How machines learn from Errors by themself?

Static concept about Parameter Estimation

INFO 7390 Advances Data Sci/Architecture SEC 03 Spring 2025

Yifan Yuan

What Does It Mean for a Machine to Learn from Errors?

First of all we should clear that the machine here is not the normal machine like vehicles in ourdaily life, instead it's represent the algorithm or model that processes data and learns from it.
So lets think about how a person learn? Assume we are a child and we are trying to learn how to ride a bike, The first few attempts often result in falls, but with each mistake, they adjust—finding better balance, controlling speed, and steering more effectively. Over time, they fall less and ride more smoothly. This process of trial and error is remarkably similar to how machines learn from their mistakes.

Machines Don’t Think, They Adjust

Human will think about why we are doing wrong but machines don't. Instead When a machine makes a mistake it won't have motions like "why I'm i doing wrong" or "what make this things bad". Instead, it analyzes the error and and keep adjusts its parameters, so it can improves its predictions using statistical methods. and this is the statistical methods what we are going talk about. The Parameter Estimation

What is Parameter Estimation

Parameter estimation refers to the process of estimating the parameters of a population distribution based on sample data. In statistics, we usually assume the form of the population distribution, such as normal distribution, Poisson distribution, etc., and assume that the parameters of the population distribution are unknown. The goal of parameter estimation is to estimate the parameters of the overall distribution through a certain method based on sample data.

Why this is important

Parameter estimation is fundamental in statistics, machine learning, and data science because it allows us to make informed decisions based on limited data. Since we rarely have access to an entire population, we rely on sample data to infer key characteristics of the population.

How we do Parameter Estimation

Maximum Likelihood Estimation(MLE)

To know what Maximum Likelihood Estimation is, we first need to understand what Likelihood means.

Likelihood Function

Likelihood refers to a function that measures how probable a given set of observed data is, given a particular statistical model with unknown parameters. Unlike probability, which describes the chance of an event occurring, likelihood evaluates how well a specific parameter value explains the observed data.

We assume there is a data set $D = x_1, x_2, \ldots, x_n$ and it is Independent and identically distributed （i.i.d.). So we could get：
$L(\theta; D) = P(D \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta)$

Log-Likelihood Function

Maximizing the product of probabilities $\prod_{i=1}^{n} P(x_i \mid \theta)$ directly can be computationally challenging due to:

1. *Numerical underflow*: Multiplying many small values (e.g., probabilities) may result in near-zero values.
2. *Complex optimization*: Derivatives of products are messy to compute.

To address this, we take the natural logarithm of the likelihood function:
$\ell(\theta; D) = \ln L(\theta; D) = \sum_{i=1}^{n} \ln P(x_i \mid \theta)$

Why logarithms?

- **Monotonicity**: $\ln$ is a strictly increasing function, so maximizing $\ell(\theta; D)$ is equivalent to maximizing $L(\theta; D)$.
- **Simplification**: Sums are easier to differentiate and optimize than products.
- **Numerical stability**: Avoids underflow by converting products into sums.

Maximum Likelihood Estimation Function

The MLE is the method that trying to estimating the parameters of a model by maximizing the likelihood function .So by doing this we could say the the parameters we estimated can observed data most probable.

To make the Likelihood Function max with θ:
$\theta_{\text{MLE}} = \arg\max_{\theta} L(\theta; D)$ If we use Log likelihood function it would be
$\theta_{\text{MLE}} = \arg\max_{\theta} \ell(\theta; D)$

Maximum A Posteriori Estimation (MAP)

Maximum A Posteriori Estimation is an extention by incorporating prior knowledge about parameters.

Posterior Probability

MAP estimation combines the likelihood of the data with a **prior distribution** over the parameters. Unlike MLE, which only considers the data, MAP introduces domain knowledge or regularization through the prior.

Given data $D = x_1, x_2, \ldots, x_n$ (i.i.d.), the posterior probability is proportional to:
$P(\theta \mid D) \propto P(D \mid \theta) \cdot P(\theta)$
where:

- $P(D \mid \theta)$: Likelihood of the data
- $P(\theta)$: Prior distribution of parameters

MAP Estimation Function

MAP aims to find the parameter value that maximizes the posterior probability:
$\theta_{\text{MAP}} = \arg\max_\theta P(\theta \mid D) = \arg\max_\theta P(D \mid \theta) \cdot P(\theta)$

Bayesian Inference

Bayesian methods go beyond point estimation (like MLE/MAP) to model the **full posterior distribution** of parameters, enabling uncertainty quantification.

Posterior Distribution

Bayes' theorem updates beliefs about θ by combining likelihood and prior:
$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$
where $P(D) = \int P(D \mid \theta) P(\theta) d\theta$ is the marginal likelihood (often intractable analytically).

Predictive Distribution

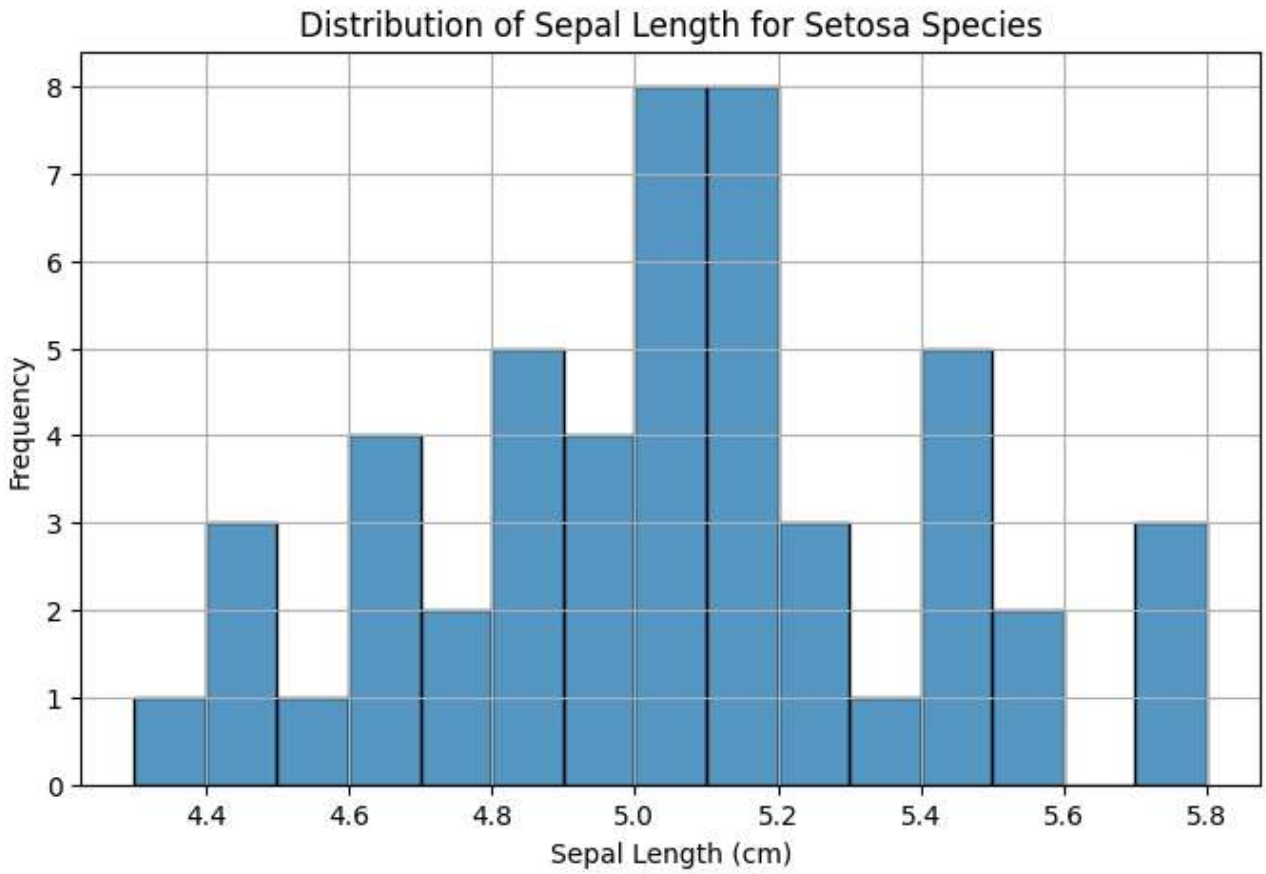
Instead of using a single θ, Bayesian methods integrate over all possible θ:
$P(y_{\text{new}} \mid D) = \int P(y_{\text{new}} \mid \theta) P(\theta \mid D) d\theta$
This accounts for parameter uncertainty in predictions.

Example using Parameter Estimation

Parameter Estimation in normal distribution

We use Iris data set as a example for it have a data that seems to be a normal distribution so we can use it to do the Parameter Estimation about it's normal distribution
We use the setosa_sepal_length as an example of how we do the Parameter Estimation in normal distribution

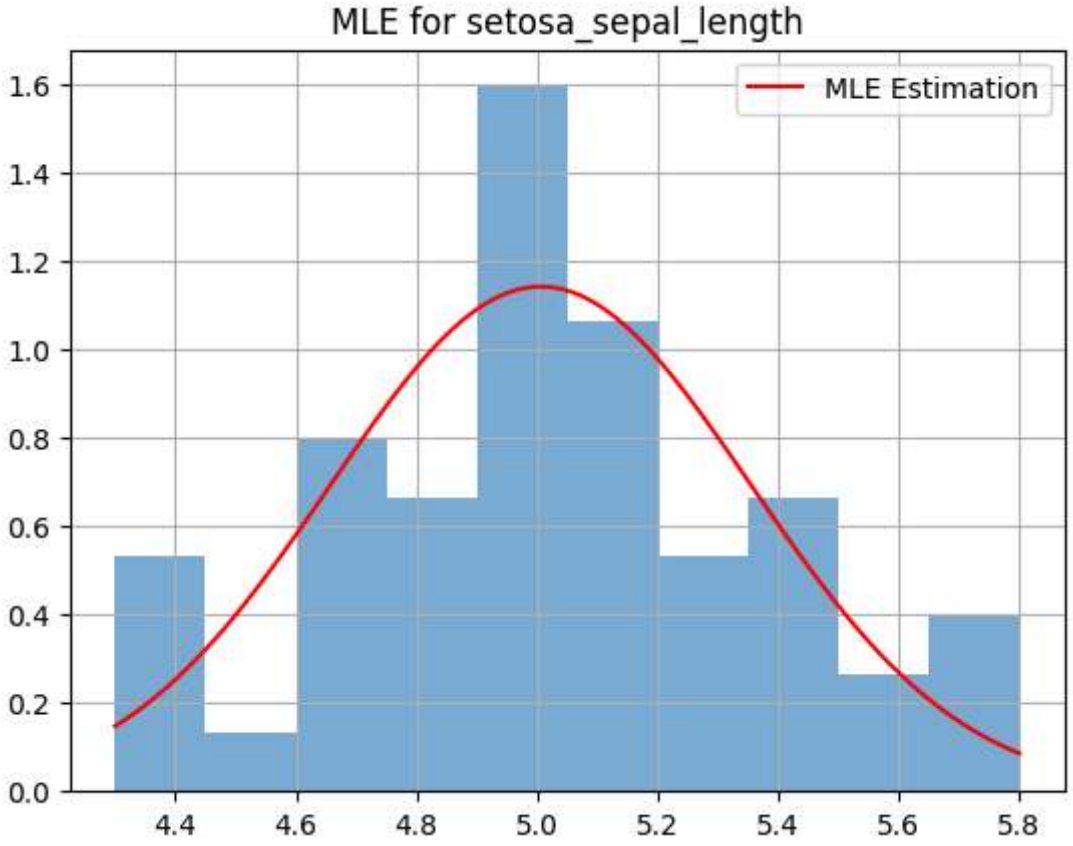
Visualize of the data of setosa sepal length



We could assume this dataset follow the Normal Distribution so we can do the Parameter Estimation

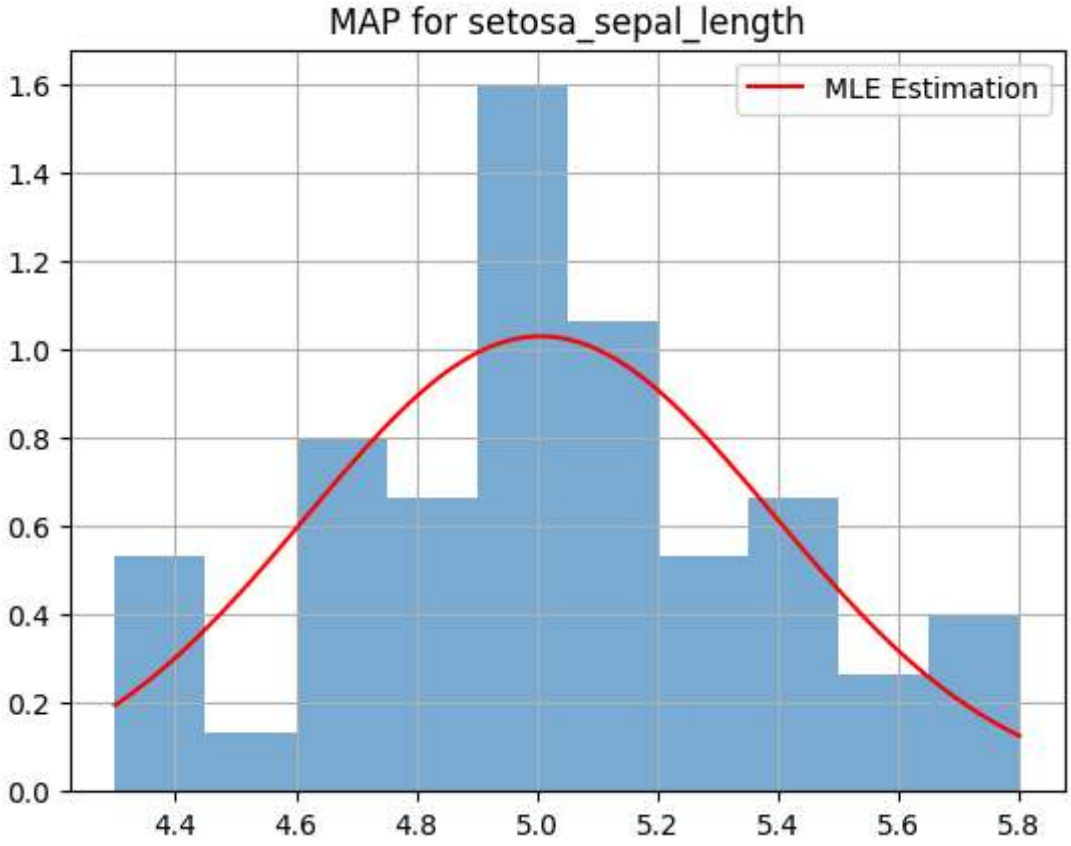
Visualize of MLE

MLE: $\mu=5.01$, $\sigma^2=0.12$



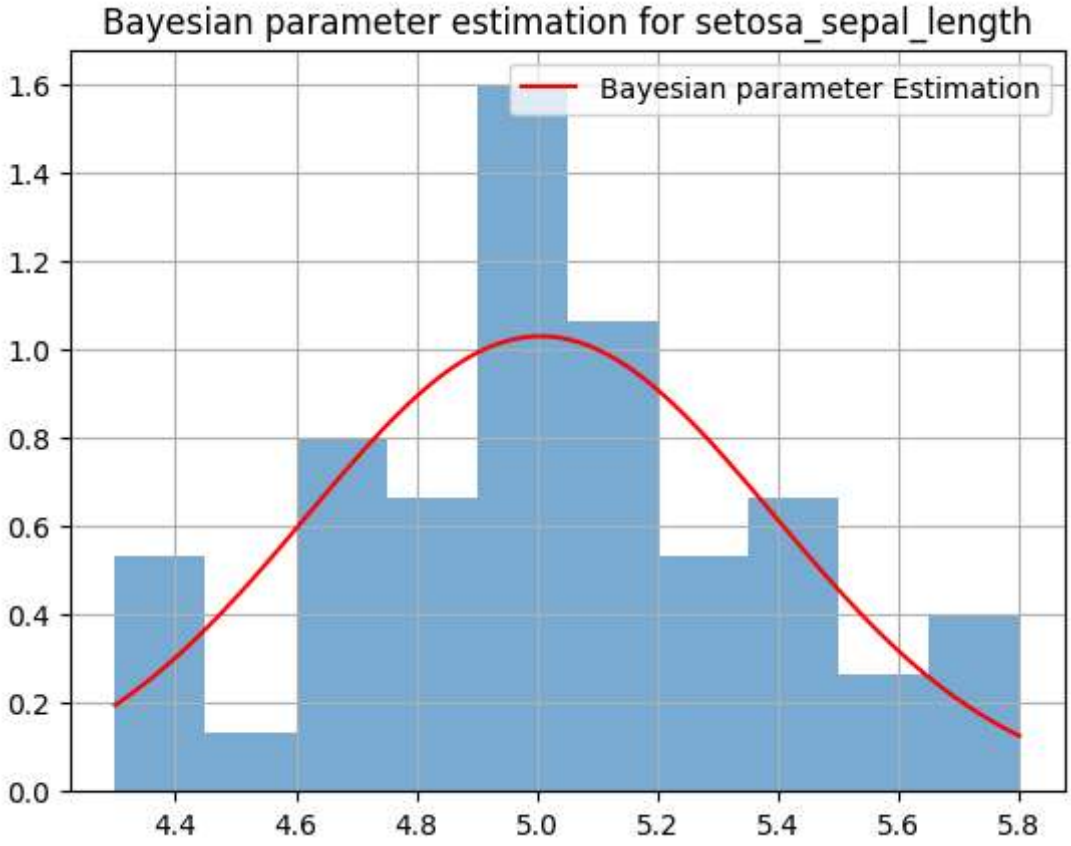
Visualize of MAP

MAP: μ=5.01, σ^2=0.15

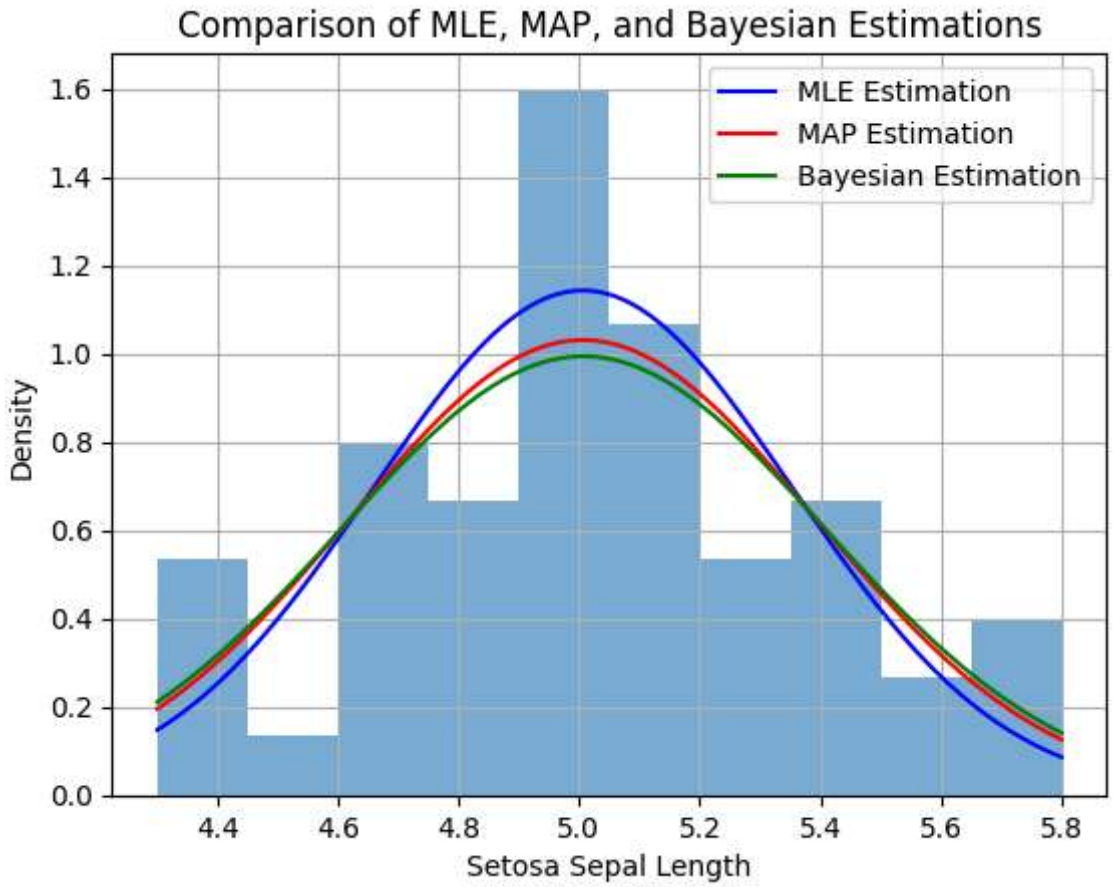


Visualize of Bayesian parameter estimation

Bayesian: $\mu=5.01$, $\sigma^2=0.16$



Comparison of MLE, MAP, and Bayesian Estimations



When we use different Parameter Estimation

Maximum Likelihood Estimation (MLE)

Key Points:

Rely entirely on data and find parameter values that maximize the probability of observing data.

Applicable scenarios

- *Large Datasets:* The asymptotic properties of MLE are most reliable with ample data.
- *No Prior Information:* When you lack strong prior beliefs or wish to remain completely data-driven.

Limitation

- It is easy to overfit in small samples
- Uncertainty in parameters cannot be quantified.

Maximum A Posteriori (MAP) Estimation

Key Points:

Introduce the prior distribution based on MLE and find the parameter value that maximizes the posterior probability

Applicable scenarios

- *Small Datasets:* The asymptotic properties of MLE are most reliable with ample data.
- *No Prior Information:* When you lack strong prior beliefs or wish to remain completely data-driven.

Limitation

- It is still a point estimate and cannot capture the parameter distribution.
- Improper prior selection may lead to bias.

Bayesian Estimation

Key Points:

Computes the posterior distribution of parameters, providing a complete probabilistic description

Applicable scenarios

- *need to quantify parameter uncertainty:*
- *Small sample learning, combined with priors to improve robustness.*
- *the posterior distribution can be updated incrementally.*
- *Generate predictive distributions.*

Limitation

- It is still a point estimate and cannot capture the parameter distribution.
- Improper prior selection may lead to bias.

Summary and Future work

Parameter Estimation can be vary different from models to models, but the main forcures is to find the parameters that can make the model represent the real dataset best
For the future we need to forcures more about when the complexity of models are increas how we do the Parameter Estimation like doing the Backpropagation inneural networks

References:

irs dataset : https://archive.ics.uci.edu/dataset/53/iris
MLE MAP Bayesian Estimation expalination :https://www.geeksforgeeks.org/parameter-estimation/