Maximum Likelihood

Likelihood: A fundamental concept in stats that measures how probable your observed data is, given specific parameter values for statistic model.

likelihood is opposite of probability

Probability: A numerical measure that quantifies how likely it is for that event to occur as a result of random experiment.

likelihood flips the question around. Instead of asking "what data will i see" we ask "what parameter explains the data i already saw?

Probability: "Given this coin, what results might i see?"

likelihood: "Given this result, what coin did i probably have?"

Probability: This is a measure of the chance that a certain event will occur out of all possible events. It's usually presented as a ratio or fraction, and it ranges from 0 (meaning the event will not happen) to 1 (meaning the event is certain to happen).

Likelihood: In statistical context, likelihood is a function that measures the plausibility of a particular parameter value given some observed data. It quantifies how well a specific outcome supports specific parameter values.

More Definitions

A probability quantifies how often you observe a certain outcome of a test, given a certain understanding of the underlying data.

A likelihood quantifies how good one's model is, given a set of data that's been observed.

Probabilities describe test outcomes, while likelihoods describe models.

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given some observed data.

Example : Coin Toss

$P(H) = 0.5$

HHHHH combination ⟶ Likelihood of this happening is very less

$L(P=0.5 \mid HHHHH) = (0.5)^5$

if $L(P=0.6 \mid HHHHH) = (0.6)^5$

This is greater..

So, by going this way, what value of p be that the MLE is maximum? off course 1

$L(P=1 \mid HHHHH) = (1)^5$

MLE for Normal distribution:

In data, when we find the MLE for a data point, what we are actually doing is finding the best values for parameters which are u and σ, Since we Always take assumption that our data follows normal distribution, but the value of (u, σ) might differ based on each data point.

u = mean (Center of bell curve)

σ = Standard deviation (width/spread of bell curve)

General formula for Normal $N(u, \sigma)$ distribution :

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

What we need is to find the perfect value for the parameters (μ, σ). To do this, we try each combination of μ and σ values and select the one with Maximum likelihood.

$$L(\mu, \sigma | x_1, x_2, x_3, \dots x_n)$$ ← data

{ Here we assume the data is independent }

$$L(\mu, \sigma | x_1, x_2, \dots x_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_2-\mu)^2}{2\sigma^2}} \times \dots \dots$$

$$\dots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}$$

likelihood

Now we take log both sides (because log has a very special property: $\log ab = \log a + \log b$)

↓ Becomes log likelihood.

$$\log(L(\mu, \sigma | x_1, x_2 \dots x_n) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\right) + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}\right) + \dots$$

$$\dots + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}\right)$$

Simplifying for one

$$\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\right) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}$$ { $\log ab = \log a + \log b$

$$= \log\left(2\pi\sigma^2\right)^{-\frac{1}{2}} - \frac{(x_1-\mu)^2}{2\sigma^2}$$ { $\log e$ cancels out each other

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}$$ { $\log a^b = b \log a$

$$= -\frac{1}{2}\left(\log 2\pi + \log \sigma^2\right) - \frac{(x_1-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{(x_1-u)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\log 2\pi - \frac{2}{2}\log \sigma - \frac{(x_1-u)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\log 2\pi - \log \sigma - \frac{(x_1-u)^2}{2\sigma^2}$$

Now applying this to n points:

$$\log(L(u,\sigma|x_1 x_2 \ldots x_n) = -\frac{1}{2}\log 2\pi - \log\sigma - \frac{(x_1-u)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \log\sigma - \frac{(x_2-u)^2}{2\sigma^2}$$

$$- \ldots - \frac{1}{2}\log 2\pi - \log\sigma - \frac{(x_n-u)^2}{2\sigma^2}$$

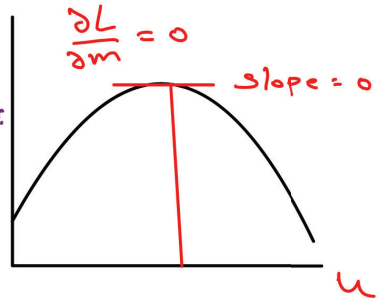
Adding these terms (total n terms $\therefore \frac{n}{2}\log 2\pi$)

Similarly $n\log\sigma$

$$\log(L) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{(x_1-u)^2}{2\sigma^2} - \frac{(x_2-u)^2}{2\sigma^2} - \ldots - \frac{(x_n-u)^2}{2\sigma^2}$$

Since we want to maximize likelihood, we try out infinite values for u and σ and find the best one. (So, we also maximize log likelihood to find maximum likelihood)

Therefore, we differentiate with u, to find the point where the slope is 0, and by doing so, we will get the maxima, which will gives us the maximum value of u with respect to L. Therefore, if u is max, than Likelihood will also be max.



Differentiating

$$\log(L) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{(x_1-u)^2}{2\sigma^2} - \frac{(x_2-u)^2}{2\sigma^2} - \ldots - \frac{(x_n-u)^2}{2\sigma^2}$$

No u $\therefore$ both will be 0

$$\frac{\partial \log(L)}{\partial u} = \frac{(x_1 - u)}{\sigma^2} + \frac{(x_2 - u)}{\sigma^2} + \cdots + \frac{(x_n - u)}{\sigma^2}$$

$$\frac{(x_1 - u)}{\sigma^2} + \frac{(x_2 - u)}{\sigma^2} + \cdots + \frac{(x_n - u)}{\sigma^2} = 0$$

$$(x_1 - u) + (x_2 - u) + \cdots + (x_n - u) = 0$$

$$x_1 + x_2 + x_3 + \cdots + x_n - nu = 0$$

$$nu = x_1 + x_2 + x_3 + \cdots + x_n$$

$$\boxed{u = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}}$$

We concluded that u of Normal distributed data will be the mean of the observed data points.

Same differentiation is applied for σ. $\therefore$ differentiating w.r.t. σ

$$\log(L) = -\underbrace{\frac{n}{2}\log 2\pi}_{=0} - n\log\sigma - \frac{(x_1-u)^2}{2\sigma^2} - \frac{(x_2-u)^2}{2\sigma^2} - \cdots - \frac{(x_n-u)^2}{2\sigma^2}$$

$$\frac{\partial \log(L)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{(x_1-u)^2}{\sigma^3} + \frac{(x_2-u)^2}{\sigma^3} + \cdots + \frac{(x_n-u)^2}{\sigma^3}$$

$$-\frac{n}{\sigma} + \frac{(x_1-u)^2}{\sigma^3} + \frac{(x_2-u)^2}{\sigma^3} + \cdots + \frac{(x_n-u)^2}{\sigma^3} = 0$$

$$n = \frac{(x_1-u)^2}{\sigma^2} + \frac{(x_2-u)^2}{\sigma^2} + \cdots + \frac{(x_n-u)^2}{\sigma^2}$$

$$\sigma^2 = \frac{(x_1-\mu)^2}{n} + \frac{(x_2-\mu)^2}{n} + \cdots + \frac{(x_n-\mu)^2}{n}$$

$$\therefore \boxed{\sigma^2 = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n}} \longrightarrow \text{This is Variance indeed.}$$

$$\boxed{\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n}}} \longrightarrow \text{This is Standard deviation}$$

This means, if you take standard deviation of all the observed data points, you will get the value of σ in normal distribution formula.

$\therefore N(\mu, \sigma) \longrightarrow \mu$: Mean, σ : Standard deviation

MLE in Machine learning (Steps)

1. **Find Data Distribution**
   Identify if it's binary (Bernoulli), continuous (Normal), etc.
2. **Choose Parametric ML Model**
   Must have fixed parameters like β_0, β_1
   - Parametric: Fixed structure (Linear Regression, Logistic Regression)
   - Non-parametric: No fixed structure (k-NN, Decision Trees)
3. **Initialize Parameter Values**
   Start with random β_0, β_1, β_2...
4. **Define Likelihood Function**
   For Bernoulli: p^k(1-p)^(1-k), for Normal: the exponential formula
5. **Find Maximum Values**
   Use calculus (derivatives = 0) or optimization algorithms
6. **Complete Model Training**
   Iterate until convergence

MLE literally trains the entire model by finding parameter values that make your training data most probable to have occurred. This gives you:

- Statistically optimal parameters
- Principled approach with theoretical backing
- Connection to modern loss functions (cross-entropy, MSE)

Parametric: Model that assumes that data has finite no. of parameters (input col)
Non parametric: Model that does not have a fixed structure, doesn't assume the distribution of the input cols ....

FAQ's

1. Is MLE a general concept applicable to all machine learning algorithms

Maximum Likelihood Estimation (MLE) is a general statistical concept that can be applied to many machine learning algorithms, particularly those that are parametric (i.e., defined by a set of parameters), but it's not applicable to all machine learning algorithms.

MLE is commonly used in algorithms such as linear regression, logistic regression, and neural networks, among others. These algorithms use MLE to find the optimal values of the parameters that best fit the training data.

However, there are some machine learning algorithms that don't rely on MLE. For example:

1. Non-parametric methods: Some machine learning methods, such as k-Nearest Neighbors (k-NN) and Decision Trees, are non-parametric and do not make strong assumptions about the underlying data distribution. These methods don't have a fixed set of parameters that can be optimized using MLE.

2. Unsupervised learning algorithms: Some unsupervised learning algorithms, like K-means clustering, use different objective functions, not necessarily tied to a probability distribution.

3. Reinforcement Learning: Reinforcement Learning methods generally don't use MLE, as they are more focused on learning from rewards and punishments over a sequence of actions rather than fitting to a specific data distribution.

2. How is MLE related to the concept of loss functions?

In machine learning, a loss function measures how well a model's predictions align with the actual values. The goal of training a machine learning model is often to find the model parameters that minimize the loss function.

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model to maximize the likelihood function, which is conceptually similar to minimizing a loss function. In fact, for many common models, minimizing the loss function is equivalent to maximizing the likelihood function.

MLE and the concept of loss functions in machine learning are closely related. Many common loss functions can be derived from the principle of maximum likelihood estimation under certain assumptions about the data or the model. By minimizing these loss functions, we're effectively performing maximum likelihood estimation.

3. Then why does loss function exist, why don't we maximize Likelihood

The confusion arises from the fact that we're using two different perspectives to look at the same problem.

In many machine learning algorithms, the aim is to minimize the difference between the predicted and actual values, and this is typically represented by a loss function. When we talk about minimizing the loss function, it's essentially the same as saying we're trying to find the best model parameters that give us the closest predictions to the actual values.

On the other hand, when we look at the problem from a statistical perspective, we talk in terms of maximizing the likelihood of seeing the observed data given the model parameters. This is represented by a likelihood function.

For many models, these two perspectives are equivalent - minimizing the loss function is the same as maximizing the likelihood function. In fact, many common loss functions can be derived from the principle of MLE under certain assumptions about the data.

So why do we often talk about minimizing the loss function instead of maximizing the likelihood?
There are a few reasons:

1. Computational reasons: It's often easier and more computationally efficient to minimize a loss function than to maximize a likelihood function. This is particularly true when working with complex models like neural networks.
2. Generalization: The concept of a loss function is more general and can be applied to a wider range of problems. Not all machine learning problems can be framed in terms of maximizing a likelihood. For example, many non-parametric methods and unsupervised learning algorithms don't involve likelihoods.
3. Flexibility: Loss functions can be easily customized to the specific needs of a problem. For instance, we might want to give more weight to certain types of errors, or we might want to use a loss function that is robust to outliers. In summary, while the concepts of loss function minimization and maximum likelihood estimation are closely related and often equivalent, the concept of a loss function is more flexible and computationally convenient, which is why it's more commonly used in the machine learning community.

4. Then why study about maximum likelihood at all?

The study of Maximum Likelihood Estimation (MLE) is essential for several reasons, despite the prevalence of loss functions in machine learning:

1. Statistical Foundation: MLE provides a strong statistical foundation for understanding machine learning models. It gives a principled way of deriving the loss functions used in many common machine learning algorithms, and it helps us understand why these loss functions work and under what assumptions.

2. Interpretability: The MLE framework gives us a way to interpret our model parameters. The MLEs are the parameters that make the observed data most likely under our model, which can be a powerful way of understanding what our model has learned.

3. Model Comparison: MLE gives us a way to compare different models on the same dataset. This can be done using tools like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which are based on the likelihood function and can help us choose the best model for our data.

4. Generalization to Other Methods: MLE is a specific case of more general methods, like Expectation-Maximization and Bayesian inference, which are used in more complex statistical modelling. Understanding MLE can provide a stepping stone to these more advanced topics.

5. Deeper Understanding: Lastly, understanding MLE can give us a deeper understanding of our models, leading to better intuition, better model selection, and ultimately, better performance on our machine learning tasks.
In short, while you can often get by with a practical understanding of loss functions and optimization algorithms in applied machine learning, understanding MLE can be extremely valuable for gaining a deeper understanding of how and why these models work.