Introduction to Probability Theory – Part III [1]

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

[1] Largely based on material from Sharon Goldwater's tutorial *Basics of Probability Theory* (henceforth abbreviated as SGT), available at: `https://homepages.inf.ed.ac.uk/sgwater/math_tutorials.html`

Bayes' Rule

- An inference rule with the goal of updating an (initial) hypothesis, called a *prior*, in the presence of more data.
- Enables a comparison between:
  - $P(A \mid B)$ and $P(B \mid A)$
  - a **prior** $P(A)$, the probability of an event A, without knowledge about whether an event B has occurred, and a **posterior**, the probability of an event A with knowledge that B has occurred.

Bayes' Rule: a First Look

Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

How to derive Bayes' Rule:

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

Bayes' Rule: a Second Look

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{18}$$

where H stands for **hypothesis** and whose probability is affected by E, which stands for **data** or **evidence**

Often we are faced with the task of evaluating the plausibility (i.e., probability) of competing hypotheses, and the task is to determine which is the most plausible (i.e., probable) hypothesis, given the data (the evidence).

Bayes' Rule: a Second Look

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{18}$$

- $P(H)$ is called the **prior probability**. It is a probability estimate without taking the probability of the data (i.e., the evidence) E into account.
- $P(H|E)$ is the **posterior probability**, which takes the probability of the data (i.e., the evidence) E into account.
- $P(E|H)$ is the called the **likelihood**: the probability of observing E given H.
- $P(E)$ is called the **evidence** or **marginal likelihood**.

Bayes' Rule: a Third Look

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{18}$$

Comparing likelihood and evidence:

If $\frac{P(E|H)}{P(E)} > 1$, then $P(E|H) > P(E)$

If $\frac{P(E|H)}{P(E)} = 1$, then $P(E|H) = P(E)$

Bayes' Rule: a Fourth Look

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)} \tag{18}$$

Bayes' Rule: a First Example[2]

We take as the prior the probability I am asleep, taking into account only the time of day. What if we know the time and have additional evidence? How would knowing that my bedroom light is on change the probability that I am asleep? This is where we use Bayes' Rule to update our estimate. For a specific time, if we know information about my bedroom light, we can use the probability from the distribution above as the prior and then apply Bayes' equation:

$$P(\mathit{sleep}|\mathit{light}) = \frac{P(\mathit{light}|\mathit{sleep}) \times P(\mathit{sleep})}{P(\mathit{light})}$$

[2] due to Will Koehrsen; `https://towardsdatascience.com/bayes-rule-applied-75965e4482ff`

Bayes' Rule: a First Example

The left side is the posterior, the conditional probability of sleep given the status of my bedroom light (either on or off). The probability at a given time will serve as our prior, P(sleep) , or the estimate we use if we have no additional information. For example, at 10:00 PM, the prior probability I am asleep is 27.34%. If we do have more information, we can update this using the likelihood, P(bedroom light |sleep) , which is derived from observed data. Based on my habits, I know the probability my bedroom light is on given that I am asleep is about 1%. That is:

$$P(light|sleep) = 0.01 \qquad P(\neg light|sleep) = 0.99$$

Bayes' Rule: a First Example

The final piece of the equation is the normalization constant $P(light)$. This represents the total probability my light is on. There are two conditions for which my light is on: I am asleep or I am awake. Therefore, if we know the prior probability of sleep, we can calculate the normalization constant as:

$$P(light) = P(light|sleep) \times P(sleep) + P(light|\neg sleep) \times P(\neg sleep)$$

Bayes' Rule: a First Example

The total probability my light is on takes into account both the chance I am asleep and my light is on and the chance I am awake and my light is on. ($P(\neg sleep) = 1 - P(sleep)$ is the probability I am awake.)

The probability my light is on given that I am not asleep, $P(light|\neg sleep)$, is also determined from observations. In my case, I know there is around a 80% probability my bedroom light is on if I am awake (which means there is a 20% chance my light is not on if I'm awake).

Using the total probability for my light being on, Bayes' equation is:

$$P(sleep|light) = \frac{P(light|sleep) \times P(sleep)}{P(light|sleep)P(sleep) + P(light|\neg sleep)P(\neg sleep)}$$

Bayes' Rule: a First Example

We will walk through applying the equation for a time of 10:30 PM if we know my light is on. First, we calculate the prior probability I am asleep using the time and get an answer of 73.90%. The prior provides a good starting point for our estimate, but we can improve it by incorporating info about my light. Knowing that my light is on, we can fill in Bayes' Equation with the relevant numbers:

$$
\begin{aligned}
P(\mathit{sleep}|\mathit{light}) &= \frac{P(\mathit{light}|\mathit{sleep}) \times P(\mathit{sleep})}{P(\mathit{light}|\mathit{sleep})P(\mathit{sleep}) + P(\mathit{light}|\neg\mathit{sleep})P(\neg\mathit{sleep})} \\
&= \frac{0.01 \times 0.7390}{0.01 \times 0.7390 + 0.8 \times (1 - 0.7390)} \\
&= .00342
\end{aligned}
$$

Bayes' Rule: a First Example

The knowledge that my light is on drastically changes our estimate of the probability I am asleep from over 70% to 3.42%. This shows the power of Bayes' Rule: we were able to update our initial estimate for the situation by incorporating more information. While we might have intuitively done this anyway, thinking about it in terms of formal equations allows us to update our beliefs in a rigorous manner.

Bayes' Rule: a Second Example

An individual has been described by a neighbor as follows: "Steve is very shy and withdrawn, invariably helpful, but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure and a passion for detail." Is Steve more likely to be a librarian or a farmer?

Please assume that Steve was selected at random from a representative sample.

This example is due to Daniel Kahneman (2011). Thinking Fast and Slow.

Applying Bayes Rule to Second Example

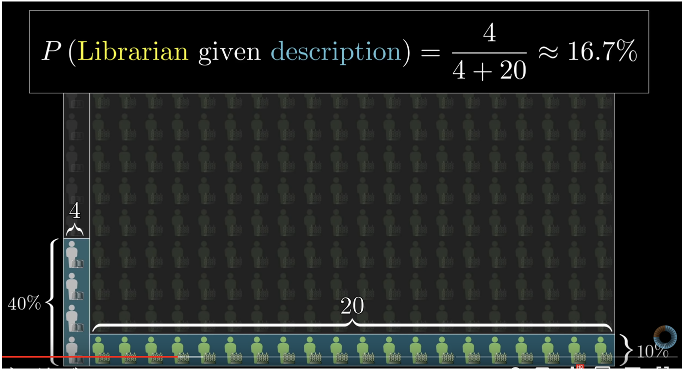

Figure: 3Blue1Brown Youtube Tutorial on Bayes Rule

Applying Bayes Rule to Second Example

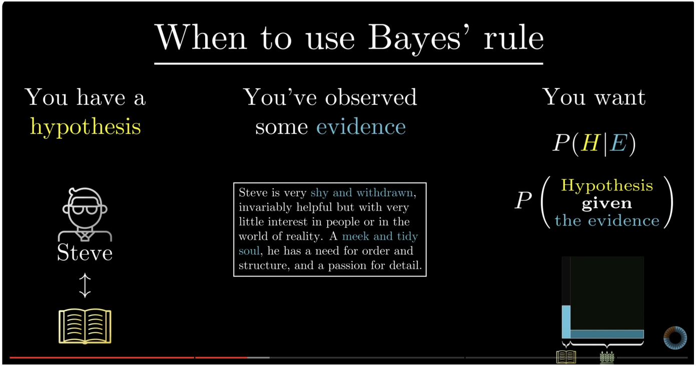

Figure: 3Blue1Brown Youtube Tutorial on Bayes Rule

Applying Bayes Rule to Second Example

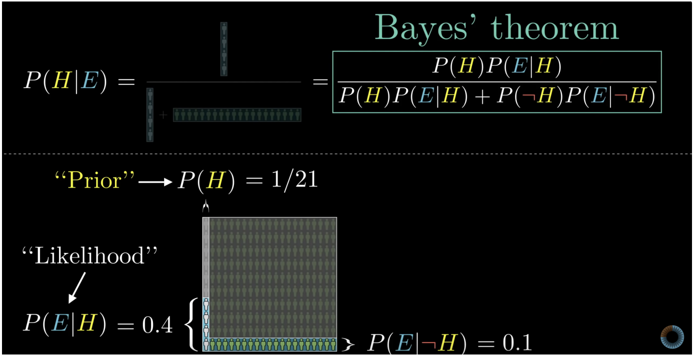

Figure: 3Blue1Brown Youtube Tutorial on Bayes Rule

Bayes' Rule: a Third Example[3]

Suppose there are two full bowls of cookies. Bowl #1 has 10 chocolate chip and 30 plain cookies, while bowl #2 has 20 of each. Our friend Fred picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a plain one. How probable is it that Fred picked it out of bowl #1?

Intuitively, it seems clear that the answer should be more than a half, since there are more plain cookies in bowl #1. The precise answer is given by Bayes' theorem. Let $H1$ correspond to bowl #1, and $H2$ to bowl #2. It is given that the bowls are identical from Fred's point of view, thus $P(H1) = P(H2)$, and the two must add up to 1, so both are equal to 0.5. The event E is the observation of a plain cookie.

[3] due to `https://en.wikipedia.org/wiki/Bayesian_inference`

Bayes' Rule: a Third Example (cntd.)

From the contents of the bowls, we know that
$P(E \mid H1) = \frac{30}{40} = 0.75$ and $P(E \mid H2) = \frac{20}{40} = 0.5$.

Bayes' formula then yields:

$$
\begin{aligned}
P(H1 \mid E) &= \frac{P(E \mid H1)P(H1)}{P(E \mid H1)P(H1) + P(E \mid H2)P(H2)} \\
&= \frac{0.75 \times 0.5}{0.75 \times 0.5 + 0.5 \times 0.5} = 0.6
\end{aligned}
$$

Bayes' Rule: a Fourth Example

Example 4.6.2. In the general population, 1/1000 infants have a particular genetic marker that is linked to cancer in later life. There is a test for this marker, but it isn't perfect. If an infant has the genetic marker, the test will always show a positive result. But it also has a false positive rate of 1% (that is, 1% of infants who do not have the marker will still show a positive test result). If we test a particular infant and the test comes out positive, what's the probability that infant actually has the genetic marker?

Bayes' Rule: a Fourth Example

We are told that P(G) = .001, where *G* stands for *infant has genetic marker*; that $P(T|\neg G) = .01$, where *T* stands for *test is positive* and $\neg G$ stands for *false positive rate*; and that $P(T|G) = 1$ (perfect detection). Now we want to compute $P(G|T)$. We begin by applying Bayes' Rule:

$$P(G|T) = \frac{P(T|G)P(G)}{P(T)}$$

Bayes' Rule: a Fourth Example

$$P(G|T) = \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\neg G)P(\neg G)}$$

$$= \frac{(1)(.001)}{(1)(.001) + (0.1)(.999)} \approx .09$$

Are you surprised that the probability is so low? It's because even though the false positive rate is low, there are far more infants who don't have the genetic marker than those who do, and altogether those infants generate quite a few false positives.

However, it's a well-known result from psychology that people are bad at estimating the answers to problems like this, so if you were surprised, you are in good company.

Bayes' Rule: a Fifth Example[4]

Suppose one is interested in a rare syntactic construction, perhaps parasitic gaps (see examples (1) below) , which occurs on average once in 100,000 sentences of English.

(1) a. Which articles did you file __ without reading them?
    b. Which articles did you file__ without reading __?
    c. * Which articles did you file them without reading __?

Joe Linguist has developed a complicated pattern matcher that attempts to identify sentences with parasitic gaps. It's pretty good, but not perfect: if a sentence has a parasitic gap, it will say so with probability .95, if it doesn't it will wrongly say it does with probability 0.005.

[4] due to: Christopher Manning and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. The MIT Press, p. 44

Bayes' Rule: a Fifth Example

Suppose the test says that a sentence contains a parasitic gap. What is the probability that this is true?

Solution: Let G be the event of the sentence having a parasitic gap, and let T be the event of the text being positive. We want to determine:

$$
\begin{aligned}
P(G|T) &= \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\neg G)P(\neg G)} \\
&= \frac{(0.95)(.00001)}{(0.95)(.00001) + (0.005)(0.99999)} \quad \approx .002
\end{aligned}
$$

Here we use having the construction or not as the partition in the denominator.

Bayes' Rule: a Fifth Example

Although Joe's test seems quite reliable, we find that using it won't help as much as one might have hoped. On average, only 1 in every 500 sentences that the test identifies will actually contain a parasitic gap. This poor result comes about because of the prior probability of a sentence containing a parasitic gap is so low.