameter Estimation

The Problem

- We observe data $X_1 = x_1, \ldots, X_n = x_n$ from distribution $f(x; \theta)$
- $X_1, \ldots, X_n$ are i.i.d.
- Parameter θ is **unknown**
- Goal: estimate θ from the data

Estimator: $\hat{\theta} = g(X_1, \ldots, X_n)$

- The estimator is a function of random variables, so it is also a random variable

Estimate: $\hat{\theta} = g(x_1, \ldots, x_n)$

- The estimate is the realized value of the estimator calculated from observed data

Properties of Estimators

Bias: $\mathrm{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

- **Unbiased**: $E[\hat{\theta}] = \theta$

Mean Squared Error:

$$\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2$$

Consistency: $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$

Efficiency: lower variance among unbiased estimators

Uniformly Minimum Variance Unbiased Estimator

UMVUE: Among all unbiased estimators of θ, there is one that has smallest variance for all θ

Cramér-Rao Lower Bound:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

where $I(\theta)$ is the **Fisher Information**:

$$I(\theta) = E\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right]$$

An unbiased estimator achieving this bound is **efficient** (and is the UMVUE)

Method of Moments (MOM)

Idea: Equate sample moments to population moments

Population moments: $\mu_k = E[X^k]$

Sample moments: $m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$

Solve: $\mu_k(\theta) = m_k$ for $k = 1, 2, \ldots$

Example: For $X \sim N(\mu, \sigma^2)$

- $\hat{\mu}_{\text{MOM}} = \bar{X}$
- $\hat{\sigma}^2_{\text{MOM}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Maximum Likelihood Estimation (MLE)

Principle

Choose $\hat{\theta}$ that **maximizes** the likelihood function:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta; \mathbf{x})$$

Interpretation:

- Select parameter value making observed data "most likely"
- Most plausible explanation for the data

The Likelihood Function

Definition: Given data $\mathbf{x} = (x_1, \ldots, x_n)$, the likelihood function is:

$$L(\theta; \mathbf{x}) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Key Insight:

- Same formula as joint PDF/PMF, but different perspective
- View as function of θ (not $\mathbf{x}$)
- Measures "plausibility" of parameter value given data
- Higher $L(\theta)$ means θ is more consistent with observed data

Likelihood vs Probability

Probability: Fixed θ, variable data

- "What data might we observe?"
- $P(X = x|\theta)$

Likelihood: Fixed data, variable θ

- "Which parameter values are consistent with observed data?"
- $L(\theta|\mathbf{x})$

Log-Likelihood Function

Definition:

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

Why use log-likelihood?

- Converts products to sums (easier computation)
- Numerically stable (avoids underflow)
- Preserves location of maximum (log is monotonic)
- Simplifies derivatives

$$\arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ell(\theta)$$

MLE Methodology

Step-by-Step Process

1. **Write the likelihood**: $L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta)$
2. **Take the log**: $\ell(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log f(x_i; \theta)$
3. **Differentiate**: $\frac{\partial \ell}{\partial \theta}$
4. **Set to zero**: $\frac{\partial \ell}{\partial \theta} = 0$ (score equation)
5. **Solve for** $\hat{\theta}$
6. **Verify maximum**: Check $\frac{\partial^2 \ell}{\partial \theta^2} < 0$ at $\hat{\theta}$

Example 1: Bernoulli Distribution

Setup: $X_1, \ldots, X_n \sim \text{Ber}(p)$ i.i.d., where $f(x; p) = p^x(1-p)^{1-x}$

Likelihood:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Log-likelihood:

$$\ell(p) = \left(\sum_{i=1}^{n} x_i\right) \log p + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p)$$

Example 1: Bernoulli (continued)

Calculate derivative with respect to p:

$$\frac{\partial \ell}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p}$$

Set derivative to zero:

$$\frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0 \Rightarrow \frac{\sum x_i}{p} = \frac{n - \sum x_i}{1-p} \Rightarrow \frac{1-p}{p} = \frac{n - \sum x_i}{1-p}$$

Solve:

$$\hat{p}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

Example 2: Normal Distribution (Known Variance)

Setup: $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ i.i.d., estimate μ

Log-likelihood:

$$\ell(\mu) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Derivative:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)$$

Example 2: Normal (continued)

Set to zero:

$$\sum_{i=1}^{n}(x_i - \mu) = 0$$

Solve:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Second derivative:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\sigma_0^2} < 0 \quad \checkmark$$

Example 3: Normal Distribution (Both Parameters)

Setup: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ i.i.d., estimate both μ and σ^2

Log-likelihood:

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

Partial derivatives:

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

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

Example 3: Normal (continued)

Solve system:

From first equation: $\hat{\mu}_{\text{MLE}} = \bar{x}$

Substitute into second:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note: MLE for σ^2 is **biased**

Bias: $\text{Bias}(\hat{\sigma}^2_{\text{MLE}}) = -\frac{\sigma^2}{n} \to 0$ as $n \to \infty$

MLE for σ^2 is Biased

Claim: $E[\hat{\sigma}^2_{\text{MLE}}] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$

Proof: Add and subtract μ inside the square:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2$$

Expand:

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

MLE for σ^2 is Biased (continued)

Take expectations:

$$E[\hat{\sigma}^2_{\text{MLE}}] = E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = E\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - nE[(\bar{X} - \mu)^2]$$

Since $E[(X_i - \mu)^2] = \sigma^2$ and $E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$:

$$= n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2$$

Therefore:

$$E[\hat{\sigma}^2_{\text{MLE}}] = \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n}\sigma^2$$

Example 4: Exponential Distribution

Setup: $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ i.i.d., where $f(x; \lambda) = \lambda e^{-\lambda x}$

Log-likelihood:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

Derivative:

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

MLE:

$$\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

Example 5: Poisson Distribution

Setup: $X_1, \ldots, X_n \sim \mathrm{Poi}(\lambda)$ i.i.d., where $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Log-likelihood:

$$\ell(\lambda) = \left(\sum_{i=1}^{n} x_i\right) \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

Derivative:

$$\frac{\partial \ell}{\partial \lambda} = \frac{\sum x_i}{\lambda} - n$$

MLE:

$$\hat{\lambda}_{\mathrm{MLE}} = \bar{x}$$

Properties of MLEs

Asymptotic Properties

1. **Consistency**: $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$
2. **Asymptotic Normality**:

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$$

3. **Asymptotic Efficiency**: Achieves Cramér-Rao lower bound asymptotically
4. **Invariance**: If $\hat{\theta}$ is MLE of θ, then $g(\hat{\theta})$ is MLE of $g(\theta)$

Numerical Optimization

When analytical solution is difficult

Newton-Raphson Method:

$$\theta^{(k+1)} = \theta^{(k)} - \left[\frac{\partial^2 \ell}{\partial \theta^2}\right]^{-1} \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta^{(k)}}$$

Gradient Ascent:

$$\theta^{(k+1)} = \theta^{(k)} + \alpha \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta^{(k)}}$$

Software: scipy.optimize, statsmodels, sklearn

MLE in Python: Example Setup

```python
import numpy as np
from scipy.optimize import minimize
import matplotlib.pyplot as plt

# Generate sample data from exponential distribution
np.random.seed(42)
true_lambda = 2.5
n = 100
data = np.random.exponential(scale=1/true_lambda, size=n)

# Analytical MLE
mle_analytical = 1 / np.mean(data)
print(f"Analytical MLE: {mle_analytical:.4f}")
print(f"True parameter: {true_lambda}")
```

MLE in Python: Numerical Optimization

```
# Define negative log-likelihood (minimize instead of maximize)
def neg_log_likelihood(lam, data):
    if lam <= 0:
        return np.inf
    return -np.sum(np.log(lam) - lam * data)

# Numerical optimization
result = minimize(neg_log_likelihood,
                  x0=[1.0],  # initial guess
                  args=(data,),
                  method='L-BFGS-B',
                  bounds=[(0.001, None)])

mle_numerical = result.x[0]
print(f"Numerical MLE: {mle_numerical:.4f}")
```

Resources

- Casella & Berger: *Statistical Inference* (Ch. 7)
- Wasserman: *All of Statistics* (Ch. 9)
- Murphy: *Probabilistic Machine Learning* (Ch. 4)