



Maximum Likelihood Estimation -Conceptual understanding using an example



Shantanu Tripathi · [Follow](#)

Published in Analytics Vidhya

5 min read · Jan 13, 2020



Listen



Share



More

Maximum Likelihood Estimation (MLE) is one of the core concepts of Machine Learning. A lot of other Machine Learning algorithms/techniques are based on results derived using MLE. Therefore, it is always suggested to have a proper understanding of the concept and have a handy example within our pockets.

This story conveys the intent behind MLE and also provides us with an easy example that we can keep handy with us for future use.

An attempt has been made to reduce the use of Mathematics to the maximum extent possible. However, it is suggested that the reader has a basic understanding of differentiation and probability.

We will start with an example and use Maximum Likelihood Estimation (MLE) to get to the desired solution. With the help of the example, we will understand the concept of MLE.



**What is
 $p(\text{head})$?**

**Let's use MLE
to derive it!!**



Example

An unfair coin is tossed 5 times. The result of the 5 tosses is as follows: HHTTH. What is the probability of getting a head?

Solution:

$$p(\text{head}) = (\text{no. of times we get head}) / (\text{total number of tosses})$$

$$p(\text{head}) = \# \text{Heads} / \# \text{Tosses}$$

$$p(\text{head}) = 3/5 = 0.6$$

Well, the solution was pretty straightforward. However, another question arises at this point:

Why $p(\text{head}) = \# \text{Heads} / \# \text{Tosses}$????

Let's see if the Maximum Likelihood Estimation can help us answer the above-mentioned question.

Before we move:

1) Let us assume that $p(\text{head}) = \Theta$ and $p(\text{tail}) = 1 - \Theta$

2) Let us call our observed dataset as 'D'. $D = \text{HHTTH}$.

Maximum Likelihood to estimate the value of Θ i.e. $p(\text{head})$

So, according to our assumption, we know that the probability of getting a head is Θ .

Our end goal is to find a value of Θ , such that this value of Θ can help us claim that if we toss the coin 5 times, we will have observations similar to the observations in our Dataset D.

Reiterating the above-mentioned lines, we need to find (estimate) a Θ for the coin, such that, on repeating the same experiment of tossing the coin 5 times, the chances (likelihood) of getting D as the final outcome is maximized.

How do we find a Θ which increases the chances of getting D as the final outcome?

1. Firstly, we try to come up with a function that gives us the probability of getting D as the final outcome. This is known as the **likelihood function**.
2. Then, we take the log of the likelihood function to get the **log-likelihood function**. (This step is done for the mathematical reasons which are stated later. It is not a part of the real concept of Maximum Likelihood.)
3. Finally, we **maximize** this log-likelihood function to maximize the probability of getting D.

1) Finding Likelihood function:

Considering the fact that we already know that $p(\text{head})$ is Θ , the **likelihood** of getting dataset D as the final outcome is the **probability** of getting the same sequence (or number) of heads and tails as present in the dataset D.

Thus,

likelihood = $p(\text{getting same head-tail sequence as in D})$

Since $D = \text{HHTTH}$,

likelihood = $p(\text{head}).p(\text{head}).p(\text{tail}).p(\text{tail}).p(\text{head})$

Let n_h = #Heads = Number of heads

Let n_t = #Tails = Number of tails

$$\text{Likelihood (L)} = p(\text{head})^{n_h} \times p(\text{tail})^{n_t}$$
$$L(\theta) = \theta^{n_h} \times (1 - \theta)^{n_t}$$

Hence we have the likelihood function $L(\theta)$ as:

$$L(\theta) = \theta^{n_h} \times (1 - \theta)^{n_t}$$

likelihood function

2) Finding the Log-Likelihood Function:

Why do we take the log?

The likelihood function $L(\theta)$ has got θ raised to the power of #heads and $(1-\theta)$ raised to the power of #tails. Since $\theta < 1$, a huge value of #heads or #tails might cause our likelihood function to take very small floating-point values. To prevent this, we take the log of the likelihood function. This is known as the log-likelihood function.

There is yet another reason for taking the log. It is observed that differentiating log-likelihood functions is easier as compared to differentiating likelihood functions.

(This will become more evident towards the end of the story.)

$$\text{Log-likelihood (l)} = \log(L(\theta))$$
$$l(\theta) = \log(\theta^{n_h} \times (1 - \theta)^{n_t})$$
$$l(\theta) = \log(\theta^{n_h}) + \log((1 - \theta)^{n_t})$$
$$l(\theta) = n_h \times \log(\theta) + n_t \times \log((1 - \theta))$$

Hence we have the log-likelihood function $l(\theta)$ as:

$$l(\theta) = n_h \times \log(\theta) + n_t \times \log((1 - \theta))$$

log-likelihood function

3) Maximizing Log-Likelihood to estimate θ

We desired to find a likelihood function that can be maximized. However, we changed the likelihood function to log-likelihood. Now we are moving forward with the aim of maximizing log-likelihood.

Is maximizing the log-likelihood the same as maximizing likelihood?

It turns out yes!!

Since 'log' is an increasing function, the value of Θ that maximizes the log-likelihood function will also maximize the likelihood function.

How do we maximize the log-likelihood function $l(\Theta)$?

Maximizing log-likelihood means, finding a value of Θ such that the value of the log-likelihood function $l(\Theta)$ maximizes.

This is a simple maximization problem where we have been given an equation $l(\Theta)$ and we need to find a value of Θ which maximizes the equation.

We can differentiate $l(\Theta)$ w.r.t Θ and equate it to 0. This helps us in finding the value of Θ .

$$\frac{dl}{d\theta} = \frac{d}{d\theta} (nh \times \log(\theta) + nt \times \log((1 - \theta)))$$

$$\frac{dl}{d\theta} = \frac{nh}{\theta} - \frac{nt}{1 - \theta}$$

Equating $\frac{dl}{d\theta} = 0$, we get

$$\frac{nh}{\theta} = \frac{nt}{1 - \theta}$$

$$\theta = \frac{nh}{nh + nt}$$

Since, $nh + nt = \text{Total tosses}$

$$p(\text{head}) = \theta = \frac{nh}{\text{Total tosses}} = \frac{\#Heads}{\#Tosses}$$

Thus, using Maximum Likelihood Estimation, we estimated the value of $p(\text{head})$ i.e. Θ and found out that it is $\#Heads / \#Tosses$. This helps us answer “Why $p(\text{head}) = \#Heads / \#Tosses$?”.

Conclusion:

Maximum likelihood is used to estimate the value of a parameter ($p(\text{head})$ in our case).

1. We assume that we know the value of the parameter and call that value to be Θ
2. Using Θ , we find the likelihood function.
3. We take the log of the likelihood function for mathematical ease.

4. We maximize the log-likelihood function to get the value of Θ .

Machine Learning

Maximum Likelihood

Likelihood



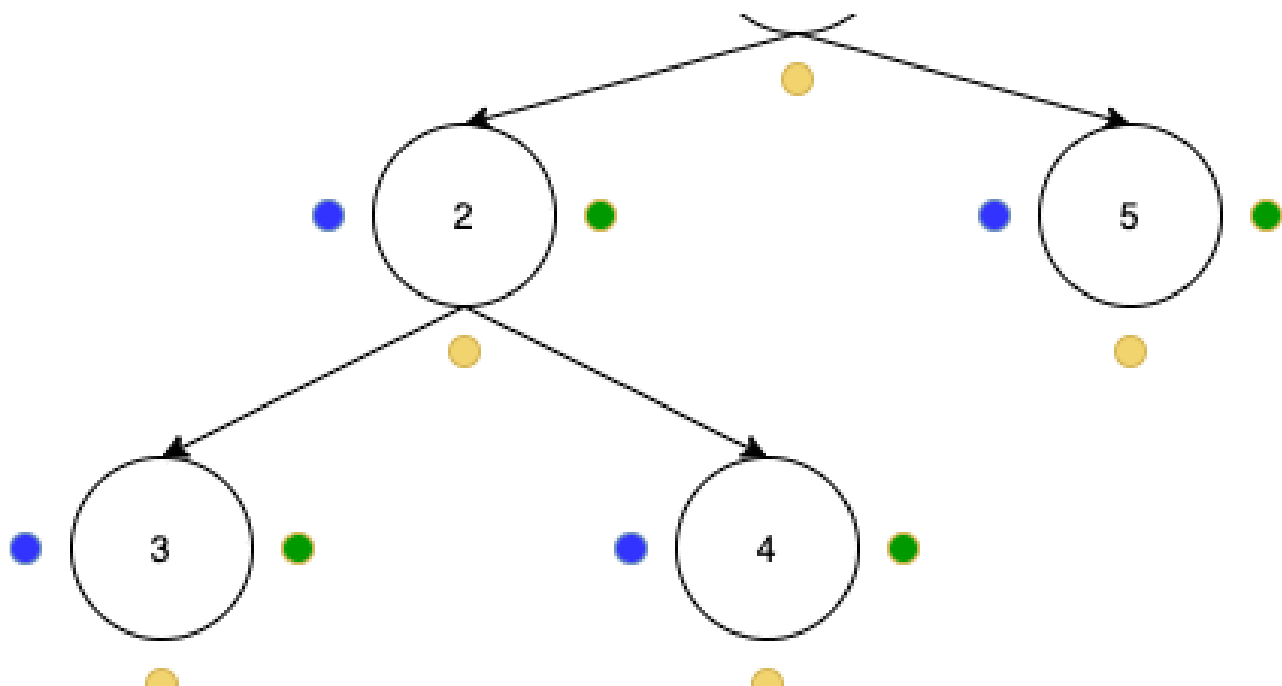
Follow


Written by Shantanu Tripathi

63 Followers · Writer for Analytics Vidhya

Deep Learning, NLP, Software dev etc. | NYU | Former SDE Intern at Amazon , AWS | Former SDE at CodeNation | Occasionally Philosophical | Mostly technical :p

More from Shantanu Tripathi and Analytics Vidhya



 Shantanu Tripathi in Analytics Vidhya