

Section 02: Supplementary Material

Prepared by: Shuyang Hou

1. Maximum Likelihood Estimation, MLE

1.1. Mathematical Definition

Maximum Likelihood Estimation (MLE) is a method for estimating unknown parameters of a probabilistic model. Given observed data, MLE selects the parameters under which those observations have the highest likelihood.

Suppose X follows a parametric distribution $p(x; \theta)$, where $\theta \in \Theta$ is unknown. Let

$$\mathcal{D} = \{x_1, x_2, \dots, x_n\}$$

be n independent and identically distributed (i.i.d.) observations. The joint probability mass function or density is

$$p(\mathcal{D}; \theta) = \prod_{i=1}^n p(x_i; \theta).$$

As a function of θ with the observations fixed, this is the **likelihood function**:

$$L(\theta; \mathcal{D}) = \prod_{i=1}^n p(x_i; \theta).$$

The maximum likelihood estimator is

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Theta} L(\theta; \mathcal{D}) = \arg \max_{\theta \in \Theta} \prod_{i=1}^n p(x_i; \theta).$$

Because the logarithm is strictly increasing, maximizing the likelihood is equivalent to maximizing the **log-likelihood**:

$$\ell(\theta) = \log L(\theta; \mathcal{D}) = \sum_{i=1}^n \log p(x_i; \theta).$$

Equivalently, MLE minimizes the **negative log-likelihood (NLL)**:

$$\hat{\theta}_{\text{MLE}} = \arg \min_{\theta \in \Theta} \left[-\frac{1}{n} \sum_{i=1}^n \log p(x_i; \theta) \right].$$

The factor $1/n$ does not change the minimizer.

1.2. Example for MLE: Gaussian Distribution

Suppose

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \sigma^2 > 0,$$

and both μ and σ^2 are unknown. The Gaussian density is

$$p(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

The log-likelihood of the observed sample is

$$\begin{aligned} \ell(\mu, \sigma^2) &= \sum_{i=1}^n \log p(x_i; \mu, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

Estimating the Mean μ

Taking the derivative with respect to μ gives

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

Setting the derivative equal to zero yields

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

Therefore,

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

That is, the MLE of the Gaussian mean is the sample mean.

Estimating the Variance σ^2

Writing $v = \sigma^2$, we obtain

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

Setting this derivative equal to zero gives

$$v = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

Substituting $\hat{\mu}_{\text{MLE}} = \bar{x}$, we obtain

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

Thus, for a nonconstant sample, the joint Gaussian MLE is

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

Important distinction. The Gaussian MLE of the variance uses $1/n$, whereas the usual unbiased sample variance uses $1/(n-1)$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

An MLE need not be unbiased. If all observations are identical, the likelihood has no maximizer over $\sigma^2 > 0$; its supremum is reached only in the limit $\sigma^2 \downarrow 0$.

1.3. MLE in Machine Learning

Many common machine learning objectives arise by minimizing negative log-likelihood under different assumptions about the data distribution.

Linear Regression and Mean Squared Error (MSE)

Consider the linear regression model

$$y_i = \mathbf{w}^\top \mathbf{x}_i + b + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Conditioned on $\mathbf{x}_i$,

$$y_i \mid \mathbf{x}_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, \sigma^2).$$

For fixed $\sigma^2 > 0$, the negative log-likelihood is

$$\begin{aligned} -\ell(\mathbf{w}, b) &= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i - b)^2 \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i - b)^2 + C, \end{aligned}$$

where C does not depend on $\mathbf{w}$ or b . Consequently,

$$(\hat{\mathbf{w}}, \hat{b})_{\text{MLE}} = \arg \min_{\mathbf{w}, b} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i - b)^2.$$

Thus, **Gaussian MLE is equivalent to minimizing MSE** for the regression parameters.

Logistic Regression and Binary Cross-Entropy

For binary classification, let

$$Y_i \mid \mathbf{x}_i \sim \text{Bernoulli}(p_i), \quad p_i = \text{sigmoid}(\mathbf{w}^\top \mathbf{x}_i + b),$$

where $\text{sigmoid}(z) = 1/(1 + e^{-z})$. The log-likelihood is

$$\ell(\mathbf{w}, b) = \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)].$$

Minimizing its negative, averaged over observations, gives

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{n} \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)].$$

This is exactly the **binary cross-entropy loss**.

1.4. Key Takeaways

- (a) **Assume a probability model** $p(x; \theta)$ for the observations.
- (b) **Construct the likelihood** from the observed data:

$$L(\theta; \mathcal{D}) = \prod_{i=1}^n p(x_i; \theta).$$

- (c) **Take logarithms** to turn the product into a sum:

$$\ell(\theta) = \sum_{i=1}^n \log p(x_i; \theta).$$

- (d) **Estimate the parameters** by maximizing log-likelihood, or equivalently minimizing negative log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \ell(\theta) = \arg \min_{\theta} [-\ell(\theta)].$$

Probability versus likelihood: With parameters fixed, probability describes how observations are distributed. With observations fixed, likelihood measures how well different parameter values explain the data.

Main idea: MLE chooses the parameters that make the observed data most likely under the assumed model.