GLeMa: Maximum Likelihood Estimation for Generalized Linear Models

Stochastic Batman

December 2025

1 Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is fundamentally a technique that operates in the reverse direction of traditional probability. While probability functions calculate the likelihood of observing data given specified parameters, MLE seeks to identify the optimal parameters that maximize the likelihood of observing the given data. This method is essential for many statistical procedures and forms the backbone of various algorithms used in data science and machine learning, such as the Naive Bayes classifier [3], Gaussian mixture models [4], and more. By efficiently identifying parameters, MLE enhances model performance and predictive accuracy, allowing for robust analysis and decision-making based on empirical data.

Intuitive Explanation

The basic intuition behind Maximum Likelihood Estimation is grounded in probability. Suppose we have a random sample of data, $X_1, X_2, \ldots, X_n$, and we assume this data comes from a specific probability distribution (like a Normal or Bernoulli distribution). This distribution depends on some unknown parameter, denoted by θ.

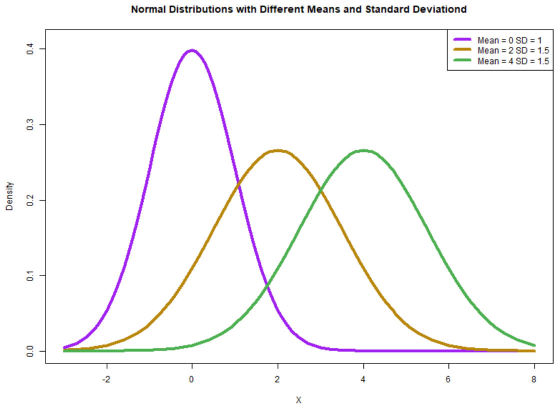

Figure 1: Normal distributions with different means and standard deviations

The method suggests that the optimal parameter value is the specific θ that maximizes the probability (or likelihood) of having generated the data we actually collected.

In a nutshell: We treat the probability of our data as a function of the unknown parameter. We then ask, "Which value of the parameter makes our observed data most likely to have happened?"

The Likelihood Function

Definition: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \ldots, \theta_m)$. Suppose that the parameter vector $(\theta_1, \ldots, \theta_m)$ is restricted to a given parameter space Ω.

The joint probability density (or mass) function of $X_1, X_2, \ldots, X_n$, when regarded as a function of $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_m$, is called the **Likelihood Function**, denoted by L:

$$L(\boldsymbol{\theta}) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

The first equality is merely the definition of the joint probability mass function. The second equality stems from the fact that we have a random sample, which inherently means that the samples are independent.

The function is defined for all parameter vectors $(\theta_1, \ldots, \theta_m)$ in the parameter space Ω.

Example: Likelihood for Bernoulli Data

Suppose we have a random sample of n independent coin flips $X_1, \ldots, X_n$, where $X_i \in \{0, 1\}$, and the probability of success is p. The probability mass function is $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$. The likelihood function $L(p)$ is the product of the individual probabilities:

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

The Necessity of Log-Likelihood

For large sample sizes, the Likelihood Function $L(\theta)$, which is a product of many probability values (each between 0 and 1), quickly becomes an extremely small number. This causes two main problems:

1. **Numerical Underflow:** Computational systems may fail to store these tiny numbers accurately, leading to a value of zero (underflow).

2. **Calculus Complexity:** Products are difficult to differentiate. By taking the logarithm, the product simplifies to a sum, which is mathematically much easier to optimize.

Mathematical Justification for Log-Likelihood

The use of the natural logarithm is justified because it is a *strictly monotonically increasing function.* This property ensures that the maximum point of the Likelihood function $L(\theta)$ is exactly the same as the maximum point of the Log-Likelihood function $\ln L(\theta)$.

If we consider any two parameter values θ_a and θ_b:

$$L(\theta_a) > L(\theta_b) \iff \ln(L(\theta_a)) > \ln(L(\theta_b))$$

Therefore, maximizing the simpler function, $\ln L(\theta)$, yields the identical maximum likelihood solution, $\hat{\theta}$. The optimization procedure involves differentiating the Log-Likelihood, setting it to zero, and solving for θ:

$$\frac{\partial \ln L(\theta)}{\partial \theta} \overset{\text{set}}{=} 0$$

Example: Log-Likelihood for Bernoulli Data

Applying the natural logarithm to the Bernoulli likelihood function $L(p)$ from the example above:

$$\ln L(p) = \ln \left(p^{\sum x_i} (1-p)^{n - \sum x_i} \right)$$

Using the logarithm property $\ln(a^b c^d) = b\ln(a) + d\ln(c)$:

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

Therefore, we have (remember, the necessary condition for a maximum is that the derivative equals 0):

$$\frac{\partial \ln L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} \text{ which we set to 0}$$

Then:

$$\left(\sum x_i \right) (1-p) - \left(n - \sum x_i \right) p = 0 \implies \sum x_i - np = 0$$

Solving for p (hat on the parameter $\hat{p}$ indicates that it is an estimate) to get an estimate:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ or, alternatively, an estimator: } \hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Estimators vs. Estimates

A common source of confusion in mathematical statistics is the distinction between an *estimator* and an *estimate*. While used interchangeably in casual conversation, they represent distinct mathematical concepts.

The Recipe vs. The Meal

Before defining formally, to understand the difference intuitively, consider the analogy of cooking:

1. **The Estimator (The Recipe):** Think of an estimator as a *rule* or a *formula*. It tells you how to process ingredients (data) to get a result. Before you actually cook, the recipe exists as a procedure. In statistics, the estimator is the formula we choose to use (e.g., "I will calculate the average of whatever data I collect"). Because we haven't collected the data yet, the result of the estimator is unknown and random.

2. **The Estimate (The Meal):** Think of the estimate as the specific result you get after you have finished cooking. It is a concrete, fixed product. In statistics, once you collect your specific data points and plug them into the formula, you get a single number (e.g., ”The average is 5.2”). This number is the estimate.

The Statistic

A **statistic** is formally defined as any function of the observable random variables in a sample that does not depend on any unknown parameters.

Let $X_1, X_2, \ldots, X_n$ be a random sample. A statistic T is any function:

$$T = g(X_1, X_2, \ldots, X_n)$$

Since T is a function of random variables, the statistic T itself is a random variable and has a probability distribution (called its sampling distribution).

The Estimator

An **estimator** is a statistic, which is a function of the random sample $X_1, X_2, \ldots, X_n$. Because the sample variables X_i are random variables (before the experiment is performed), the estimator is also a random variable.

Let Θ be the parameter of interest. An estimator $\hat{\Theta}$ is defined as:

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

The function h is the mathematical **rule or procedure** that maps the values of the random sample space (the observed data) to a value in the parameter space (the estimate). It represents the specific formula chosen for estimation (e.g., the formula for the sample mean or the sample variance). The choice of h determines the properties of the estimator $\hat{\Theta}$.

The Estimate

An **estimate** is a specific realization of the estimator. It is calculated using the observed values $x_1, x_2, \ldots, x_n$ (after the experiment is performed).

Let $x_1, \ldots, x_n$ be the observed data. The estimate $\hat{\theta}$ is:

$$\hat{\theta} = h(x_1, x_2, \ldots, x_n)$$

Feature	**Estimator**	**Estimate**
Symbol	$\hat{\Theta}$ or $\bar{X}$ (Capital)	$\hat{\theta}$ or $\bar{x}$ (Lowercase)
Mathematical Nature	Random Variable (Function)	Constant (Real Number)
Input	Random Sample $X_1, \ldots, X_n$	Observed Data $x_1, \ldots, x_n$
Timing	Pre-data collection	Post-data collection
Properties	Bias, Variance, Consistency	Accuracy (for a specific case)

Table 1: Comparison of Estimators and Estimates

References

[1] **PennState - Eberly College of Science.** STAT 415 - Introduction to Mathematical Statistics, 1.2 - Maximum Likelihood Estimation. From `https://online.stat.psu.edu/stat415/lesson/1/1.2`

[2] **GLeMa GitHub Repository.** MLE meets GLM. `https://github.com/Stochastic-Batman/GLeMa`

[3] Langley, P., & Sage, S. (1994). **The importance of naive Bayes**. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 333-337). Morgan Kaufmann.

[4] McLachlan, G. J., & Basford, K. E. (1988). **Mixture Models: Inference and Applications to Clustering**. Wiley.