18.650: Fundamentals of Statistics November 3, 2023

Lecture 20 — Bayesian inference

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

Overview. So far we have been working with frequentist inference. In frequentist inference, the interpretation of an event having probability 0.9 is that if you were to repeat the experiment many times, then 90% of the time the event would occur. For example, $\mathbb{P}(a(X_{1:n}) < \theta^* < b(X_{1:n})) = 0.9$ means for 90% of the datasets $X_{1:n}$, the point θ^* will be trapped in the interval $(a(X_{1:n}), b(X_{1:n}))$.

The Bayesian approach is an alternative method to produce estimators, confidence intervals, and tests. The interpretation of probability in Bayesian inference has more to do with degree of belief, as we will see.

To explain Bayesian inference, let's review how we computed the MLE, which is a frequentist concept. Write $f(x|\theta)$ to denote $f_\theta(x)$ (we are simply switching notation). Now, define the joint n-dimensional pdf

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta).$$

Note that this is simply the joint pdf of $X_1, \dots, X_n$ in the case that each of the X_i has pdf $f(x \mid \theta)$ (because the joint pdf of independent random variables is the product of the pdfs of each of the individual random variables.) But now, recall that the log likelihood $\ell_n(\theta)$ is the log of the likelihood $L_n(\theta)$, which is precisely given by

$$L_n(\theta) = \prod_{i=1}^n f(X_i|\theta) = f(X_1, \dots, X_n|\theta).$$

Since $\hat{\theta}^{\text{MLE}} = \text{argmax}_\theta\, \ell_n(\theta) = \text{argmax}_\theta\, L_n(\theta)$, we see that

$$\hat{\theta}^{\text{MLE}} = \text{argmax}_\theta\, L_n(\theta) = \text{argmax}_\theta\, f(X_1, \dots, X_n|\theta).$$

Therefore, *the MLE is simply the parameter value θ which maximizes the probability of our observed data if the ground truth parameter were equal to θ.*

1 The Bayesian method

In the Bayesian world, we weight the likelihood by a prior pdf $f(\theta)$, which indicates our prior belief about θ *before* seeing the data $X_1, \dots, X_n$.

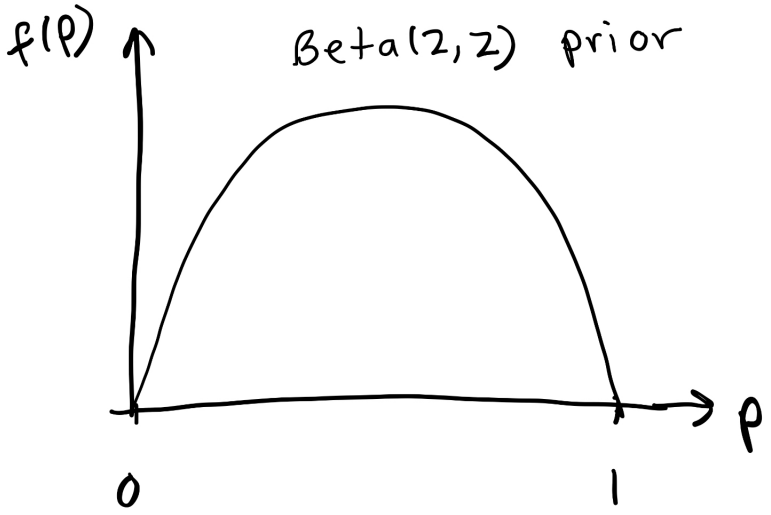

Figure 1: A Beta$(2, 2)$ prior on the parameter p of a Bernoulli distribution

Example.

Recall the kiss example: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$. If we believe apriori that there is no preference for head turning, our prior pdf $f(p)$ might look like the one in Figure 1. This pdf is symmetric about $p = 1/2$ and is maximized at $p = 1/2$, though it still allows for the possibility that p might not equal $1/2$. This is a Beta$(2, 2)$ distribution; the pdf is $f(p) = 6p(1 - p)$. More generally, a Beta(a, b) distribution has pdf

$$f(p) = \frac{1}{K} p^{a-1}(1-p)^{b-1}, \tag{1}$$

where $1/K$ is a normalizing constant.

After seeing the data, we update our prior belief into a posterior belief. The posterior belief is based on Bayes rule. Suppose for now that we are working with discrete pmf's. Then Bayes rule gives us that

$$\mathbb{P}(\theta \mid \text{data}) = \frac{\mathbb{P}(\text{data} \mid \theta)\mathbb{P}(\theta)}{\mathbb{P}(\text{data})},$$

where "data" refers to $X_1, \ldots, X_n$. With continuous pdfs, the same rule applies. We get that the *posterior* pdf is

$$f(\theta \mid X_1, \ldots, X_n) \overset{\text{Bayes rule}}{=} \frac{f(X_1, \ldots, X_n \mid \theta) f(\theta)}{f(X_1, \ldots, X_n)} = \frac{L_n(\theta) f(\theta)}{c_n},$$

In the second equality, we (1) observed that $f(X_1, \ldots, X_n \mid \theta)$ is precisely the likelihood $L_n(\theta)$, and (2) have written the denominator simply as c_n. In particular, *c_n does not depend on θ, so it is just the normalizing constant!* We can compute it by noting that

$$1 = \int f(\theta \mid X_1, \ldots, X_n) d\theta = \frac{1}{c_n} \int L_n(\theta) f(\theta) d\theta$$
$$\implies c_n = \int L_n(\theta) f(\theta) d\theta.$$

However, we usually do not write out this normalizing constant explicitly and simply write $f(\theta \mid X_1, \ldots, X_n) \propto L_n(\theta) f(\theta)$ where $\propto$ means "proportional to".

Definition 1.1: Posterior

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

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

Example.

Let us go back to the kiss example with prior $f(p) = 6p(1-p)$, the pdf of Beta$(2, 2)$. Note that the likelihood is

$$L_n(p) = f(X_1, \ldots, X_n \mid p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_i X_i}(1-p)^{n-\sum_i X_i},$$

which is just the product of the pmfs. Therefore, the posterior is

$$\begin{aligned} f(p \mid X_1, \ldots, X_n) &\propto L_n(p) f(p) \\ &\propto p^{\sum_i X_i}(1-p)^{n-\sum_i X_i} p(1-p) \\ &= p^{1+\sum_i X_i}(1-p)^{n+1-\sum_i X_i}. \end{aligned} \tag{2}$$

Note that we dropped the 6 from $f(p)$ because it's just a constant of proportionality. Now if we look at the third line of (2) and compare to the general formula for a beta distribution given in (1), we see that the posterior (2) is itself a Beta!

$$f(p \mid X_1, \ldots, X_n) = \text{pdf of Beta}\left(2 + \sum_i X_i, n + 2 - \sum_i X_i\right)$$

Figure 2 shows what the posterior looks like in the case $\sum_i X_i = n$, corresponding to all couples turning their heads to the right when they kiss.

Note that we didn't even need to compute the normalizing constant to get the pos-

terior in the kiss example. We just recognized that the posterior has the form of the beta distribution, up to normalization constant (which is uniquely determined).

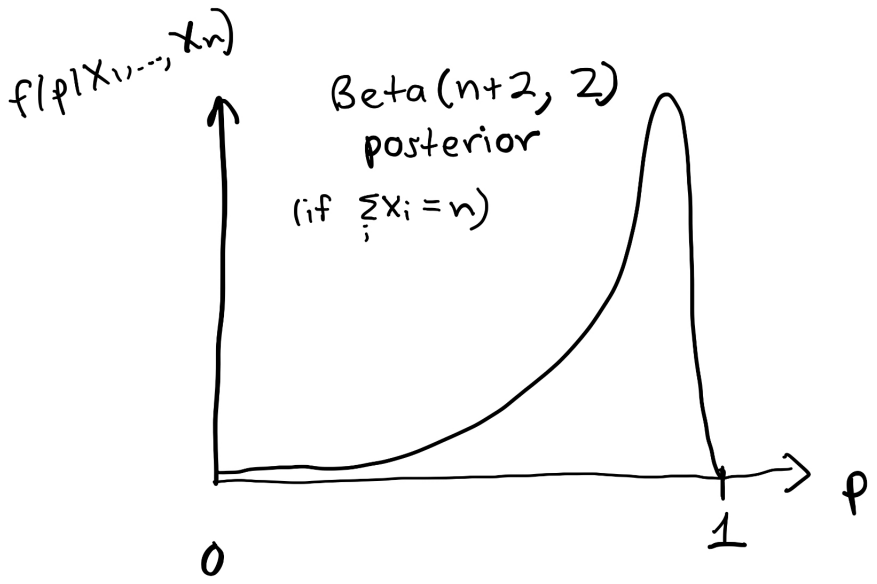

Figure 2: The posterior $\text{Beta}(n+2, 2)$ updated from the prior $\text{Beta}(2, 2)$, supposing we observed $\sum_i X_i = n$. This conveys the intuition about prior vs posterior beliefs: if we observed all couples turning their heads to the right, our posterior belief is much more skewed toward $p = 1$. However, the posterior still assigns positive probability to p's less than 1. This shows that our prior belief (depicted in Figure 1) still has some effect on the posterior.

If the prior and posterior turn out to be in the same family of distributions, as we saw with the Beta in the kiss example, we call the prior a "conjugate" prior.

Definition 1.2: Conjugate prior

When the prior and the posterior are in the same family of distributions, we say that we have a conjugate prior.

Remark.

Whether or not the prior and posterior are in the same family depends on what the likelihood is. For example, the beta is a conjugate prior when the samples are i.i.d. Bernoulli, but may not be a conjugate prior when the data has some other distribution.

The most natural way to summarize the posterior pdf is to use the mean.

Definition 1.3: Bayes estimator

The Bayes estimator is the expectation (mean) of the posterior.

It can be really hard to compute the mean of the posterior, because doing so requires knowing the normalizing constant which is often difficult to compute. In contrast, the maximum of the posterior (the mode) does not require the normalizing constant.

Definition 1.4: MAP (max a posteriori)

The MAP is the mode of the posterior, given by

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

Because it is simpler to compute, the MAP is also a popular summary statistic to describe the posterior.

Remark.

> Note from Figure 2 that the mode $\hat{\theta}^{\text{MAP}}$ of the posterior in the kiss example is a bit less than 1, even though we observed all couples turning their heads to the right. In contrast, for the MLE we have $\hat{\theta}^{\text{MLE}} = \bar{X}_n = 1$. The difference is that $\hat{\theta}^{\text{MAP}} = \text{argmax}_\theta \, L_n(\theta) f(\theta)$ while $\hat{\theta}^{\text{MLE}} = \text{argmax}_\theta \, L_n(\theta)$. The fact that $\hat{\theta}^{\text{MAP}} < 1$ is due to the effect of the prior $f(\theta)$.