Cross Validated

Why Normalizing Factor is Required in Bayes Theorem?

Asked 4 years, 11 months ago Active 3 years, 5 months ago Viewed 10k times

20

6

Bayes theorem goes

$$P(\text{model}|\text{data}) = \frac{P(\text{model}) \times P(\text{data}|\text{model})}{P(\text{data})}$$

This is all fine. But, I've read somewhere:

> Basically, P(data) is nothing but a normalising constant, i.e., a constant that makes the posterior density integrate to one.

We know that $0 \leq P(\text{model}) \leq 1$ and $0 \leq P(\text{data}|\text{model}) \leq 1$.

Therefore, $P(\text{model}) \times P(\text{data}|\text{model})$ must be between 0 and 1 as well. In such a case, **why do we need a normalizing constant to make the posterior integrate to one?**

probability bayesian conditional-probability bayes

edited Dec 19 '14 at 7:27
Matt Krause
16.1k 2 47 90

asked Dec 18 '14 at 21:38
Sreejith Ramakrishnan
309 1 2 4

4 When you are working with probability *densities*, as mentioned in this post, you can no longer conclude `0 <= P(model) <= 1` nor `0 <= P(data/model) <= 1`, because either (or even both!) of those could exceed 1 (and even be infinite). See stats.stackexchange.com/questions/4220. – whuber ♦ Dec 18 '14 at 22:35

1 It is not the case that

$$P(\text{data}|\text{model}) \leq 1$$

because this vague notation represents the integrated likelihood of the data, not a probability. – Xi'an Dec 19 '14 at 11:29

3 Answers

15

First, *the integral of "likelihood x prior" is not necesserely 1*.

It is not true that if:

$0 \leq P(\text{model}) \leq 1$ and $0 \leq P(\text{data}|\text{model}) \leq 1$

then the integral of this product with respect to the model (to the parameters of the model, indeed) is 1.

Demonstration. Imagine two discrete densities:

$$P(\text{model}) = [0.5, 0.5] \text{ (this is called "prior")}$$
$$P(\text{data} \mid \text{model}) = [0.80, 0.2] \text{ (this is called "likelihood")}$$

If you multiply them both you get:

$$[0.40, 0.25]$$

which is not a valid density since it does not integrate to one:

$$0.40 + 0.25 = 0.65$$

So, what should we do to force the integral to be 1? Use the normalizing factor, which is:

$$\sum_{\text{model\_params}} P(\text{model})P(\text{data} \mid \text{model}) = \sum_{\text{model\_params}} P(\text{model, data}) = P(\text{data}) = 0.65$$

(sorry about the poor notation. I wrote three different expressions for the same thing since you might see them all in the literature)

Second, *the "likelihood" can be anything, and even if it is a density, it can have values higher than 1*.

As @whuber said this factors do not need to be between 0 and 1. They need that their integral (or sum) be 1.

Third [extra], *"conjugates" are your friends to help you find the normalizing constant*.

You will often see:

$$P(\text{model}|\text{data}) \propto P(\text{data}|\text{model})P(\text{model})$$

because the missing denominator can be easily get by integrating this product. Note that this integration will have one well known result if the prior and the likelihood are *conjugate*.

edited Dec 19 '14 at 8:35

answered Dec 19 '14 at 8:20
alberto
2,346 10 35

+1. This is the only answer that actually addresses the original question of why the normalization constant is needed *to make the posterior integrate to one*. What you do with the posterior later (e.g. MCMC inference or calculating absolute probabilities) is a different matter. – Pedro Mediano May 30 '17 at 13:07

I have difficulties following the notation. What does it mean $P(model) = [0.5, 0.5]$? Suppose we use a Gaussian model with given variance $\sigma^2 = 1$, and we wish to find the mean μ. What does it mean $P(\mu) = [0.5, 0.5]$? – rtrtrt Apr 27 at 9:57

I tried to follow the notation of the OP. I mean that you have two possible values for your model parameter, and that both are equally likely. For instance, is there a teapot revolving around the Earth? yes/no? In your case it does not make sense because your parameter μ is continuous and can take infinite values. – alberto May 4 at 19:01

12

The short answer to your question is that without the denominator, the expression on the right-hand side is merely a *likelihood*, not a *probability*, which can only range from 0 to 1. The "normalizing constant" allows us to get the probability for the occurrence of an event, rather than merely the relative likelihood of that event compared to another.

edited Dec 18 '14 at 22:21
Horst Grünbusch **4,512** 14 21

answered Dec 18 '14 at 22:12
heropup **3,948** 1 9 17

8

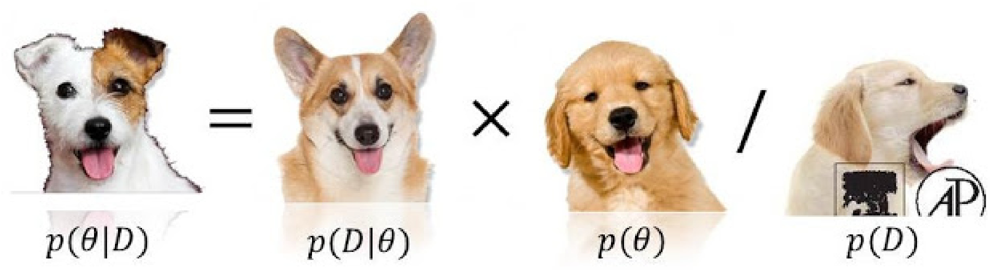
You already got two valid answers but let me add my two cents.

Bayes theorem is often defined as:

$$P(\text{model} \mid \text{data}) \propto P(\text{model}) \times P(\text{data} \mid \text{model})$$

because the only reason why you need the constant is so that it integrates to 1 (see the answers by others). This is not needed in most MCMC simulation approaches to Bayesian analysis and hence the constant is dropped from the equation. So for most simulations it is *not* even required.

I *love* the description by Kruschke: the last puppy (constant) is sleepy because he has nothing to do in the formula.



Also some, like Andrew Gelman, consider the constant as "overrated" and "basically meaningless when people use flat priors" (check the discussion here).

edited Jun 4 '16 at 11:47

answered Dec 19 '14 at 8:46
Tim ♦ **71.3k** 14 152 253

9 +1 to the introduction of puppies. "No animals were harmed in the writing of this answer" :) – alberto Dec 19 '14 at 8:50