1 The Basics of Bayesian Statistics

In this course, we will apply statistics to draw conclusions on astrophysical data. For the first few lectures, we therefore study Bayesian statistics to equip ourselves with the tools and abilities that we later apply to cosmology.

First of all let's introduce notations. We denote the probability that some hypothesis "X" ist true given some information "I" with

$$\text{prob}(X \mid I)$$

probability that "X" is true "given" I

So the bar "|" is read as "given".
The exact opposite of X is denoted by "not-X": $\overline{X}$

Without loss of generality, we can normalize

$$\text{prob}(X \mid I) \in [0, 1]$$

There are two axioms that are needed for logical consistency. They are called sum-rule and product rule respectively.

The sum-rule says that the probability of X being true plus the probability of X being false is 100%.

sum rule: $\text{prob}(X \mid I) + \text{prob}(\overline{X} \mid I) = 1$

The second axiom, the product rule says:

product rule: $\text{prob}(X, Y \mid I) = \text{prob}(X \mid Y, I)\ \text{prob}(Y \mid I)$

As an example, consider:

X: "it's raining"
Y: "there are clouds on the sky"
I: It's the 15th of October & you are in Germany

Then the probability

$\text{prob}(X, Y \mid I)$: "It's raining and there are clouds on the sky"

is the product of

$\text{prob}(X \mid Y, I)$: "It's raining given that there are clouds and I"

and

$\text{prob}(Y \mid I)$: "There are clouds given I"

1.1 Bayes' theorem and marginalization

Bayes was the first to state an in principle simple corollary that follows from the product rule and the observation that

$$\text{prob}(X, Y \mid I) = \text{prob}(Y, X \mid I)$$

So:

$$\text{prob}(X \mid I, Y)\,\text{prob}(Y \mid I) = \text{prob}(Y \mid X, I)\,\text{prob}(X \mid I)$$

which leads to Bayes ~~the~~ Theorem:

$$\text{prob}(X \mid Y, I) = \frac{\text{prob}(Y \mid X, I)\,\text{prob}(X \mid I)}{\text{prob}(Y \mid I)}$$

Bayes theorem is extremely powerful! You'll immediately see this, when we use "hypothesis" and "data" instead of X and Y:

$$\text{prob}(\text{hypothesis} \mid \text{data}, I) \propto \text{prob}(\text{data} \mid \text{hypothesis}, I) \times \text{prob}(\text{hypothesis} \mid I)$$

So you learn about a hypothesis (model) you have given the data and additional information by computing how probable it is to measure the data you took given the hypothesis (and I) times the probability that the hypothesis is valid given I.

Bayes' Theorem is so important, that each term has its own name:

$$\text{prob}(X|Y,I) = \frac{\text{prob}(Y|X,I)\ \text{prob}(X|I)}{\text{prob}(Y|I)}$$

"posterior" ← prob(X|Y,I); "likelihood" ← prob(Y|X,I); "prior" ← prob(X|I); "evidence" ← prob(Y|I)

So you infer the "posterior" probability that your hypothesis X is right by the measurement yielding the "likelihood" that your hypothesis explains the (data) Y times your "prior" knowledge or prejudice that your hypothesis X is right divided by the "evidence" (more of this strange name later) that the (data) Y exists given the Information I.

Very often, the data Y is measured already and you want to test a hypothesis X. So the evidence prob(Y|I), which is unchanged by your hypothesis can is treated as a rather uninteresting nuisance.

A very useful tool is <u>marginalization</u>. Suppose first that you have a discrete set of propositions

$$\{Y_k\} = Y_1, Y_2, \ldots Y_H$$

For instance with 5 presidential candidates, Y_1 could be the probability proposition that candidate "1" wins etc.

The probability that X is true, for instance that unemployment is lower in a year's time irrespective of who is president is then

$$\text{prob}(X \mid I) = \sum_{k=1}^{M} \text{prob}(X, Y_k \mid I)$$

While this seems sensible, let us prove it by using the product rule. First, we have

$$\text{prob}(X, Y_1 \mid I) = \text{prob}(Y_1, X \mid I) = \text{prob}(Y_1 \mid X, I)\ \text{prob}(X \mid I)$$

Now add all k propositions of Y_k:

$$\text{prob}(X, Y_1 \mid I) + \text{prob}(X, Y_2 \mid I) + \ldots + \text{prob}(X, Y_k \mid I)$$

$$= \left[\text{prob}(Y_1 \mid X, I) + \text{prob}(Y_2 \mid X, I) + \ldots + \text{prob}(Y_k \mid X, I)\right] \times \text{prob}(X \mid I)$$

But as $\{Y_k\}$ is a mutually exclusive and exhaustive set, i.e.

$$\sum_{k=1}^{M} \text{prob}(Y_k \mid X, I) = 1,$$

the quantity in square brackets above is 1 and we obtain our desired result.

Now suppose that instead of 5 presidential candidates, we'd have a continuous set of propositions, for instance if Y were the amplitude of some nuisance noise of a telescope.

in this case, it is obvious to generalize the concept of probability to <u>probability density functions</u> (pdf's):

$$\text{pdf}(X, Y=y \mid I) = \lim_{\delta y \to 0} \frac{\text{prob}(X, y \leq Y < y+\delta y \mid I)}{\delta y}$$

Yet, we'll use the notation prob() instead of pdf() for notational convenience!
So finally, we can generalize the sum we computed above to the important statement of marginalization over some pa. proportion (parameter) Y which we are not interested in:

<u>Marginalization:</u> $$\text{prob}(X \mid I) = \int_{-\infty}^{\infty} \text{prob}(X, Y \mid I)\, dY$$

2 Parameter Estimation I

Suppose we would like to find out, if a certain coin is fair in the sense that flipping it yields the same chance for heads as it does for tails. Let us denote the bias of the coin (if any) by a parameter $H \in [0,1]$.

If $H = 0 \Rightarrow$ all tails

" $H = 1 \Rightarrow$ all heads

and an unbiased coin would have $H = 1/2$.

Suppose further that we really have no opinion on wether it is biased and on how much it is. So our prior could be:

$$\text{prob}(H \mid I) = \begin{cases} 1 & ; \; 0 \leq H \leq 1 \\ 0 & ; \; \text{otherwise} \end{cases}$$

and Baye's theorem says

$$\text{prob}(H \mid \{\text{data}\}, I) = \frac{\text{prob}(\{\text{data}\} \mid H, I)\, \text{prob}(H \mid I)}{\text{prob}(\{\text{data}\} \mid I)}$$

Please note that since our posterior will be normalized,

$$\int_0^1 \text{prob}(H \mid \{\text{data}\}, I)\, dH = 1$$

we can neglect the evidence $\text{prob}(\{\text{data}\} \mid I)$, because it does not depend on H and can be retrieved from the normalization condition later, if we really want to know it.

So Bayes's theorem says

$$\text{prob}(H \mid \{\text{data}\}, I) \propto \text{prob}(\{\text{data}\} \mid H, I)\ \text{prob}(H \mid I)$$

and the likelihood to obtain 'R' heads in 'N' tosses is given by

$$\text{prob}(\{\text{data}\} \mid H, I) \propto H^R (1-H)^{N-R}$$

So

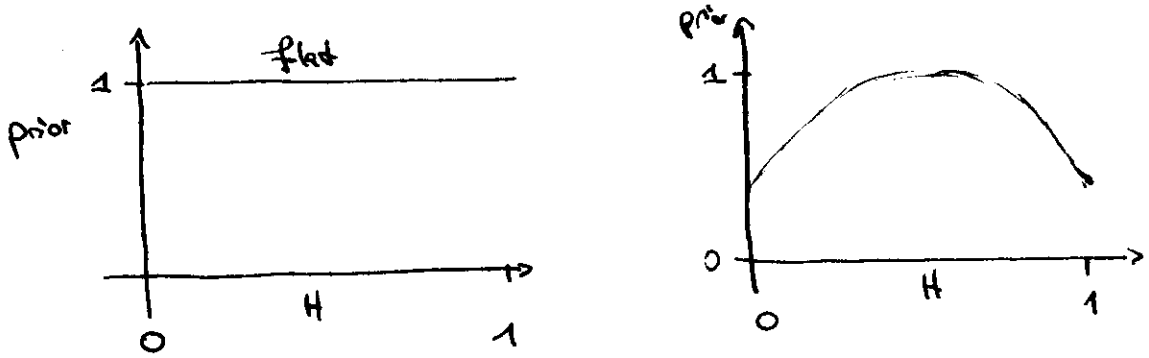
$$\text{prob}(H \mid \{\text{data}\}, I) \propto H^R (1-H)^{N-R} \quad ; 0 \le H \le 1$$

and 0 otherwise.

Figure?

The choice of priors

What if we chose a different prior? Say one which gives more weight to $H = 1/2$ than to $H = 0$ or $H = 1$:



As the prior still has $\text{prob}(H \mid I) \neq 0$ everywhere, it does not exclude the case of $H = 0$ or $H = 1$. In the long run, the data-taking will win over the prior, for as long as the prior is not too narrow or excludes values of H that are actually the coins true value.

On the other hand, a sensible prior can improve the quality of a deduction: Suppose you measured the length of some sort of fish in a river by taking a random sample of 10 such fishes. If you'd look up the usual length (plus deviations) of that sort of fish, you could improve your estimate of the average length by this prior knowledge, i.e. if you denote the average length as L, then

$$\text{prob}(L \mid \{\text{data}\}, I) \propto \text{prob}(\{\text{data}\} \mid L, I)\ \text{prob}(L \mid I)$$

and your prior knowledge of $\text{prob}(L \mid I)$ could help you here! You can also use the prior like here to incorporate prior experiments. Say you'd like to infer the Hubble parameter H from the WMAP measurements of the CMB. You could then add the HST measurement of H, to improve on your deduction and account for the prior measurement of H.