Lecture 9: Bayesian hypothesis testing

5 November 2007

In this lecture we'll learn about Bayesian hypothesis testing.

1 Introduction to Bayesian hypothesis testing

Before we go into the details of Bayesian hypothesis testing, let us briefly review frequentist hypothesis testing. Recall that in the Neyman-Pearson paradigm characteristic of frequentist hypothesis testing, there is an *asymmetric* relationship between two hypotheses: the NULL hypothesis H_0 and the ALTERNATIVE hypothesis H_A. A decision procedure is devised by which, on the basis of a set of collected data, the null hypothesis will either be *rejected* in favor of H_A, or *accepted*.

In Bayesian hypothesis testing, there can be more than two hypotheses under consideration, and they do not necessarily stand in an asymmetric relationship. Rather, Bayesian hypothesis testing works just like any other type of Bayesian inference. Let us consider the case where we are considering only two hypotheses: H_1 and H_2. We know we will collect some data $\vec{x}$ but we don't yet know what that data will look like. We are interested in the posterior probabilities $P(H_1|\vec{x})$ and $P(H_2|\vec{x})$, which can be expressed using Bayes rule as follows:

$$P(H_1|\vec{x}) = \frac{P(\vec{x}|H_1)P(H_1)}{P(\vec{x})} \tag{1}$$

$$P(H_2|\vec{x}) = 1 - P(H_1|\vec{x}) \tag{2}$$

Crucially, the probability of our data $P(\vec{x})$ takes into account the possibility of each hypothesis under consideration to be true:

$$P(\vec{x}) = P(\vec{x}|H_1)P(H_1) + P(\vec{x}|H_2)P(H_2) \tag{3}$$

In other words, we are *marginalizing* over the possible hypotheses to calculate the data probability:

$$P(\vec{x}) = \sum_i P(\vec{x}|H_i)P(H_i) \tag{4}$$

As an example, we will return once more to the case of the possibly weighted coin as a case study. We will call hypothesis 1 the "fair coin" hypothesis, that the binomial parameter π is 0.5. In Bayesian statistics, model parameters have probabilities, so we state the fair coin hypothesis as:

$$H_1 : P(\pi|H_1) = \begin{cases} 1 & \pi = 0.5 \\ 0 & \pi \neq 0.5 \end{cases}$$

The probability above is a PRIOR PROBABILITY on the binomial parameter π.

Hypothesis 2 is the "weighted coin" hypothesis. For this hypothesis we must place a non-trivial probability distribution on π. Suppose that we did not entertain the possibility that the coin was two-headed or two-tailed, but we did consider it possible that the coin was weighted so that two out of three tosses turned up either heads or tails, and that each of these two possibilities was equally likely. This gives us:

$$H_2 : P(\pi|H_2) = \begin{cases} 0.5 & \frac{1}{3} \\ 0.5 & \frac{2}{3} \end{cases} \tag{5}$$

In order to complete the comparison in Equation (1), we need prior probabilities on the hypotheses themselves, $P(H_1)$ and $P(H_2)$. If we had strong beliefs one way or another about whether this coin was fair (e.g., from prior experience with the coin vendor), we might set one of these prior probabilities close to 1. For these purposes, we will use $P(H_1) = P(H_2) = 0.5$.

Now suppose we flip the coin six times and observe the sequence

HHTHTH

We can summarize this dataset as Does this data favor H_1 or H_2? We answer this question by completing Equation (1). We have:

$$P(H_1) = 0.5 \tag{6}$$

$$P(\vec{x}|H_1) = \binom{6}{4}(\frac{1}{2})^4(\frac{1}{2})^2 \tag{7}$$

$$= \binom{6}{4}0.0156 \qquad P(H_2) = 0.5 \tag{8}$$

Now to complete the calculation of $P(\vec{x})$ in Equation (3), we need $P(\vec{x}|H_2)$. To do this, we need to consider all possible values of π given H_2—that is, MARGINALIZE over π just as we are marginalizing over H to get the probability of the data. We have:

$$P(\vec{x}|H_2) = \sum_i P(\vec{x}|\pi_i)P(\pi_i) \tag{9}$$

$$= P(\vec{x}|\pi = \frac{1}{3})P(\pi = \frac{1}{3}) + \tag{10}$$

$$P(\vec{x}|\pi = \frac{2}{3})P(\pi = \frac{2}{3}) + \tag{11}$$

$$= \binom{6}{4}(\frac{1}{3})^4(\frac{2}{3})^2 \times 0.5 + \binom{6}{4}(\frac{2}{3})^4(\frac{1}{3})^2 \times 0.5 \tag{12}$$

$$= \binom{6}{4} \times 0.0137 \tag{13}$$

thus

$$P(\vec{x}) = \overbrace{\binom{6}{4}0.0156}^{P(\vec{x}|H_1)} \times \overbrace{0.5}^{P(H_1)} + \overbrace{\binom{6}{4}0.0137}^{P(\vec{x}|H_2)} \times \overbrace{0.5}^{P(H_2)} \tag{14}$$

$$= \binom{6}{4}0.01465 \tag{15}$$

Therefore

$$P(H_1|\vec{x}) = \frac{\binom{6}{4}0.0156 \times 0.5}{\binom{6}{4}0.01465} \tag{16}$$

$$= 0.53 \tag{17}$$

Note that even though the maximum-likelihood estimate of $\hat{\pi}$ from the data we observed hits one of the two possible values of π under H_2 on the head, our data actually supports the "fair coin" hypothesis H_1 – its support went up from a prior probability of $P(H_1) = 0.5$ to a posterior probability of $P(H_1|\vec{x}) = 0.53$.

1.1 More complex hypotheses

We might also want to consider more complex hypotheses than H_2 above as the "weighted coin" hypothesis. For example, we might think all possible values of π in $[0, 1]$ are equally probable *a priori*:

$$H_3 : P(\pi|H_2) = 1 \quad 0 \leq \pi \leq 1$$

(In Hypothesis 3, the probability distribution over π is continuous, not discrete, so H_3 is still a proper probability distribution.) Let us discard H_2 and now compare H_1 against H_3.

Let us compare H_3 against H_1 for the same data. To do so, we need to calculate the likelihood $P(\vec{x}|H_2)$, and to do this, we need to marginalize over π. Since π can take on a continuous range of values, this marginalization takes the form of an integral:[1]

[1] In general, the following relation holds:

$$\int_0^1 \pi^a (1-\pi)^b d\pi = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

$$= \frac{a!b!}{(a+b+1)!} \qquad \text{when } a \text{ and } b \text{ are integers}$$

The quantity $\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b+1)}$ is also called the BETA FUNCTION with parameters $a+1$,$b+1$, accessible in `R` as `beta()`.

$$P(\vec{x}|H_2) = \binom{6}{4} \int_0^1 \pi^4 (1-\pi)^2 d\pi$$
$$= \binom{6}{4} 0.0095$$

If we plug this result back in, we find that

$$P(H_1|\vec{x}) = \frac{\binom{6}{4} 0.0156 \times 0.5}{\binom{6}{4} 0.01255}$$
$$= 1.243$$

So H_3 fares even worse than H_2 against the fair-coin hypothesis H_1. Correspondingly, we would find that H_2 is favored over H_3.

> Would our hypothesis-testing results be changed at all if we did not consider the data as summarized by the number of successes and failures, and instead used the likelihood of the specific sequence `HHTHTH` instead?

1.2 Bayes factor

Sometimes we do not have strong feelings about the prior probabilities $P(H_i)$. Nevertheless, we can quantify how strongly a given dataset supports one hypothesis over another in terms of

$$\frac{P(\vec{x}|H_1)}{P(\vec{x}|H_2)}$$

that is, the LIKELIHOOD RATIO for the two hypotheses. This likelihood ratio is also called the BAYES FACTOR.

2 Learning contextual contingencies in sequences

Consider a sequence (e.g., phonemes) of length 20.

> ABABBAAAAABBBABBBAAAA

Let us entertain two hypotheses. The first hypothesis H_1, is that the probability of an A is independent of the context. The second hypothesis, H_2, is that the probability of an A is dependent on the preceding token. How might this data influence the learner?

We can make these hypotheses precise in terms of the parameters that each entails. H_1 involves only one parameter $P(A)$, which we will call π. H_2 involves three parameters:

1. $P(A|\emptyset)$ (the probability that the sequence will start with A), which we will call $\pi_\emptyset$;
2. $P(A|A)$ (the probability that an A will appear after an A), which we will call π_A
3. $P(A|B)$ (the probability that an A will appear after an B), which we will call π_B.

Let us assume that H_1 and H_2 are equally likely; we will be concerned with the likelihood ratio between the two hypotheses. We will put a uniform prior distribution on all model parameters.

There are 21 observations, 12 of which are A and 9 of which are B . The likelihood of H_1 is therefore simply

$$\int_0^1 \pi^{12}(1-\pi^9)d\pi = 1.546441 \times 10^{-07} \tag{18}$$

To calculate the likelihood of H_2 it helps to break down the results into a table:

	Outcome	
	A	B
$\emptyset$	1	0
A	7	4
B	4	5

So the likelihood of H_2 is

$$\int_0^1 \pi_\emptyset^1 d\pi_\emptyset \times \int_0^1 \pi_A^7(1-\pi_A^4)d\pi_A \times \int_0^1 \pi_B^4(1-\pi_B^5)d\pi_B \tag{19}$$

$$= 0.5 \times 0.00025 \times 0.00079 \tag{20}$$

$$= 1.002084 \times 10^{-07} \tag{21}$$

This dataset provides some support for the simpler hypothesis of statistical independence—the Bayes factor is about 1.5 in favor of H_1.