Bayesian Statistics IV

Matthieu Vignes

April 2024

Bayesian calculation for comparing K models

Suppose we have an observation X, which we believe to have come from one of K models, $M_1, \ldots, M_K$. Suppose we can compute the likelihood for any models. We present the computation of the posterior probability that the observation came from model K, and emphasize that the result depends only on the likelihood ratios. This is a straightforward extension of the 2-class calculations (`BayesIII` document)

Elephant example

In a previous example, we considered the question of whether a tusk came from one of two classes: a savanna elephant or a forest elephant, based on its DNA. In practice we might be interested in finer-scale distinctions than this. For example, forest elephants from West Africa actually differ somewhat genetically from those in Central Africa. And Savanna elephants from North Africa differ from those in the East and the South. (Actually elephant genetics within each subspecies varies roughly continuously across the continent; and any division into discrete groups can be seen as a convenient approximation.)

So what if now we have allele frequencies for "North Savanna", "South Savanna", "East Savanna", "West Forest", and "Central Forest" groups? How do we decided which group a tusk likely came from? Now we have five models, but the calculation is the same for K models, so we may as well do it for a general value of K. Here is the general outline:

Suppose we are presented with a series of observations $x_1, \ldots, x_n$, each of which are generated from a model M_k for some $k \in \{1, \ldots, K\}$. Let $Z_i \in \{1, \ldots, K\}$ indicate which model the ith observation came from, and let π_k denote the proportion of the observations that came from model M_k. Bayes Theorem says tha:

$$P(Z_i = k|x_i) = P(x_i|Z_i = k)P(Z_i = k)/P(x_i).$$

By using the law of total probability on $P(x_i)$ and recalling that $L_{ik} := P(x_i|Z_i = k)$ is the "likelihood" for model k given data x_i. Also $P(Z_j = k) = \pi_k$, so altogether:

$$P(Z_i = k|x_i) = \frac{L_{ik}\pi_k}{\sum_{k'} L_{ik'}\pi_{k'}} \propto L_{ik}\pi_k,$$

and the last proportionality statement stems from the fact that all numerators are the same for all k; the normalisation constant is here for the probabilities to sum to 1. This way of applying Bayes theorem is very common and convenient in practice, so you should get used to it. In words, this formula can be said:

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

Numerical application

The five rows of the matrix `ref_freqs` represent the allele frequencies in five groups: "North Savanna", "South Savanna", "East Savanna", "West Forest", and "Central Forest". The calculation presented here assumes that

the population of tusks we are looking at is equally drawn from all four groups, so $\pi = (0.2, 0.2, 0.2, 0.2, 0.2)$, but it would of course be easy to change to any other value of π.

```
x <- c(1,0,1,0,0,1)
ref_freqs <- rbind(
  c(0.39, 0.14,0.22,0.12,0.03,0.38),
  c(0.41, 0.10,0.18, 0.12,0.02,0.28),
  c(0.40, 0.11,0.22, 0.11,0.01,0.3),
  c(0.75,0.25,0.11,0.18,0.25,0.25),
  c(0.85,0.15,0.11,0.16,0.21,0.26)
)

# define functions for computing posterior from Likelihood vector and pi vector
normalize <- function(x){return(x/sum(x))}
posterior_prob <- function(L_vec, pi_vec){ return(normalize(L_vec*pi_vec)) }

# define likelihood function
L <- function(f,x){ prod(f^x*(1-f)^(1-x)) }

# compute the likelihoods for each model by applying L to rows of ref_freqs
L_vec <- apply(ref_freqs, 1, L, x = x)
print(L_vec)
```

```
## [1] 0.023934466 0.016038570 0.020702326 0.009513281 0.013712299
```

```
posterior_prob(L_vec, c(0.2,0.2,0.2,0.2,0.2))
```

```
## [1] 0.2852705 0.1911608 0.2467472 0.1133871 0.1634344
```

Remarks

1. Remember that when comparing two models, only the likelihood ratios matter, not the actual likelihoods. In fact the same is true when comparing K models, as we can see by examining the calculation above. Specifically, imagine multiplying all the likelihoods by some positive constant c, and notice that this would not change the final answer, because of the normalization step.
2. Notice that, just as with the 2-model case, the calculation involves weighing the relative support from the data for each model (from the likelihood function) against the "prior" plausibility of each model (from the vector π).
3. In practice we might not know π. And although in such a case it might seem natural to assume that all the values of π are equal, one has to be careful to note that this is still an assumption, and such assumptions may have unforeseen implications. For example, in this case, this assumption implies that 60% of the tusks are from savanna elephants and 40% from forest elephants, not 50%-50% (because three of our five groups are savanna). The difference between 60-40 and 50-50 is probably not a big deal in most applications, but imagine that we had 20 different savanna groups and 2 forest groups. Would we still be happy to assume that every group was equally common (and so savanna tusks are 10 times as common as forest tusks)? The answer would depend on the context, but quite possibly not.