Maximum Likelihood vs. Bayesian Estimation

A comparison of parameter estimation met

Lulu Ricketts
Apr 20, 2021 11 min read

Hands-on Tutorials

Photo by kazuend on Unsplash

At its core, machine learning is about models. How
represent data? In what ways can we group data to
comparisons? What distribution or model does our

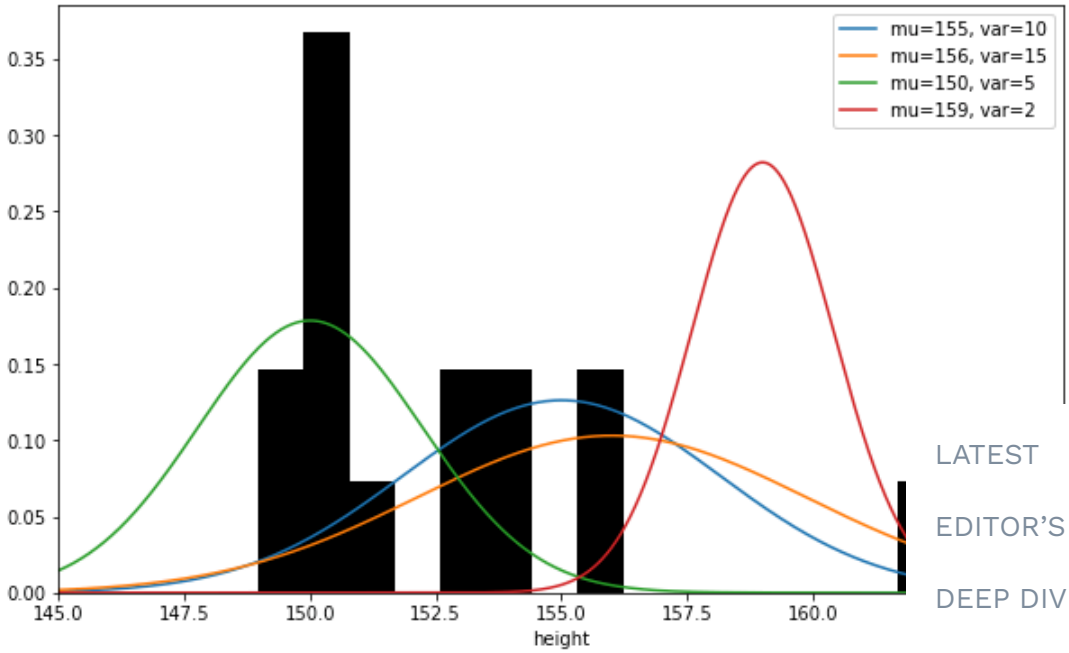
from? These questions (and many many more) drive data processes, but the latter is the basis of parameter estimation.

Maximum likelihood estimation (MLE), the frequentist view, and Bayesian estimation, the Bayesian view, are perhaps the two most widely used methods for parameter estimation, the process by which, given some data, we are able to estimate the model that produced that data. Why's this important? Data collected in the real world is almost never representative of the enti (imagine how much we would need to collect!), and estimating distribution parameters from an observe population, we can gain insight to unseen data.

As a prerequisite to this article, it is important that understand concepts in calculus and probability the joint and conditional probability, random variables, a probability density functions.

Parameter estimation deals with approximating *parc* distribution, meaning the type of distribution is typi beforehand, which determines what the unknown pa will be estimating are (λ for Poisson, μ and σ^2 for Ga example I use in this article will be Gaussian.

> Sample problem: Suppose you want to know the of tree's heights in a forest as a part of an longitu ecological study of tree health, but the only data you for the current year is a sample of 15 trees a recorded. The question you wish to answer is: "Wi distribution can we model the entire forest's tree



histogram of observed data (15 samples), and 4 examples of Gaussian curves that co
Image by author.

Quick note on notation

- θ is the unknown variables, in our Gaussian case
- *D* is all observed data, where _D = (x_1, x_2,...,x*n*

Likelihood Function

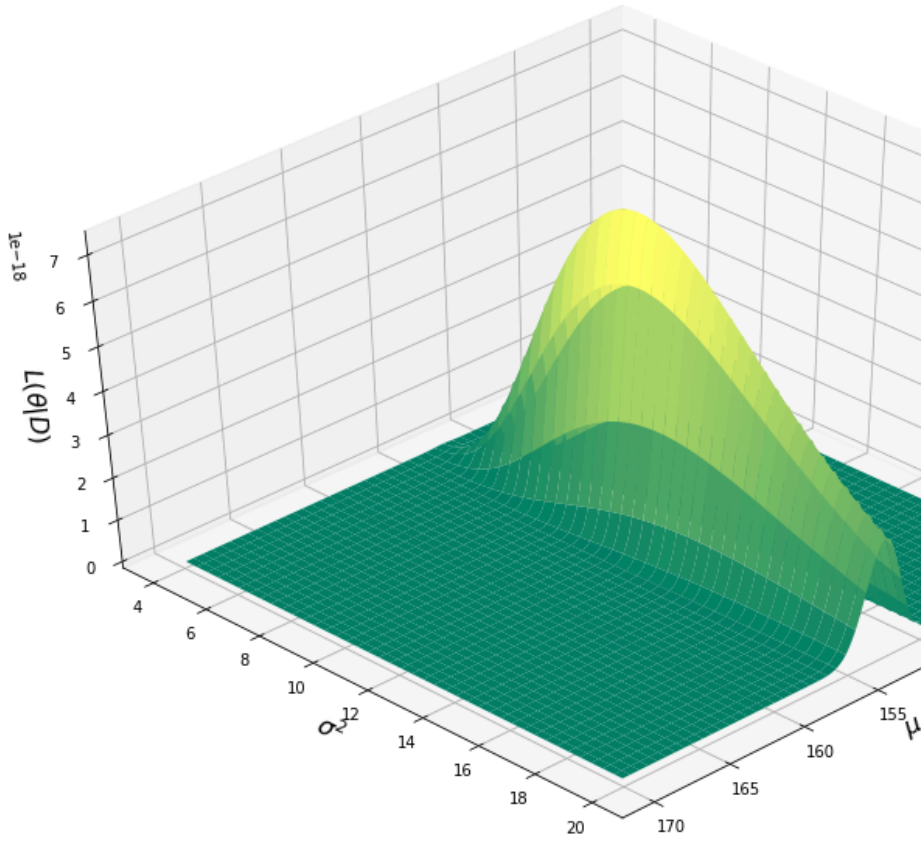
The (pretty much only) commonality shared by MLE estimation is their dependence on the **likelihood** of our case, the 15 samples). The likelihood describes t that each possible parameter value produced the da observed, and is given by:

$$L(\theta|D) = f(D|\theta) = \prod_{i=1}^{N} f$$

likelihood function. Image by author.

Thanks to the wonderful i.i.d. assumption, all data samples are considered independent and thus we are able to forgo messy conditional probabilities.

Let's return to our problem. All this entails is knowing the values of our 15 samples, what are the probabilities that each combination of our unknown parameters (μ,σ²) produced this set of data? By using the Gaussian distribution function, our likelihood function is:



likelihood function over μ,σ². Image by author.

Maximum Likelihood Estimation (MLE

Awesome. Now that you know the likelihood functio
the maximum likelihood solution is *really easy*. It's i
To get our estimated parameters (θ), all we have to
parameters that yield the maximum of the likelihoo

other words, what combination of (μ,σ^2) give us that brightest yellow point at the top of the likelihood function pictured above?

To find this value, we need to apply a bit of calculus and derivatives:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} f(x_i|\theta) = \frac{\partial}{\partial\theta} \prod_{i=1}^{N} f($$

(problematic) calculation of θ. Image by author.

As you may have noticed, we run into a problem. Tak
derivatives of products can get really complex and w
avoid this. Luckily, we have a way around this issue:
use the log likelihood function. Recall that (1) the log
is the sum of logs, and (2) taking the log of any func
change the values, but does not change *where* the n
that function occurs, and therefore will give us the s
solution.

log likelihood:

$$\ell(\theta) = \ln f(D|\theta) = \ln \prod_{i=1}^{N} f(x_i|\theta) = \sum_{i=1}^{N} \ln$$

$$\text{therefore, } \hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta)$$

correct calculation of θ using log likelihood. Image by author.

It turns out that for a Gaussian random variable, the
solution is simply the mean and variance of the obs
Therefore, for our problem, the MLE solution modeli
distribution of tree heights is a Gaussian distributior
μ=152.62 and σ^2=11.27.

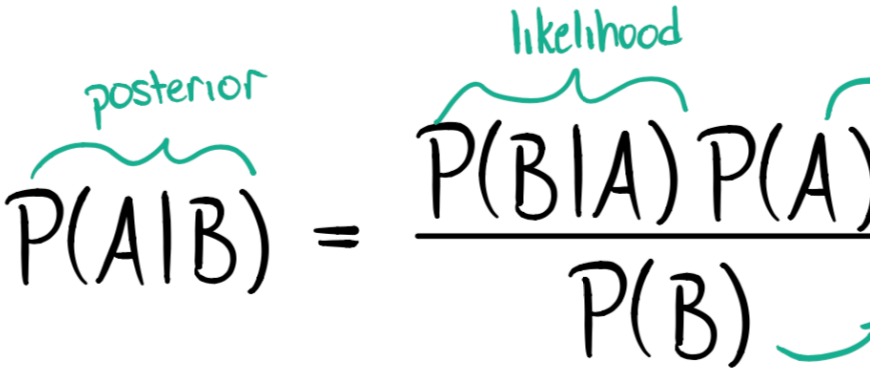
Bayesian Estimation

> Great news! To help you on your search for the distribution of tree heights in this forest, your coworker has managed to go into the data archives and dig up the mean of tree heights in the forest for the past 10 years. With this information, you can now additionally use Bayesian estimation to solve this problem.

The central idea behind Bayesian estimation is that seen any data, we already have some prior knowledg distribution it came from. Such prior knowledge usu from experience or past experiments. Before jumpin nitty gritty of this method, however, it is vitally impo the concept of Bayes' Theorem.

Bayes' Theorem

Hopefully you know, or at least heard of, Bayes' Theo probabilistic context, where we wish to find the prol one event conditioned on another event. Here, I hop in a way that'll give insight into Bayesian parameter and the significance of priors.



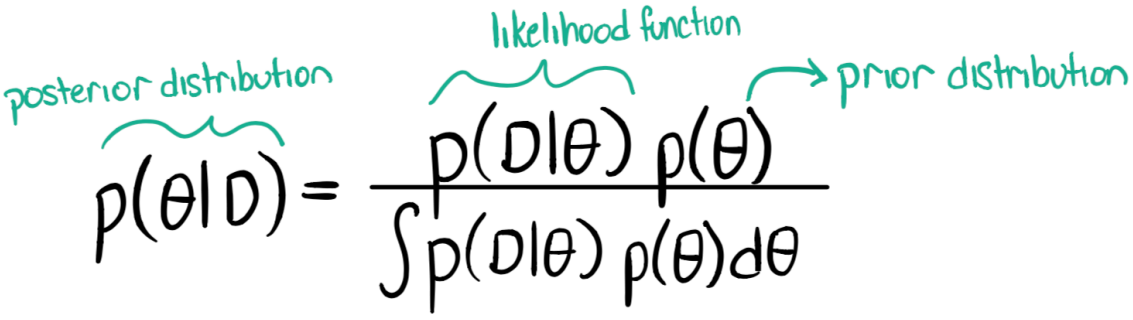
Bayes' Theorem. Image by author.

To illustrate this equation, consider the example that event *A = "it rained earlier today"*, and event *B = "the grass is wet"*, and we wish to calculate *P(A|B)*, the probability that it rained earlier given that the grass is wet. To do this, we must calculate *P(B|A)*, *P(B)*, and *P(A)*. The conditional probability *P(B|A)* represents the probability that the grass is wet given that it rained. In other words, it is the ***likelihood*** that the grass would be wet, given it is the case that it rained.

The value of *P(A)* is known as the ***prior***: the probabil
rained regardless of whether the grass is wet or not
knowing the state of the grass). This prior knowledg
because it determines how strongly we weight the li
we are somewhere that doesn't rain often, we woul
inclined to attribute wet grass to something other th
as dew or sprinklers, which is captured by a low *P(A*
However, if we were somewhere that constantly rain
probable that wet grass is a byproduct of the rain, a
will reflect that.

All that's left is *P(B)*, also known as the ***evidence***: th
that the grass is wet, this event acting as the eviden
fact that it rained. An important property of this valu
normalizing constant for the final probability, and as
soon see in Bayesian estimation, we substitute a no
factor in place of the traditional "evidence."

The equation used for Bayesian estimation takes on
form as Bayes' theorem, the key difference being tha
using models and probability density functions (pdfs
numerical probabilities.



Bayesian Estimation. Image by author.

Notice that first, the likelihood is equivalent to the l
used in MLE, and second, the evidence typically use
Theorem (which in this case would translate to *P(D)*
with an integral of the numerator. This is because (1)
extremely difficult to actually calculate, (2) *P(D)* doe
which is what we really care about, and (3) its usabi
normalizing factor can be substituted for the integra
which ensures that the integral of the posterior dist

Recall that to solve for parameters in MLE, we took
the log likelihood function to get numerical solution
Bayesian estimation, we instead compute a distribut
parameter space, called the ***posterior pdf***, denoted
distribution represents how strongly we believe eacl
value is the one that generated our data, after takin
account both the observed data and prior knowledg

The prior, *p(θ)*, is also a distribution, usually of the s
the posterior distribution. I won't get into the details
when the distribution of the prior matches that of th
it is known as a conjugate prior, and comes with ma
computational benefits. Our example will use conjug

Let's return to our problem concerning tree heights
time. In addition to the 15 trees recorded by the hike
have means for tree heights over the past 10 years.

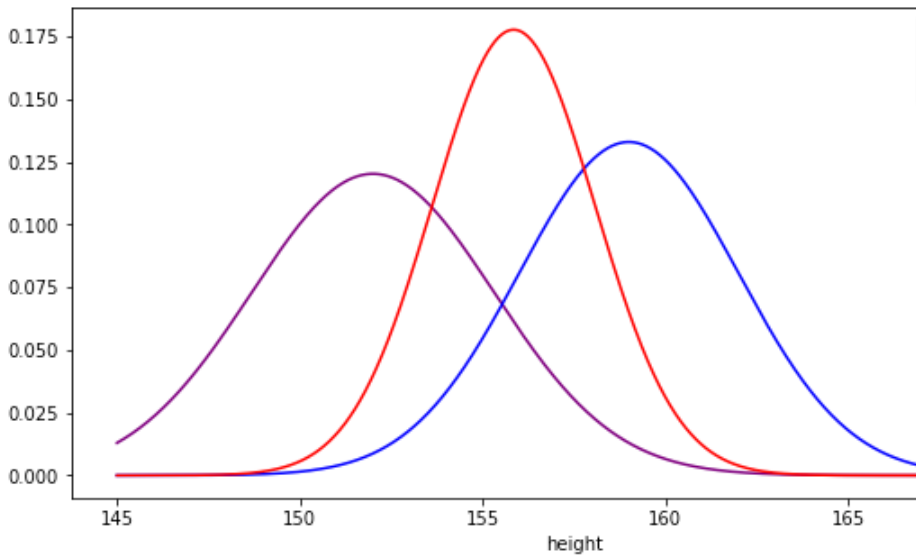
Prev. Year	1	2	3	4	5	6	7	8	9	10
-	160.5	153.8	157.3	160.5	164.0	156.4	157.1	158.1	163.5	160.6

tree heights for the previous 10 years. Image by author.

Following the assumption that this year, tree heights should fall into the distribution of all the previous year's, our prior is the Gaussian distribution with $\mu=159.2$ and $\sigma^2=9.3$.

All that's left is to calculate our posterior pdf. For th
calculation, I assume a fixed $\sigma^2 = \sigma^2\_MLE = 11.27$. In
would not solve Bayesian estimation this way, but m
Gaussians of different dimensions, like our likelihoo
extremely complex and I believe simplifying calculat
case is sufficient for understanding the process and
visualize. If you'd like more resources on how to exe
calculation, check out these two links.

Multiplying the univariate likelihood and prior and th
normalizing the result, we end up with a posterior G
distribution with $\mu=155.85$ and $\sigma^2=7.05$.



likelihood, prior, and posterior distributions. Image by author

And there you have the result of Bayesian estimatio
see, the posterior distribution takes into account bo

and likelihood to find a middle ground between them. In light of new observed data, the current posterior becomes the new prior, and a new posterior is calculated with the likelihood given by the novel data.

Predictions

We have models to describe our data, so what can v
them? The foremost usage of these models is to ma
predictions on unseen future data, which essentially
likely an observation is to have come from this distr
won't explicitly go through the calculations for our e
the formulas are below if you'd like to on your own.

MLE Prediction

$$p(x|D) = p(x|\hat{\theta}$$

MLE prediction. Image by author.

Maximum likelihood predictions utilize the predictio
latent variables in the density function to compute a
For instance, in the Gaussian case, we use the maxi
likelihood solution of (μ,σ^2) to calculate the predicti

Bayesian Prediction

$$p(x|D) = \int p(\theta|D)\,p(x|\theta)$$

Bayesian prediction. Image by author.

As you probably guessed, Bayesian predictions are a little more complex, using both the posterior distribution and the distribution over the random variable θ to yield the prediction of a new sample.

Concluding Remarks

When to use MLE? Bayesian estimation?

We've seen the computational differences between parameter estimation methods, and a natural questi When should I use one over the other? While there's fast rule when selecting a method, I hope you can u following questions as rough guidelines to steer you direction:

- **How much data are you working with?**

MLE, which depends solely on the outcomes of *obse* notorious for becoming easily biased when the data Consider an experiment where you flip a fair coin 3 t each flip comes up heads. While you know a fair coi up heads 50% of the time, the maximum likelihood you that *P(heads)* = 1, and *P(tails)* = 0. In situations v observed data is sparse, Bayesian estimation's incor prior knowledge, for instance knowing a fair coin is 5 help in attaining a more accurate model.

- **Do you have reliable prior knowledge about you**

As I just mentioned, prior beliefs can benefit your m certain situations. However, unreliable priors can lea slippery slope of highly biased models that require l of seen data to remedy. Make sure that if you are us they are well defined and ***contain relevant insight t***

you're trying to solve. If you are unsure about the reliability of your priors, MLE may be a better option, especially if you have a sufficient amount of data.

- **Are you limited in computational resources?**

A recurring theme in this article is that Bayesian computations are more complex than those for MLE. With modern computational power, this difference may be incons
however if you do find yourself constrained by resou
may be your best bet.

Some interesting properties

Finally, having already discussed many of the differe
between MLE and Bayesian estimation, I just want t
some interesting connections between these two m

1. If the Bayesian prior is uniform over all values (a
informative prior"), Bayesian predictions will be
not equal to, MLE predictions.
2. If the Bayesian prior is well-defined and non-zer
points, then, as the amount of observed data ap
infinity, MLE and Bayesian predictions will conve
same value.

That's all! If you got this far, thank you for reading. A
is appreciated!

- Link to code

WRITTEN BY

Lulu Ricketts

See all from Lulu Ricketts

Bayesian Statistics | Deep Dives | Hands On Tutorials | Parameter Estimation | Probability Distributions

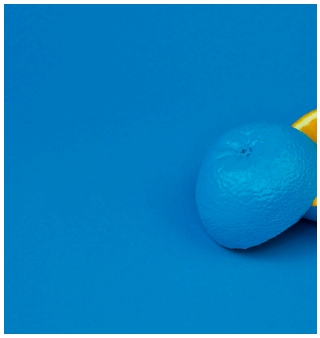
Share This Article

Related Articles

ARTIFICIAL INTELLIGENCE

How to Forecast Hierarchical Time Series

A beginner's guide to forecast reconciliation



DATA SCIENCE

Hands-on Time S Detection using A with Python

Dr. Robert Kübler

August 20, 2024 13 min read

Here's how to use Autoencoders to detect signals with anomalies in a few lines of...

Piero Paialunga

August 21, 2024 12 min read

DATA SCIENCE

Solving a Constrained Project Scheduling Problem with Quantum Annealing

Solving the resource constrained project scheduling problem (RCPSP) with D-Wave's hybrid constrained quadratic model (CQM)

Luis Fernando PÉREZ ARMAS, Ph.D.

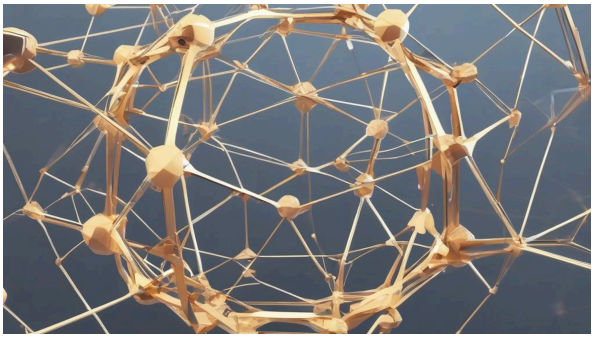
August 20, 2024 29 min read

Integrating LLM A LangChain into V

Learn how we use LLM improve and customis chatbot!

Ng Wei Cheng

August 20, 2024 17 min



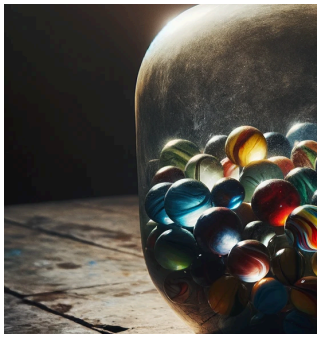
DATA SCIENCE

Towards Generalization on Graphs: From Invariance to Causality

This blog post shares recent papers on out-of-distribution generalization on graph-structured data

Qitian Wu

July 18, 2024 19 min read

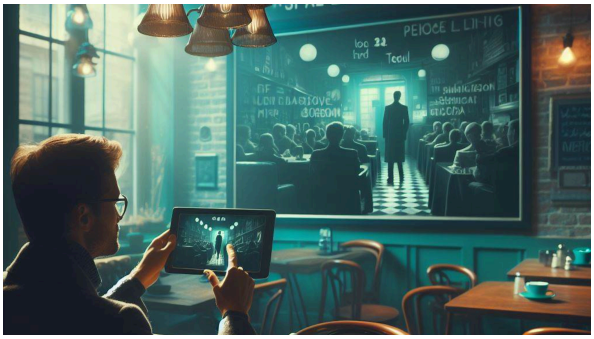


Monte Carlo Meth

Solving complex probl simulations

Hennie de Harder

February 15, 2024 19 mi



CINEMA

Evaluating Cinematic Dialogue - Which syntactic and semantic features are predictive of genre?

This article explores the relationship between a movie's dialogue and its genre, leveraging domain-driven data...

Christabelle Pabalan

January 20, 2024 18 min read