Likelihood vs. Probability

Asked 6 years, 11 months ago Active 5 years, 4 months ago Viewed 3k times

8

I have difficulties with *Likelihoods*. I do understand Bayes' Theorem

$$p(A|B,\mathcal{H}) = \frac{p(B|A,\mathcal{H})p(A|\mathcal{H})}{p(B|\mathcal{H})}$$

which can be directly deduced from applying $p(A,B) = p(B) \cdot p(A|B) = p(A)p(B|A) = p(B,A)$. Thus in my interpretation, the $p(\cdot)$ functions in Bayes Theorem are somehow all probabibilities, either marginal or conditional. So I have actually thought that Likelihood as a concept was more of a frequentist view of the inverse probability.

However, I have now repeatedly seen statements in *Bayesianists'* books that say that the likelihood is not a probability distribution. Reading MacKay's book yesterday, I stumbled over the following statement

> "[...] it is important to note that the terms likelihood and probability are not synonyms. The quantity $P(n_b|u,N)$ is afunction of both n_B and u. For fixed u, $P(n_b|u,N)$ defines a probability over n_B, for fixed n_B, $P(n_B|u,N)$ defines the likeihood of u."

- I understand this as follows: $p(A|B)$ is a probability of A under given B, thus a function $\text{probability} : \mathcal{A} \rightarrow [0,1]$. But considering a given value $a \in A$ and evaluating $p(A=a|B)$'s dependency on different $b \in \mathcal{B}$'s we are actually using a different function $L : \mathcal{B} \rightarrow [0,1]$.
- Is this interpretation correct?
- Can one then say that maximum likelihood methods could be motivated by the Bayesian theorem, where the prior is chosen to be constant?

probability likelihood

Share Cite Edit Follow Flag

asked Apr 4 '14 at 5:51
wirrbel
525 3 14

1 As an element of answer, I advice you the answer with links of Stephane Laurent in mathoverflow.net/questions/10971/.... Hope it helps. – peuhp Apr 4 '14 at 7:34

Add a comment

Start a bounty

1 Answer

Active Oldest Votes

8

I think maybe the best way to explain the notion of likelihood is to consider a concrete example. Suppose I have a sample of IID observations drawn from a Bernoulli distribution with unknown probability of success p: $X_i \sim \operatorname{Bernoulli}(p)$, $i = 1, \ldots, n$, so the joint probability mass function of the sample is

$$\Pr[\boldsymbol X = \boldsymbol x \mid p] = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}.$$

This expression also characterizes the likelihood of p, given an observed sample $\boldsymbol x = (x_1, \ldots, x_n)$:

$$L(p \mid \boldsymbol x) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}.$$

But if we think of p as a random variable, this likelihood is not a density:

$$\int_{p=0}^1 L(p \mid \boldsymbol x) \, dp \ne 1.$$

It is, however, *proportional* to a probability density, which is why we say it is a likelihood of p being a particular value given the sample--it represents, in some sense, the relative plausibility of p being some value for the observations we made.

For instance, suppose $n = 5$ and the sample was $\boldsymbol x = (1,1,0,1,1)$. Intuitively we would conclude that p is more likely to be closer to 1 than to 0, because we observed more ones. Indeed, we have

$$L(p \mid \boldsymbol x) = p^4 (1 - p).$$

If we plot this function on $p \in [0,1]$, we can see how the likelihood confirms our intuition. Of course, we do not know the true value of p--it could have been $p = 0.25$ rather than $p = 0.8$, but the likelihood function tells us that the former is much less likely than the latter. But if we want to determine a *probability* that p lies in a certain interval, we have to normalize the likelihood: since $\int_{p=0}^1 p^4(1-p) \, dp = \frac{1}{30}$, it follows that in order to get a *posterior density* for p, we must multiply by 30:

$$f_p(p \mid \boldsymbol x) = 30p^4(1-p).$$

In fact, this posterior is a beta distribution with parameters $a = 5, b = 2$. Now the areas under the density correspond to probabilities.

So, what we have essentially done here is applied Bayes' rule:

$$f_{\boldsymbol \Theta}(\boldsymbol \theta \mid \boldsymbol x) = \frac{f_{\boldsymbol X}(\boldsymbol x \mid \boldsymbol \theta)f_{\boldsymbol \Theta}(\boldsymbol \theta)}{f_{\boldsymbol X}(\boldsymbol x)}.$$

Here, $f_{\boldsymbol \Theta}(\boldsymbol \theta)$ is a *prior* distribution on the parameter(s) $\boldsymbol \theta$, the numerator is the likelihood $L(\boldsymbol \theta \mid \boldsymbol x) = f_{\boldsymbol X}(\boldsymbol x \mid \boldsymbol \theta) f_{\boldsymbol \Theta}(\boldsymbol \theta) = f_{\boldsymbol X, \boldsymbol \Theta}(\boldsymbol x, \boldsymbol \theta)$ which is also the joint distribution of $\boldsymbol X, \boldsymbol \Theta$, and the denominator is the marginal (unconditional) density of $\boldsymbol X$, obtained by integrating the joint distribution with respect to $\boldsymbol \theta$ to find the normalizing constant that makes the likelihood a probability density with respect to the parameter(s). In our numerical example, we implicitly took the prior for $f_{\boldsymbol \Theta}$ to be uniform on $[0,1]$. It can be shown that, for a Bernoulli sample, if the prior is $\operatorname{Beta}(a,b)$, the posterior for $f_{\boldsymbol \Theta}$ is also Beta, but with parameters $a^* = a + \sum x_i$, $b^* = b + n - \sum x_i$. We call such a prior *conjugate* (and refer to this as a Bernoulli-Beta conjugate pair).

Share Cite Edit Follow Flag

edited Nov 5 '15 at 9:24
Felipe
1 2

answered Apr 4 '14 at 7:52
heropup
4,606 1 12 21

Add a comment

Your Answer

Community wiki

Post Your Answer

Not the answer you're looking for? Browse other questions tagged probability likelihood or ask your own question.

Featured on Meta

State of the Stack Q1 2021 Blog Post

Should we replace the "data set request" with distinct "this is an off-topic..."

Related

29 Wikipedia entry on likelihood seems ambiguous
2 On the relation between conditional distribution and likelihood
2 How to calculate likelihood from Bayes' theorem without the normalisation?
13 A Measure Theoretic Formulation of Bayes' Theorem
1 Likelihood in Bayes theorem vs in MLE

Hot Network Questions

Блин, everyone?
Ancient temple booby traps designed for dragons
New DM on House Rules, concerning Nat20 & Rule of Cool
Hello, Permutations!
I don't understand why it is necessary to use a trigger on an oscilloscope for data acquisition
Idiom similar to "if the shoe fits"
Why does water weaken ion ion attractions?
Can a Lan Adapter cause a whole home network to crash?
How can I draw the trefoil knot in 3D co-ordinates in Latex using these parametric equations?
How to change some numbers in a font?
This ISS trash deployment looks more like 2 feet than 2 inches per second, was it too fast or are these articles incorrect?
Retaining permissions when copying a folder
Is a comment aligned with the element being commented a good practice?
Why might radios not be effective in a post-apocalyptic world?
Sun as water pump in underpressure water power circle
Using TikZ to draw a relation
Term to describe paradox where those with less subject matter expertise can sometimes make better teachers?
Self-Replicating Numbers
n-th Fibonacci number with memoization
Word for the animal providing motive power for a vehicle?
How did James Potter get his Invisibility Cloak?
Recommendations for OR video channels (YouTube etc)
Does either 'messy' or 'untidy' necessarily imply 'dirty'?
How long would it take for inbreeding issues to arise for a family that practiced inbreeding?

Question feed