Derivation of Maximum Likelihood Estimators (MLE)

Teaching team of STA314

Oct 21, 2024

MLE for Bernoulli Distribution

Problem: Flipping a coin with outcomes heads (1) and tails (0). p denotes the probability of getting head. $(x_1, \cdots x_n)$ independent samples.

Write out likelihood, then solve for its derivative, get closed form solution.

MLE for Bernoulli Distribution

Problem: Flipping a coin with outcomes heads (1) and tails (0). p denotes the probability of getting head. $(x_1, \cdots x_n)$ independent samples.

Likelihood:

$$L(p \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Log-Likelihood:

$$\ell(p) = \sum_{i=1}^{n} [x_i \ln p + (1-x_i) \ln(1-p)]$$

Derivative and Solution:

$$\frac{d\ell(p)}{dp} = \sum_{i=1}^{n} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p} \right) = 0$$

$$p^* = \frac{1}{n} \sum_{i=1}^{n} x_i$$

MLE for Uniform(0, θ) Distribution

Problem: Observations $x_1, x_2, \ldots, x_n$ are drawn from a uniform distribution $U(0, \theta)$.

Parameter: θ is the unknown upper bound, and we assume all x_i are between 0 and θ.

MLE for θ? Don't need strict mathematics derivation, just give a guess first.

MLE for Uniform(0, θ) Distribution

Problem: Observations $x_1, x_2, \ldots, x_n$ are drawn from a uniform distribution $U(0, \theta)$. **Parameter:** θ is the unknown upper bound, and we assume all x_i are between 0 and θ.

Likelihood:

$$L(\theta \mid x_1, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n}, & \text{if } \theta \geq \max(x_1, \ldots, x_n) \\ 0, & \text{otherwise} \end{cases}$$

Log-Likelihood:

$$\ell(\theta) = \begin{cases} -n \ln \theta, & \text{if } \theta \geq \max(x_1, \ldots, x_n) \\ -\infty, & \text{otherwise} \end{cases}$$

MLE Solution: To maximize the likelihood, θ must be as small as possible while still being at least the largest observed value:

$$\theta^* = \max(x_1, \ldots, x_n).$$

MLE for Linear Regression

Model: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The errors are i.i.d., normally distributed with mean 0 and variance σ^2.
Parameter: β.

Write out likelihood using pdf of normal distribution, then take derivative with respect to β.

MLE for Linear Regression - Likelihood

Model: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The errors are i.i.d., normally distributed with mean 0 and variance σ^2.
Parameter: β.

Likelihood:

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right)$$

Log-Likelihood: Taking the logarithm:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

MLE for Linear Regression - Direct Solution

Step-by-Step Derivation:

1. Focus on minimizing the sum of squared errors (OLS):

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

2. Take the gradient of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$:

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2X^\top (y - X\boldsymbol{\beta}).$$

3. Set the gradient to zero to find the optimal solution:

$$X^\top X\boldsymbol{\beta} = X^\top y.$$

4. Solve for $\boldsymbol{\beta}$ (assuming $X^\top X$ is invertible):

$$\boldsymbol{\beta}^* = (X^\top X)^{-1} X^\top y.$$

This is the MLE for linear regression.

Gradient Descent for Linear Regression

Objective: Minimize the sum of squared errors:

$$J(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

Gradient:

$$\nabla_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) = -X^\top (y - X\boldsymbol{\beta}).$$

Gradient Descent Update Rule:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \eta X^\top (y - X\boldsymbol{\beta}_t),$$

where η is the learning rate.

Steps: 1. Initialize $\boldsymbol{\beta}_0$ randomly. 2. Update $\boldsymbol{\beta}$ iteratively using the rule above. 3. Stop when the gradient norm is small or the change in $J(\boldsymbol{\beta})$ is negligible.

MLE for Logistic Regression

Model: $P(y_i = 1 \mid \boldsymbol{x}_i) = \frac{1}{1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})}$.

Write out Log-likelihood, get the gradient for β, write out gradient descent update rule.

MLE for Logistic Regression - Model and Log-Likelihood

Model: $P(y_i = 1 \mid \boldsymbol{x}_i) = \frac{1}{1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})}$.

Likelihood:

$$L(\boldsymbol{\beta} \mid y, X) = \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})} \right)^{y_i} \left(1 - \frac{1}{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})} \right)^{1-y_i}$$

Log-Likelihood: Taking the logarithm:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \ln \left(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \right) \right]$$

Gradient Descent for Logistic Regression

Objective: Maximize the log-likelihood function:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \ln\left(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right) \right]$$

Gradient:

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \frac{1}{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})} \right).$$

Gradient Descent Update Rule:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \eta \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \frac{1}{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta}_t)} \right),$$

where η is the learning rate.

Steps: 1. Initialize $\boldsymbol{\beta}_0$ randomly. 2. Update $\boldsymbol{\beta}$ iteratively using the rule above. 3. Stop when the gradient norm is small or the log-likelihood stabilizes.

Note: Logistic regression has no closed-form solution, so gradient descent (or other optimization methods) is necessary.

Conclusion

- MLE provides a principled approach for parameter estimation.
- Some models (e.g., Bernoulli) have closed-form solutions.
- Others (e.g., Logistic Regression) require optimization techniques.