SCHOOL OF DATA
by Springboard

Data Science Career Track

The Art of Statistics, Chapter 11: Learning from experience the Bayesian way

Take-Away Notes

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$$

Bayes' Theorem performs a remarkable feat: it gives us a scientifically correct, general algorithm for updating our beliefs in the light of new evidence. We are shown how the evidence B updates our degree of belief in a given proposition A. The initial (or **prior)** probability distribution for the unknown parameters is revised to the **posterior** distribution using Bayes' theorem.

A **likelihood ratio** is a measure of the relative support that some data provides for two competing hypotheses. For hypotheses H0 and H1, the likelihood ratio given by data *x* is just P(*x*|H0)/P(*x*|H1). Using likelihood ratios, we can express the useful equation that:
(The initial odds for a hypothesis) x (the likelihood ratio) = (the final odds for the hypothesis).

Part of the power of Bayes' Theorem derives from its dispelling confusion around particularly puzzling cases ordinarily treated with *expected frequency trees*.
- **Bayes' Theorem *reverses the order*** of the expected frequency tree, putting testing first, and following this with the revelation of truth. This reversal (known as **'inverse probability'** until the 1950s) respects the temporal order in which we discover things.

Most of the controversy around Bayesian analysis is around what the source is for the prior distribution values.

- Suggestions for sources of the prior distributions include **subjective judgment**, **learning from historical data**, and specifying **objective priors** (that is, prior values that represent ignorance about parameters, and thereby, supposedly, let the data speak for themselves. No procedure for getting objective priors has been established, however).

Hierarchical modeling takes these ideas to another level: if the parameters underlying a number of units (such as areas or schools) are themselves assumed to be drawn from a common prior distribution, this results in **shrinkage** of the parameter estimates for individual units towards an overall mean. We can see the power of such models in (for example) pre-election polls.