Prepared for submission to JHEP

Methods for Data

Aditya Vijaykumar

International Centre for Theoretical Sciences, Bengaluru, India.

E-mail: aditya.vijaykumar@icts.res.in

Contents

1 A Few Basics

Physicists believe in a *cause and effect* philosophy of the world. If we are given a fair coin, we can assign equal probabilities to *head* and *tail* and predict the outcomes given the number of times the coin is tossed. We could deal with unfair coins equally well, and given the properties of the unfair coin (*ie.* probabilities of landing heads/tails/neither) are specified, we can make deductions about the outcomes of our *experiment*.

But, generally, we have information about the observed effects, from which we are supposed to ascertain what models/scenarios these effects would have arisen from. Consider tossing a single coin ten times. The nature of the coin is not specified, but we find that all ten times the coin lands heads up. Are we to believe that the coin is *biased*? Not quite. The most one can do is make some inference based on the data and our prior knowledge (*ie.* we say that the coin is *most likely* biased), adding a caveat that we reserve the right to alter our result if new information pops up in the future (*ie.* we have the data from more tosses of the same coin).

This makes the whole problem of *data analysis* open-ended. Through this exploration, we hope to understand the mathematical rules governing the analyses.

1.1 Operations on Probabilities

All the definitions of probability made below are with respect to some background information denoted by I.

- **Sum Rule** - Probability of an X being true and that of it being false should add up to 1.

$$P(X|I) + P(\overline{X}|I) = 1$$

- **Product Rule** - Probability of both X and Y being true should be the product of probability of Y being true and the probability of X being true given Y is true.

$$P(X, Y|I) = P(X|Y, I) \times P(Y|I) = P(Y|X, I) \times P(X|I)$$

  The symmetry property of the AND operation on X and Y means that we can interchange X and Y on the LHS.

- **Bayes' Theorem** - The probability of X conditioned on Y is proportional to the probability of Y conditioned on X.

$$P(X|Y,I) = \frac{P(Y|X,I) \times P(X|I)}{P(Y|I)}$$

The proof follows from the second and third expressions in the product rule. The magic of Bayes' theorem becomes evident when we replace X with *hypothesis* and Y with *data*.

$$\underbrace{P(hypothesis|data,I)}_{\text{posterior}} \propto \underbrace{P(data|hypothesis,I)}_{\text{likelihood}} \times \underbrace{P(data|hypothesis,I)}_{\text{prior}}$$

The statement of the Bayes' theorem now becomes a powerful tool for inversion! Some definitions follow,

  - **Prior** - This encodes our degree of ignorance about the hypothesis before we have even touched the data.
  - **Likelihood Function** - This encodes how well data agrees with a particular model (*Details Later*).
  - **Posterior** - This gives us the degree of knowledge of the truth of the model in light of the data.
  - **Evidence** - $P(data|I)$ which we omitted while writing the expression of the Bayes' theorem. In most cases, it will remain a proportionality constant.

- **Marginalization** - The probability of X is the same as the probability of X given Y_k, summed over all k. Here, $\{Y_k\}$ is the set of all possibilities in Y.

$$P(X|I) = \sum_k P(X|Y_k,I)$$

A continuous, integral form of the equation can be written as follows,

$$P(X|I) = \int_{-\infty}^{\infty} P(X,Y|I)\,\mathrm{d}Y$$