18.650: Fundamentals of Statistics November 6, 2023

Lecture 21 — Bayesian inference, cont'd.

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Summarizing the posterior

Recall the following key formula from Lecture 20:

Definition 1.1: Posterior

The posterior $f(\theta \mid X_1, \ldots, X_n)$ is the density which is proportional to the prior f times the likelihood L_n:

$$f(\theta \mid X_1, \ldots, X_n) \propto L_n(\theta) f(\theta).$$

We should keep in mind Figure 1: we get the posterior by updating our prior after having observed the data $X_1, \ldots, X_n$. Observing the data reduces our uncertainty in θ, so the spread of the posterior is narrower than the spread of the prior. In fact, one can show that the standard deviation of the posterior distribution decreases as $1/\sqrt{n}$ with the number of samples n.

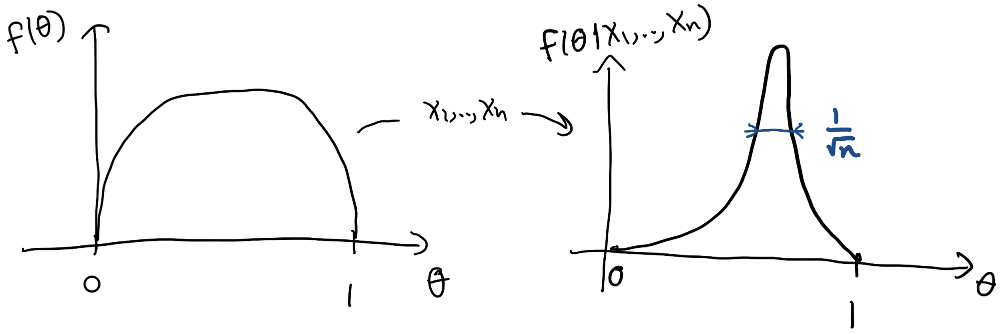

Figure 1: The posterior $f(\theta \mid X_1, \ldots, X_n)$ is an update to the prior $f(\theta)$ after having observed the data $X_1, \ldots, X_n$.

The posterior fully captures all we know about θ^*. However, we often want just a single estimator of θ^*, rather than a whole distribution. There are two commonly used estimators to summarize the posterior.

1.1 Bayes estimator (mean)

Definition 1.2: Bayes Estimator

The Bayes estimator is the mean of the posterior, or mean *a posteriori*:

$$\hat{\theta}^{\text{Bayes}} = \int \theta f(\theta | X_1, \dots, X_n) d\theta.$$

The mean can be hard to compute explicitly because it requires knowing the normalizing constant of $f(\theta|X_1, \dots, X_n)$. However, we can *approximately* compute the mean using Markov Chain Monte Carlo (MCMC), a popular numerical method.

1. **Markov chain step**: draw $\theta_1, \dots, \theta_T$ which are approximately iid from $f(\theta \mid X_1, \dots, X_n)$, by simulating a Markov chain (a sequence of random variables).

2. **Monte Carlo step**: compute

$$\hat{\theta} = \frac{1}{T} \sum_{t=1}^{T} \theta_t \approx \int \theta f(\theta | X_1, \dots, X_n) d\theta.$$

1.2 MAP (mode)

Definition 1.3: MAP

The maximum *a posteriori*, or MAP is the mode of the posterior:

$$\hat{\theta}^{\text{MAP}} = \text{argmax}_\theta\, f(\theta \mid X_1, \dots, X_n).$$

Note that the mode $\hat{\theta}^{\text{MAP}}$ is not necessarily close to the mean $\hat{\theta}^{\text{Bayes}}$. This can happen if e.g. the posterior is bimodal, as shown in the righthand plot in Figure 2.

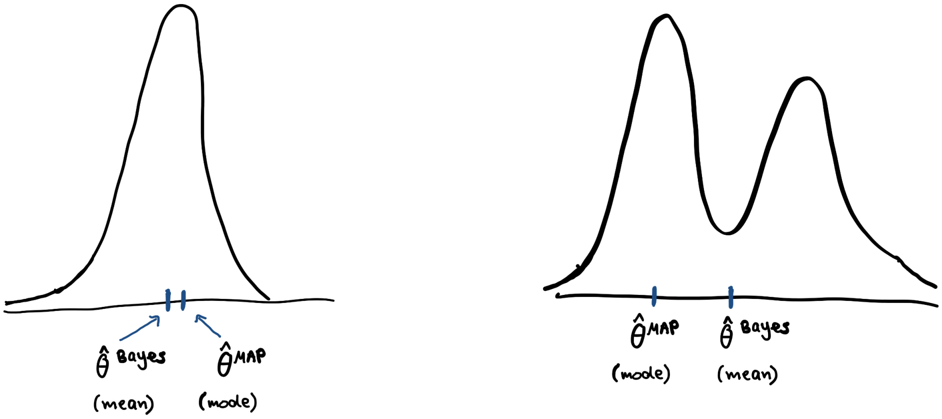

Figure 2: The mode and the mean of the posterior are sometimes close to one another, but can also be far away, e.g. if the posterior is bimodal (right).

To compute $\hat{\theta}^{\mathrm{MAP}}$, we can maximize the log posterior, which takes the form

$$\log f(\theta \mid X_1, \ldots, X_n) = \ell_n(\theta) + \log f(\theta) - \log c_n.$$

Recall that c_n is the unknown normalizing constant. However, when we set the theta derivative of $\log f(\theta \mid X_1, \ldots, X_n)$ to zero, the $\log c_n$ vanishes. If it is not possibly to find the maximum by hand, we can run gradient ascent. The gradient ascent updates will take the form

$$\theta^{(k+1)} = \theta^{(k)} + \eta_k(\nabla \ell_n(\theta) + \nabla \log f(\theta)).$$

Here, too, the log of the normalizing constant vanishes.

1.3 Measuring uncertainty: posterior interval

An estimator only gives us a single number to approximate the ground truth. It's better to have a whole interval in which the true parameter lies with high probability. In frequentist inference, we constructed a *confidence* interval for this purpose. In Bayesian inference, the analogue of a confidence interval is the posterior interval.

Definition 1.4: Posterior interval

A $(1-\alpha)$-posterior interval is an interval $[a, b]$ such that the probability (area) under the posterior curve between a and b is $1 - \alpha$. In other words,

$$\int_a^b f(\theta | X_1, \ldots, X_n) d\theta = 1 - \alpha.$$

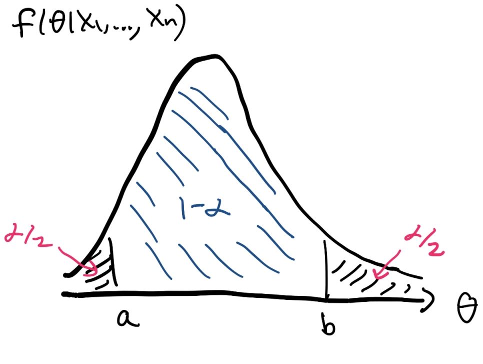

Figure 3: A $1 - \alpha$ posterior interval.

2 How to choose the prior

There are several considerations we can use to choose the prior.

1. *True prior knowledge.*

2. *Convenient calculation:* by choosing a conjugate prior, we can ensure the posterior lies in the same family as the prior. This is useful if there are known formulas for the mean and mode of probability distributions in this family, since we don't have to resort to MCMC or gradient ascent. For example, we saw in Lecture 20 that Beta prior → Beta posterior, if we observe Bernoulli data. In Section 3 below, we will also see that Gaussian prior → Gaussian posterior, if we observe Gaussian data.

3. *No knowledge: uninformative prior.* If we have no prior belief about θ, then we can choose a uniform prior, e.g. $\text{Unif}([0,1])$ in the kiss example. More generally, if the parameter space is $\Theta = [a,b]$ then we can take f to be $\text{Unif}([a,b])$, which has pdf $f(\theta) = \frac{1}{b-a}$ for $\theta \in [a,b]$.

   If Θ is the whole real line, then there's no way to get a valid uniform distribution because $\int_{-\infty}^{\infty} 1 d\theta = \infty$. However, it's still valid to take $f(\theta) = 1$. This is called an "**improper prior**", because the constant function 1 is not a valid probability distribution on the real line. Nevertheless, it leads to a valid posterior distribution:

$$f(\theta \mid X_1, \ldots, X_n) = \frac{L_n(\theta)}{\int L_n(\theta) d\theta}.$$

Remark.

> Remember that $f(\theta)$ weights the likelihood $L_n(\theta)$. When $f(\theta)$ is the constant 1, then we're not weighting L_n at all. Therefore, $\hat{\theta}^{\text{MAP}} = \hat{\theta}^{\text{MLE}}$. In words, the MAP (maximum of the posterior) equals the MLE (maximum of the likelihood). On the other hand, the mean of the posterior $\hat{\theta}^{\text{Bayes}}$ is different from both the MAP and the MLE, as the next example shows.

Example.

> Recall from the kiss example that we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$. Suppose we take an uninformative prior, the uniform distribution $f(p) = \text{Unif}([0,1])$. Then the posterior is proportional to the likelihood:
>
> $$f(p \mid X_1, \ldots, X_n) \propto p^{\sum_i X_i}(1-p)^{n - \sum_i X_i}, \quad p \in [0,1].$$
>
> We recognize that this is the density of the Beta distribution:
>
> $$\text{posterior} = \text{Beta}\left(1 + \sum_i X_i,\ n + 1 - \sum_i X_i\right).$$
>
> It's known that the mean of Beta(a,b) is $a/(a+b)$, so we get
>
> $$\hat{\theta}^{\text{Bayes}} = \frac{1 + \sum_i X_i}{n+2} = \frac{\bar{X}_n + n^{-1}}{1 + 2n^{-1}}.$$
>
> .
>
> This is different from the MLE $\bar{X}_n$ (also the MAP), although for large n, the Bayes estimator becomes close to the MLE. Note that $\hat{\theta}^{\text{Bayes}}$ is a biased estimator of p, since we have $\mathbb{E}[\hat{\theta}^{\text{Bayes}}] = \frac{1+np}{n+2} \neq p$.

Remark.

> The uniform prior is not "transformation invariant": if we change variables then a uniform prior becomes non-uniform. The so-called Jeffery's prior is transformation invariant, on the other hand. See Section 11.6 in AoS for more on this.

3 The posterior for a Gaussian prior and Gaussian data

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ and suppose our prior distribution is $\theta \sim \mathcal{N}(0, \sigma^2)$. The pdf f is then

$$f(\theta) \propto e^{-\frac{\theta^2}{2\sigma^2}}$$

(we ignore the normalization constant), and the likelihood is

$$L_n(\theta) \propto \prod_{i=1}^{n} e^{-\frac{1}{2}(X_i - \theta)^2} = \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right).$$

Putting the two together, we get

$$\begin{aligned} f(\theta \mid X_1, \ldots, X_n) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2\right) e^{-\frac{\theta^2}{2\sigma^2}} \\ &= \exp\left(-\frac{1}{2}\left[\sum_i X_i^2 - 2\sum_i X_i\theta + n\theta^2 + \frac{\theta^2}{\sigma^2}\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(n+\frac{1}{\sigma^2}\right)\theta^2 - 2\theta\sum_i X_i\right]\right) \end{aligned}$$

To get the last line we got rid of $e^{-\frac{1}{2}\sum_i X_i^2}$, since it's a constant (it doesn't depend on θ). We now complete the square inside the brackets. In general, we have

$$\begin{aligned} a\theta^2 - 2b\theta = a\left(\theta^2 - 2\frac{b}{a}\theta\right) &= a\left(\theta^2 - 2\frac{b}{a}\theta + \frac{b^2}{a^2}\right) - \frac{b^2}{a} \\ &= a\left(\theta - \frac{b}{a}\right)^2 - \frac{b^2}{a}. \end{aligned}$$

We apply this formula with $a = n + 1/\sigma^2$ and $b = \sum_i X_i$. Note that the $-\frac{b^2}{a}$ part will not depend on θ, and since it's inside the exponent, we have an overall $e^{\frac{b^2}{2a}}$ multiplicative factor, which we can throw out. Therefore, we finally get

$$\text{posterior } f(\theta \mid X_1, \ldots, X_n) \propto \exp\left(-\frac{n+\frac{1}{\sigma^2}}{2}\left(\theta - \frac{\sum_i X_i}{n+1/\sigma^2}\right)^2\right) \quad (1)$$

But recall that the pdf of $\mathcal{N}(\mu, \tau^2)$ is

$$f_{\mu,\tau^2}(x) \propto \exp\left(-\frac{1}{2\tau^2}(x-\mu)^2\right). \quad (2)$$

Comparing the posterior (1) to the density (2), we see that

$$\text{posterior } = \mathcal{N}\left(\frac{\sum_i X_i}{n+1/\sigma^2}, \left(n+\frac{1}{\sigma^2}\right)^{-1}\right) \quad (3)$$

Since the mean and the mode of a normal distribution are equal to each other, we get

$$\hat{\theta}^{\text{Bayes}} = \hat{\theta}^{\text{MAP}} = \frac{\sum_i X_i}{n+1/\sigma^2} = \frac{\bar{X}_n}{1+1/n\sigma^2}.$$

We see that when $\sigma \to \infty$, the Bayes/MAP estimator gets closer to $\bar{X}_n$, the MLE. This is to be expected: as σ gets larger, the prior becomes higher variance, so there is less and less information contained in the prior (like a flat prior). So the posterior approaches the unweighted likelihood, which is maximized at the MLE $\bar{X}_n$.