MLE, MAP and Bayesian Inference

Grasp the idea of Bayesian inference by focusing on the difference from MLE and MAP

Shota Horii Sep 21, 2019 · 7 min read ★

Photo by Aysha Begum on Unsplash

MLE/MAP and Bayesian inference.

In this article, I'm going to introduce Bayesian inference by focusing on the difference between MLE/MAP and Bayesian inference.

Note: Preliminary knowledge of MLE and MAP is assumed in this article. If you're not familiar with those methods, please refer to the following article.

. . .

The difference between MLE/MAP and Bayesian inference

Let's start from the recap of MLE and MAP.
Given the observed data D, estimations of a probabilistic model's parameter θ by MLE and MAP are the following.

$$\hat{\theta}_{ML} = argmax_\theta \{P(D|\theta)\}$$

$$\hat{\theta}_{MAP} = argmax_\theta \{P(\theta|D)\} = argmax_\theta \{\frac{P(D|\theta)P(\theta)}{P(D)}\} = argmax_\theta \{P(D|\theta)P(\theta)\}$$

MLE gives you the value which maximises the Likelihood $P(D|\theta)$. And MAP gives you the value which maximises the posterior probability $P(\theta|D)$. As both methods give you a single fixed value, they're considered as **point estimators**.

On the other hand, Bayesian inference fully calculates the posterior probability distribution, as below formula. Hence the output is not a single value but a probability density function (when θ is a continuous variable) or a probability mass function (when θ is a discrete variable).

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

This is the difference between MLE/MAP and Bayesian inference. MLE and MAP returns a single fixed value, but Bayesian inference returns probability density (or mass) function.

But why we even need to fully calculate the distribution, when we have MLE and MAP to determine the value of θ ? To answer this question, let's see the case when MAP (and other point estimators) doesn't work well.

. . .

A case when MAP (or point estimators in general) doesn't work well

Assume you're in a casino with full of slot machines with *50%* winning probability. After playing for a while, you heard the rumour that there's one special slot machine with *67%* winning probability.
Now, you're observing people playing 2 suspicious slot machines (you're sure that one of those is the special slot machine!) and got the following data.

Machine A: 3 wins out of 4 plays
Machine B: 81 wins out of 121 plays

By intuition, you would think *machine B* is the special one. Because 3 wins out of 4 plays on *machine A* could just happen by chance. But *machine B*'s data doesn't look like happening by chance.

But just in case, you decided to estimate those 2 machines' winning probabilities by MAP with hyperparameters $\alpha=\beta=2$. (Assuming that the results (k wins out of n plays) follow binomial distribution with the slot machine's winning probability θ as its parameter.)

The formula and results are below.

$$\hat{\theta}_{MAP} = \frac{k+\alpha-1}{}$$

Machine A: (3+2–1)/(4+2+2–2) = 4/6 = ***66.7%***
Machine B: (81+2–1)/(121+2+2–2) = 82/123 = ***66.7%***

Unlike your intuition, estimated winning probability θ by MAP for the 2 machines are exactly same. Hence, by MAP, you cannot determine which one is the special slot machine.

But really? Isn't it looking obvious that *Machine B* is more likely to be the special one?

. . .

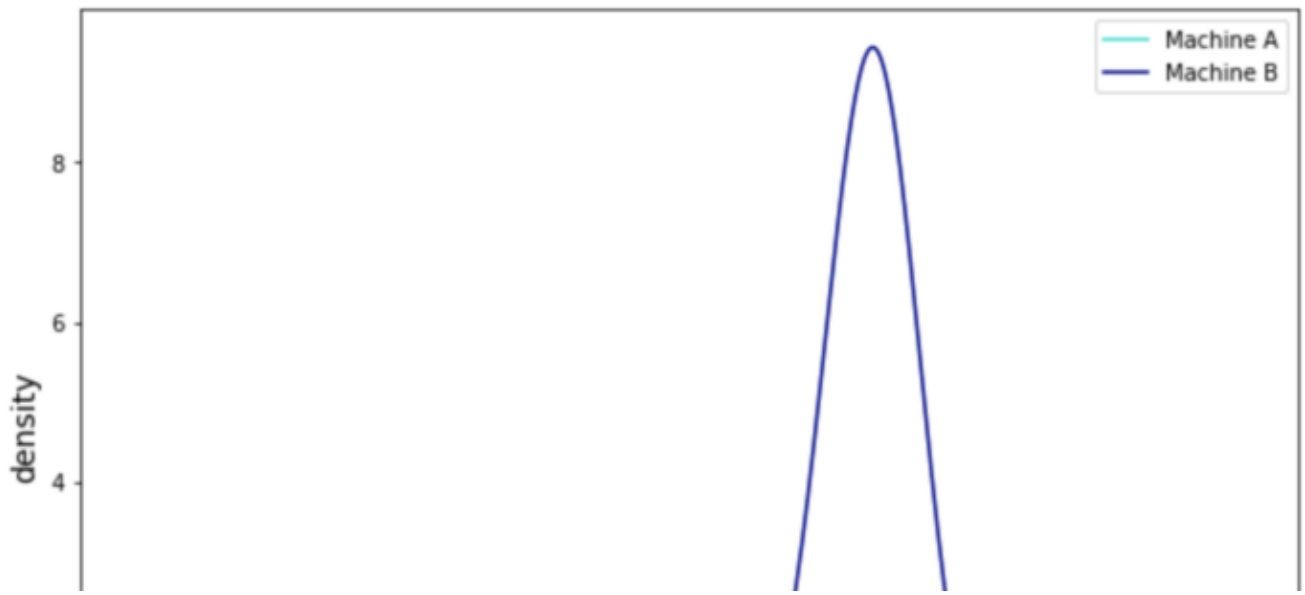
Bayesian Inference

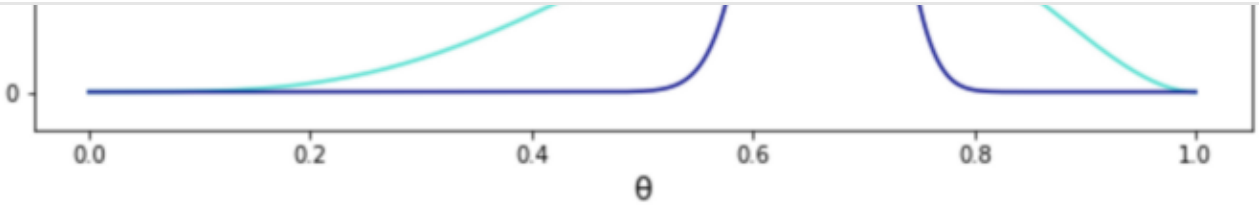
To see if there really be no difference between *machine A* and *machine B*, let's fully calculate the posterior probability distribution, not only MAP estimates.

In the case above, the posterior probability distribution $P(\theta|D)$ is calculated as below. (Detailed computation will be covered in the next section.)

$$P(\theta|D) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

And $P(\theta|D)$ for *machine A* and *machine B* are drawn as below.





Although both distributions have their ***mode*** on $\theta=0.666\ldots$ (that's why their MAP estimates are the same value), the shapes of the distributions are quite different. Density around the mode is much higher in the distribution of *machine B* than the one of *machine A*. This is why you want to calculate full distribution, not only the MAP estimate.

. . .

Computation of Bayesian Inference

As we skipped the computation of $P(\theta|D)$ in the previous section, let's go through the detailed calculation process in this section.

Both MAP and Bayesian inference are based on Bayes' theorem. The computational difference between Bayesian inference and MAP is that, in Bayesian inference, we need to calculate $P(D)$ called **marginal likelihood** or **evidence**. It's the denominator of Bayes' theorem and it assures that the integrated value* of $P(\theta|D)$ over all possible θ becomes 1. (* Sum of $P(\theta|D)$, if θ is a discrete variable.)

$P(D)$ is obtained by marginalisation of joint probability. When θ is a continuous variable, the formula is as below.

$$P(D) = \int_{\theta} P(D, \theta)\, d\theta$$

Considering the chain rule, we obtain the following formula.

$$P(D) = \int_{\theta} P(D|\theta)P(\theta)d\theta$$

Now, put this into the original formula of the posterior probability distribution. Calculating below is the goal of Bayesian Inference.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\int_\theta P(D|\theta)P(\theta)d\theta}$$

. . .

Let's calculate $P(\theta|D)$ for the case above.

Beginning with $P(D|\theta)$ — **Likelihood** — which is the probability that data D is observed when parameter θ is given. In the case above, D is "*3 wins out of 4 matches*", and parameter θ is the winning probability of *machine A*. As we assume that the number of wins follows binomial distribution, the formula is as below, where n is the number of matches and k is the number of wins.

$$P(D|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Then $P(\theta)$ — the **prior probability distribution** of θ — which is the probability distribution expressing our prior knowledge about θ. Here, specific probability distributions are used corresponding to the probability distribution of Likelihood $P(D|\theta)$. It's called conjugate prior distribution.

Since the conjugate prior of binomial distribution is Beta distribution, we use Beta distribution to express $P(\theta)$ here. Beta distribution is described as below, where α and β are hyperparameters.

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$P(D\,|\,\theta)P(\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

Then, *P(D)* — the denominator of the formula — is calculated as follows. Note that the possible range of θ is $0 \leq \theta \leq 1$.

$$P(D) = \int_\theta P(D|\theta)P(\theta)d\theta$$

$$= \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$

$$= \binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}d\theta$$

With Euler integral of the first kind, the above formula can be deformed to below.

$$P(D) = \binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)}$$

Finally, we can obtain *P(θ|D)* as below.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$= \frac{\binom{k}{k}\Gamma(\alpha)\Gamma(\beta)}{\binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)}}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

. . .

Expected A Posteriori (EAP)

As you may have noticed, the estimate by MAP is the ***mode*** of the posterior distribution. But we can also use other statistics for the point estimation, such as ***expected value*** *of* $\theta|D$. The estimation using the expected value of $\theta|D$ is called **Expected A Posteriori**.

$$\hat{\theta}_{EAP} = E[\theta|D]$$

$$= \int_{\theta} \theta\, P(\theta|D) d\theta$$

. . .

Let's estimate the winning probability of the 2 machines using EAP. From the discussion above, $P(\theta|D)$ in this case is below.

$$P(\theta|D) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

Thus the estimate is described as below.

$$\hat{\theta}_{EAP} = \int_{\theta} \theta P(\theta|D) d\theta$$

$$= \int_0^1 \theta \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)} \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} d\theta$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)} \int_0^1 \theta^{k+\alpha}(1-\theta)^{n-k+\beta-1} d\theta$$

With Euler integral of the first kind and the definition of Gamma function, above formula can be deformed to below.

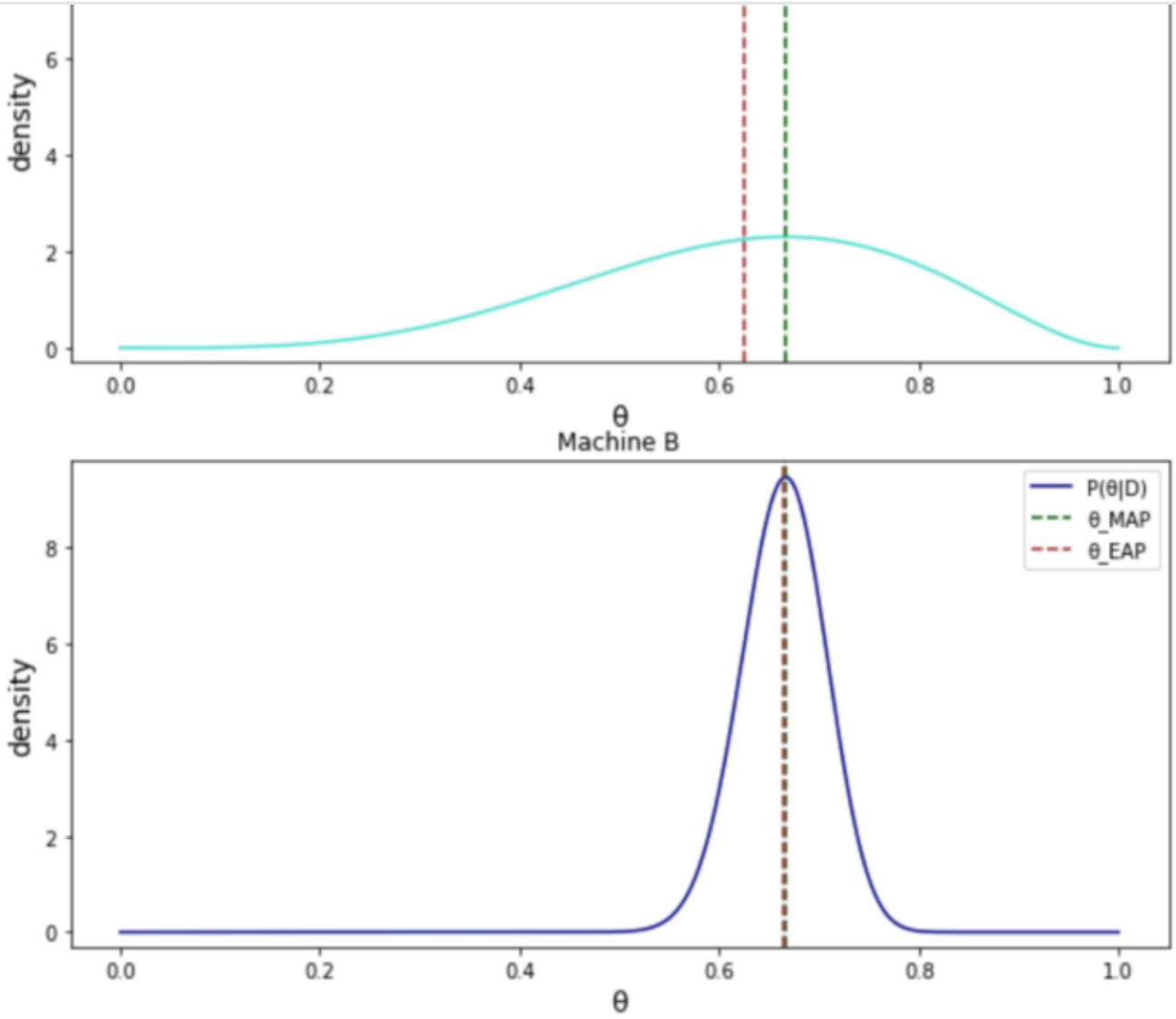
$$\hat{\theta}_{EAP} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)} \frac{\Gamma(k+\alpha+1)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta+1)}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)} \frac{(k+\alpha)\Gamma(k+\alpha)\Gamma(n-k+\beta)}{(n+\alpha+\beta)\Gamma(n+\alpha+\beta)}$$

$$= \frac{k+\alpha}{n+\alpha+\beta}$$

Hence, EAP estimate of 2 machines' winning probabilities with hyperparameters $\alpha=\beta=2$ are below.

Machine A: (3+2)/(4+2+2) = 5/8 = ***62.5%***
Machine B: (81+2)/(121+2+2) = 83/125 = ***66.4%***



. . .

Conclusion

As seen above, Bayesian inference provides much more information than point estimators like MLE and MAP. However, it also has a drawback — the complexity of its integral computation. The case in this article was quite simple and solved analytically, but it's not always the case in real-world applications. We then need to use MCMC or other algorithms as a substitute for the direct integral computation.
Hope this article helped you to understand Bayesian inference.

Machine Learning Statistics Bayesian Statistics