Data Analysis and Machine Learning: Elements of Bayesian theory

Christian Forssén[1]

Morten Hjorth-Jensen[2,3]

[1]Department of Physics, Chalmers University of Technology, Sweden
[2]Department of Physics, University of Oslo
[3]Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University

Mar 10, 2018

What is Bayesian Statistics

Morten's original plan: Reminder about probabilities from the statistics section

1. Product rule
2. Binomial distribution
3. Gaussian PDF
4. other PDFs
5. Bayesian regression analysis

Inference

Inference: "the act of passing from one proposition, statement or judgment considered as true to another whose truth is believed to follow from that of the former" (Webster)
Do premises $A, B, \ldots \to$ hypothesis, H?

Deductive inference: Premises allow definite determination of truth/falsity of H (syllogisms, symbolic logic, Boolean algebra)
$B(H|A,B,...) = 0$ or 1

Inductive inference: Premises bear on truth/falsity of H, but don't allow its definite determination (weak syllogisms, analogies)
A, B, C, D share properties x, y, z; E has properties x, y
$\to E$ probably has property z.

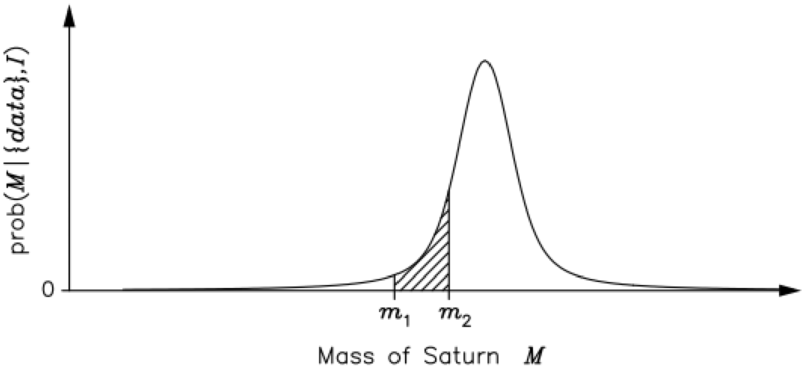
Statistical Inference

- Quantify the strength of inductive inferences from facts, in the form of data (D), and other premises, e.g. models, to hypotheses about the phenomena producing the data.
- Quantify via probabilities, or averages calculated using probabilities. Frequentists ($\mathcal{F}$) and Bayesians ($\mathcal{B}$) use probabilities very differently for this.
- To the pioneers such as Bernoulli, Bayes and Laplace, a probability represented a *degree-of-belief* or plausability: how much they thought that something as true based on the evidence at hand. This is the Bayesian approach.
- To the 19th century scholars, this seemed too vague and subjective. They redefined probability as the *long run relative frequency* with which an event occurred, given (infinitely) many repeated (experimental) trials.

Some history

Adapted from D.S. Sivia[1]:

> Although the frequency definition appears to be more objective, its range of validity is also far more limited. For example, Laplace used (his) probability theory to estimate the mass of Saturn, given orbital data that were available to him from various astronomical observatories. In essence, he computed the posterior pdf for the mass M , given the data and all the relevant background information I (such as a knowledge of the laws of classical mechanics): prob(M|data,I); this is shown schematically in the figure [Fig. 1.2].



[1] Sivia, Devinderjit, and John Skilling. Data Analysis : A Bayesian Tutorial, OUP Oxford, 2006

To Laplace, the (shaded) area under the posterior pdf curve between m_1 and m_2 was a measure of how much he believed that the mass of Saturn lay in the range $m_1 \leq M \leq m_2$. As such, the position of the maximum of the posterior pdf represents a best estimate of the mass; its width, or spread, about this optimal value gives an indication of the uncertainty in the estimate. Laplace stated that: ' . . . it is a bet of 11,000 to 1 that the error of this result is not 1/100th of its value.' He would have won the bet, as another 150 years' accumulation of data has changed the estimate by only 0.63%!

According to the frequency definition, however, we are not permitted to use probability theory to tackle this problem. This is because the mass of Saturn is a constant and not a random variable; therefore, it has no frequency distribution and so probability theory cannot be used.

If the pdf [of Fig. 1.2] had to be interpreted in terms of the frequency definition, we would have to imagine a large ensemble of universes in which everything remains constant apart from the mass of Saturn.

As this scenario appears quite far-fetched, we might be inclined to think of [Fig. 1.2] in terms of the distribution of the measurements of the mass in many repetitions of the experiment. Although we are at liberty to think about a problem in any way that facilitates its solution, or our understanding of it, having to seek a frequency interpretation for every data analysis problem seems rather perverse. For example, what do we mean by the 'measurement of the mass' when the data consist of orbital periods? Besides, why should we have to think about many repetitions of an experiment that never happened? What we really want to do is to make the best inference of the mass given the (few) data that we actually have; this is precisely the Bayes and Laplace view of probability.

Faced with the realization that the frequency definition of probability theory did not permit most real-life scientific problems to be addressed, a new subject was invented — statistics! To estimate the mass of Saturn, for example, one has to relate the mass to the data through some function called the statistic; since the data are subject to 'random' noise, the statistic becomes the random variable to which the rules of probability the- ory can be applied. But now the question arises: How should we choose the statistic? The frequentist approach does not yield a natural way of doing this and has, therefore, led to the development of several alternative schools of orthodox or conventional statis- tics. The masters, such as Fisher, Neyman and Pearson, provided a variety of different principles, which has merely resulted in a plethora of tests and procedures without any clear underlying rationale. This lack of unifying principles is, perhaps, at the heart of the shortcomings of the cook-book approach to statistics that students are often taught even today.

The Bayesian recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information using the rules of probability theory.

Probability Theory Axioms:

Product (AND) rule : $p(A,B|I) = p(A|I)p(B|A,I) = p(B|I)p(A|B,I)$
Should read $p(A,B|I)$ as the probability for propositions A AND B being true given that I is true.

Sum (OR) rule: $p(A+B|I) = p(A|I) + p(B|I) - p(A,B|I)$
$p(A+B|I)$ is the probability that proposition A OR B is true given that I is true.

Normalization: $p(A|I) + p(\bar{A}|I) = 1$
$\bar{A}$ denotes the proposition that A is false.

Bayes' theorem

Bayes' theorem follows directly from the product rule

$$p(A|B,I) = \frac{p(B|A,I)p(A|I)}{p(B|I)}.$$

The importance of this property to data analysis becomes apparent if we replace A and B by hypothesis(H) and data(D):

$$p(H|D,I) = \frac{p(D|H,I)p(H|I)}{p(D|I)}.$$

The power of Bayes' theorem lies in the fact that it relates the quantity of interest, the probability that the hypothesis is true given the data, to the term we have a better chance of being able to assign, the probability that we would have observed the measured data if the hypothesis was true.

The various terms in Bayes' theorem have formal names.

- The quantity on the far right, $p(H|I)$, is called the *prior* probability; it represents our state of knowledge (or ignorance) about the truth of the hypothesis before we have analysed the current data.
- This is modified by the experimental measurements through $p(D|H,I)$, the *likelihood* function,

- The denominator $p(D|I)$ is called the *evidence*. It does not depend on the hypothesis and can be regarded as a normalization constant.
- Together, these yield the *posterior* probability, $p(H|D,I)$, representing our state of knowledge about the truth of the hypothesis in the light of the data.

In a sense, Bayes' theorem encapsulates the process of learning.

The friends of Bayes' theorem

Normalization: $\sum_i p(H_i|\ldots) = 1$.

Marginalization: $\sum_i p(A, H_i|I) = \sum_i p(H_i|A, I)p(A|I) = p(A|I)$.

Marginalization (continuum limit): $\int dx p(A, H(x)|I) = p(A|I)$.

In the above, H_i is an exclusive and exhaustive list of hypotheses. For example,let's imagine that there are five candidates in a presidential election; then H_1 could be the proposition that the first candidate will win, and so on. The probability that A is true, for example that unemployment will be lower in a year's time (given all relevant information I, but irrespective of whoever becomes president) is then given by $\sum_i p(A, H_i|I)$.

In the continuum limit of propositions we must understand $p(\ldots)$ as a pdf (probability density function).

Marginalization is a very powerful device in data analysis because it enables us to deal with nuisance parameters; that is, quantities which necessarily enter the analysis but are of no intrinsic interest. The unwanted background signal present in many experimental measurements are examples of nuisance parameters.

Inference With Parametric Models

Inductive inference with parametric models is a very important tool in the natural sciences.

- Consider N different models M_i $(i = 1, \ldots, N)$, each with parameters $\boldsymbol{\alpha}_i$. Each of them implies a sampling distribution (conditional predictive distribution for possible data)

$$p(D|\boldsymbol{\alpha}_i, M_i)$$

- The $\boldsymbol{\alpha}_i$ dependence when we fix attention on the actual, observed data (D_{obs}) is the likelihood function:

$$\mathcal{L}_i(\boldsymbol{\alpha}_i) \equiv p(D_{\text{obs}}|\boldsymbol{\alpha}_i, M_i)$$

- We may be uncertain about i (model uncertainty),
- or uncertain about $\boldsymbol{\alpha}_i$ (parameter uncertainty).

Parameter Estimation: Premise = choice of model (pick specific i)
$\Rightarrow$ What can we say about $\boldsymbol{\alpha}_i$?

Model comparison: Premise = $\{M_i\}$
$\Rightarrow$ What can we say about i?

Model adequacy: Premise = M_1
$\Rightarrow$ Is M_1 adequate?

Hybrid Uncertainty: Models share some common params: $\boldsymbol{\alpha}_1 = \{\boldsymbol{\varphi}, \boldsymbol{\eta}_i\}$
$\Rightarrow$ What can we say about $\boldsymbol{\varphi}$? (Systematic error is an example)

Illustrative examples with python code

- Is this a fair coin? (analytical)
- Flux from a star (single parameter, MCMC)
- The lighthouse problem (two parameters, MCMC)
- Linear fit with outliers (nuisance parameters)
- ...