Bayesian Inference

Introduction

The frequentist approach to inference holds that probabilities are intrinsicially tied (unsurprisingly) to frequencies. This interpretation is actually quite natural. What, according to a frequentist, does it mean to say that the probability a fair coin will come up heads is 1/2? Well, simply that in an infinite sequence of independent tosses of the same coin, half will come up heads (loosely speaking). Many random experiments are in fact repeatable, and the frequentist paradigm readily applies in such situations.

It is often desirable, however, to assign probabilities to events that are not repeatable. When the weather forecast tells you that there is a 90% chance of rain tomorrow, for example, it is assigning a probability to a one-off event, since tomorrow only happens once! What's more, there are many scenarios in which we would like to assign probabilities to non-random events that nevertheless involve uncertainty. A bank might be interested in designing an automated system that computes the probability that a signature on a check is genuine. Even though there is an underlying ground truth (the signature is either genuine or not), there is uncertainty from the bank's point of view, so the use of probability is justified. The pure frequentist interpretation of probabilities cannot be squared up with either of these use cases.

Bayesian inference takes a subjective approach and views probabilities as representing degrees of belief. It is thus perfectly valid to assign probabilities to non-repeating and non-random events, so long as there is uncertainty that we wish to quantify. The fact that Bayesian probabilities are subjective does not mean they are arbitrary. The rules for working with Bayesian probabilities are identical to those for working with the frequentist variety. Bayesians are simply happy to assign probabilities to a larger class of events than frequentists are.

The essential spirit of Bayesian inference is encapsulated by Bayes'

theorem.

Bayes' theorem

Suppose that during a routine medical examination, your doctor informs you that you have tested positive for a rare disease. You are initially distressed, but as a good statistician, you are also aware that these tests can be finicky and there is some uncertainty in their results. Unfortunately for you, this test is quite accurate — it reports a positive result for 95% of the patients with the disease, and a negative result for 95% of the healthy patients. The outlook does not appear to be good.

As a good *Bayesian* statistician, however, you realize that these test accuracies are not quite the bottom line, as far as your health is concerned. If we let "+" and "−" denote a positive and negative test result, respectively, then the test accuracies are the conditional probabilities

$$P(+ \mid \text{disease}) = 0.95,$$
$$P(- \mid \text{healthy}) = 0.95.$$

But what you are really interested in is

$$P(\text{disease} \mid +).$$

In order to compute this last quantity, we need to "turn around" the conditional probabilities encoded in the test accuracies. This is achieved by Bayes' theorem.

Theorem 0.0.5 (Bayes' Theorem)**.** *Let $Y_1, \ldots, Y_k$ be a partition of the sample space Ω and let X be any event. Then*

$$P(Y_j|X) = \frac{P(X|Y_j)P(Y_j)}{\sum_{i=1}^{k} P(X|Y_i)P(Y_i)}.$$

Since "disease" and "healthy" partition the sample space of outcomes, we have

$$P(\text{disease}|+) = \frac{P(+|\text{disease})P(\text{disease})}{P(+|\text{disease})P(\text{disease}) + P(+|\text{healthy})P(\text{healthy})}.$$

Importantly, Bayes' theorem reveals that in order to compute the conditional probability that you have the disease given the test was positive, you need to know the "prior" probability you have the disease $P(\text{disease})$, given no information at all. That is, you need to know the overall incidence of the disease in the population to which you belong. We mentioned earlier that this is a rare disease. In fact, only 1 in 1,000 people are affected, so $P(\text{disease}) = 0.001$, which

in turn implies $P(\text{healthy}) = 0.999$. Plugging these values into the equation above gives

$$P(\text{disease} \mid +) \approx 0.019.$$

In other words, despite the apparent reliability of the test, the probability that you actually have the disease is still less than 2%. The fact that the disease is so rare means that most of the people who test positive will be healthy, simply because most people are healthy in general. Note that the test is certainly not useless; getting a positive result increases the probability you have the disease by about 20-fold. But it is incorrect to interpret the 95% test accuracy as the probability you have the disease.

The Bayesian procedure

The above example is illustrative of the general procedure for doing Bayesian inference. Suppose you are interested in some parameter θ.

1. Encode your initial beliefs about θ in the form of a *prior distribution* $P(\theta)$.

2. Collect data X via experimentation, observation, querying, etc.

3. Update your beliefs using Bayes' theorem to the *posterior distribution*

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

4. Repeat the entire process as more data become available.

Prior, likelihood, posterior

As it turns out, Bayes' theorem is so fundamental to Bayesian inference that special names are given to the terms in the equation.

Prior

The prior distribution is the unconditional distribution $P(\theta)$. The goal of the prior is to capture our pre-existing knowledge about θ, before we see any data. In the medical testing example, we used the incidence of the disease in the population as the prior probability that any particular individual has the disease.

Likelihood

In Bayesian and frequentist statistics alike, the likelihood of a parameter θ given data X is $P(X|\theta)$. The likelihood function plays such an important role in classical statistics that it gets its own letter:

$$L(\theta|X) = P(X|\theta).$$

This notation emphasizes the fact that we view the likelihood as a function of θ for some fixed data X.

Figure 4 shows a random sample x of 8 points drawn from a standard normal distribution, along with the corresponding likelihood function of the mean parameter.

Figure 4: The orange curve shows the likelihood function for the mean parameter of a normal distribution with variance 1, given a sample of 8 points (middle) from a standard normal density (top).

In general, given a sample of n independent and identically distributed random variables $X_1, \ldots, X_n$ from some distribution $P(X|\theta)$, the likelihood is

$$\begin{aligned} L(\theta|X_1, \ldots, X_n) &= P(X_1, \ldots, X_n|\theta) \\ &= \prod_{i=1}^{n} P(X_i|\theta). \end{aligned}$$

In the case of the normal distribution with variance 1 and unknown mean θ, this equation suggests a way to visualize how the likelihood function is generated. Imagine sliding the probability density function of a Normal$(\theta, 1)$ distribution from left to right by gradually increasing θ. As we encounter each sample X_i, the density "lifts" the point off the x-axis. The dotted lines in the middle panel of Figure 5 represent the quantities $P(X_i|\theta)$. Their product is precisely the likelihood, which is plotted in orange at the bottom of Figure 5.

3slide.png

Figure 5: Each orange dotted line in the middle panel represents the quantity $P(x_i|\theta)$. The product of the lengths of these dotted lines is the likelihoood for the value of θ that produced the density in blue.

We can see that the likelihood is maximized by the value of θ for which the density of a Normal$(\theta, 1)$ distribution is able to lift the most points the furthest off the x-axis. It can be shown that this maximizing value is given by the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

In this case we say that the sample mean is the *maximum likelihood estimator* of the parameter θ.

In Bayesian inference, the likelihood is used to measure quantify the degree to which a set of data X supports a particular parameter value θ. The essential idea is that if the data could be generated by a given parameter value θ with high probability, then such a value of θ is favorable in the eyes of the data.

Posterior

The goal of Bayesian inference is to update our prior beliefs $P(\theta)$ by taking into account data X that we observe. The end result of

this inference procedure is the posterior distribution $P(\theta|X)$. Bayes' theorem specifies the way in which the posterior is computed,

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

Since in any particular inference problem, the data is fixed, we are often interested in only the terms which are functions of θ. Thus, the essence of Bayes' theorem is

$$P(\theta|X) \propto P(X|\theta)P(\theta),$$

or in words,

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior},$$

where all the terms above are viewed as functions of θ. Our final beliefs about θ combine both the relevant information we had *a priori* and the knowledge we gained *a posteriori* by observing data.

Coin Tosses

To get an understanding of what the Bayesian machinery looks like in action, let us return to our coin toss example. Suppose you just found a quarter lying on the sidewalk. You are interested in determining the extent to which this quarter is biased. More precisely, you wish to determine the probability p that the coin will come up heads. The most natural way to determine the value of p is to start flipping the coin and see what happens. So you flip the coin once and observe that the coin comes up heads. What should you conclude?

It is tempting to say that we cannot conclude anything from a single coin toss. But this is not quite true. The result of this toss tells us at the very least that $p \neq 0$, whereas before the toss it was certainly possible that $p = 0$ (perhaps both sides were tails). Furthermore, we should now be slightly more inclined to believe that p takes on larger values than we were before the toss. Which values we believe are reasonable depends on what our prior beliefs were. Most of the coins I have encountered in my life have been fair, or at least very close to fair. So my prior distribution on the value of p for any particular coin might look something like this.