Bayes' Theorem

Bayes' Theorem is one of the most ubiquitous results in probability for computer scientists. In a nutshell, Bayes' theorem provides a way to convert a conditional probability from one direction, say $P(E|F)$, to the other direction, $P(F|E)$.

Bayes' theorem is a mathematical identity which we can derive ourselves. Start with the definition of conditional probability and then expand the and term using the chain rule:

$$P(F|E) = \frac{P(F \text{ and } E)}{P(E)} \qquad \text{Def of conditional probability}$$

$$= \frac{P(E|F) \cdot P(F)}{P(E)} \qquad \text{Substitute the chain rule for } P(F \text{ and } E)$$

This theorem makes no assumptions about E or F so it will apply for any two events. Bayes' theorem is exceptionally useful because it turns out to be the ubiquitous way to answer the question: "how can I update a belief about something, which is not directly observable, given evidence." This is for good reason. For many "noisy" measurements it is straightforward to estimate the probability of the noisy observation given the true state of the world. However, what you would really like to know is the conditional probability the other way around: what is the probability of the true state of the world given evidence. There are countless real world situations that fit this situation:

Example 1: Medical tests
What you want to know: Probability of a disease given a test result
What is easy to know: Probability of a test result given the true state of disease
Causality: We believe that diseases influences test results

Example 2: Student ability
What you want to know: Student knowledge of a subject given their answers
What is easy to know: Likelihood of answers given a student's knowledge of a subject
Causality: We believe that ability influences answers

Example 3: Cell phone location
What you want to know: Where is a cell phone, given noisy measure of distance to tower
What is easy to know: Error in noisy measure, given the true distance to tower
Causality: We believe that cell phone location influences distance measure

There is a pattern here: in each example we care about knowing some unobservable -- or hard to observe -- state of the world. This state of the world "causes" some easy-to-observe evidence. For example: having the flu (something we would like to know) ***causes*** a fever (something we can easily observe), not the other way around. We often call the unobservable state the "belief" and the observable state the "evidence". For that reason lets rename the events! Lets call the unobservable thing we want to know B for belief. Lets call the thing we have evidence of E for evidence. This makes it clear that Bayes' theorem allows us to calculate an updated belief given evidence: $P(B|E)$

Definition: Bayes' Theorem

The most common form of Bayes' Theorem is **Bayes' Theorem Classic**:

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

There are names for the different terms in the Bayes' Rule formula. The term $P(B|E)$ is often called the "posterior": it is your updated belief of B after you take into account evidence E. The term $P(B)$ is often called the "prior": it was your belief before seeing any evidence. The term $P(E|B)$ is called the update and $P(E)$ is often called the normalization constant.

There are several techniques for handling the case where the denominator is not known. One technique is to use the law of total probability to expand out the term, resulting in another formula, called **Bayes' Theorem with Law of Total Probability**:

$$P(B|E) = \frac{P(E|B) \cdot P(B)}{P(E|B) \cdot P(B) + P(E|B^C) \cdot P(B^C)}$$

Recall the law of total probability which is responsible for our new denominator:

$$P(E) = P(E|B) \cdot P(B) + P(E|B^C) \cdot P(B^C)$$

A common scenario for applying the Bayes' Rule formula is when you want to know the probability of something "unobservable" given an "observed" event! For example, you want to know the probability that a student understands a concept, given that you observed them solving a particular problem. It turns out it is much easier to first estimate the probability that a student can solve a problem given that they understand the concept and then to apply Bayes' Theorem. Intuitively, you can think about this as updating a belief given evidence.

Bayes' Theorem Applied

Sometimes the (correct) results from Bayes' Theorem can be counter intuitive. Here we work through a classic result: Bayes' applied to medical tests. We show a dynamic solution and present a visualization for understanding what is happening.

Example: Probability of a disease given a noisy test

In this problem we are going to calculate the probability that a patient has an illness given test-result for the illness. A positive test result means the test thinks the patient has the illness. You know the following information, which is typical for medical tests:

Natural % of population with illness: 13

Probability of a positive result given the patient has the illness 0.92

Probability of a positive result given the patient does not have the illness 0.10

The numbers in this example are from the Mammogram test for breast cancer. The seriousness of cancer underscores the potential for bayesian probability to be applied to important contexts. The natural occurrence of breast cancer is 8%. The mammogram test returns a positive result 95% of the time for patients who have breast cancer. The test resturns a positive result 7% of the time for people who do not have breast cancer. In this demo you can enter different input numbers and it will recalculate.

Answer

The probability that the patient has the illness given a positive test result is: 0.5789

Natural Frequency Intuition

One way to build intuition for Bayes Theorem is to think about "natural frequences". Let's take another approach to answer the probability question in the above example on belief of illness given a test. In this take, we are going to imagine we have a population of 1000 people. Let's think about how many of those have the illness and test positive and how many don't have the illness and test positive. This visualization is based off the numbers in the fields above. Feel free to change them!

There are many possibilities for how many people have the illness, but one very plausible number is 1000, the number of people in our population, multiplied by the probability of the disease.
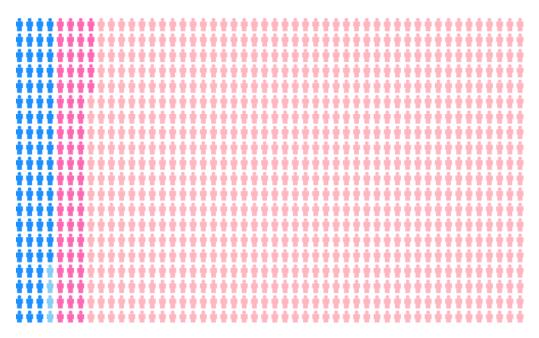
$1000 \times P(\text{Illness})$ people have the illness

$1000 \times (1 - P(\text{Illness}))$ people do not have the illness.

We are going to color people who **have the illness** in blue and those **without the illness** in pink (those colors do not imply gender!).

A certain number of people **with the illness will test positive** (which we will draw in Dark Blue) and a certain number of people **without the illness will test positive** (which we will draw in Dark Pink):

$1000 \times P(\text{Illness}) \times P(\text{Positive}|\text{Illness})$ people have the illness and test positive

$1000 \times P(\text{Illness}^C) \times P(\text{Positive}|\text{Illness}^C)$ people do not have the illness and test positive.

Here is the whole population of 1000 people:

The number of people who **test positive and have the illness** is 76.

The number of people who **test positive and don't have the illness** is 65.

The total number of people who test positive is 141.

Out of the subset of people who test positive, the fraction that have the illness is 76/141 = 0.5390 which is a close approximation of the answer. If instead of using 1000 imaginary people, we had used more, the approximation would have been even closer to the actual answer (which we calculated using Bayes Theorem).

Bayes with the General Law of Total Probability

A classic challenge when applying Bayes' theorem is to calculate the probability of the normalization constant $P(E)$ in the denominator of Bayes' Theorem. One common strategy for calculating this probability is to use the law of total probability. Our expanded version of Bayes' Theorem uses the simple version of the total law of probability: $P(E) = P(E|F) P(F) + P(E|F^c) P(F^c)$. Sometimes you will want the more expanded version of the [law of total probability](#): $P(E) = \sum_i P(E|B_i) P(B_i)$. Recall that this only works if the events B_i are mutually exclusive and cover the sample space.

For example say we are trying to track a phone which could be in any one of n discrete locations and we have prior beliefs $P(B_1) \ldots P(B_n)$ as to whether the phone is in location B_i. Now we gain some evidence (such as a particular signal strength from a particular cell tower) that we call E and we need to update all of our probabilities to be $P(B_i|E)$. We should use Bayes' Theorem!

The probability of the observation, assuming that the the phone is in location B_i, $P(E|B_i)$, is something that can be given to you by an expert. In this case the probability of getting a particular signal strength given a location B_i will be determined by the distance between the cell tower and location B_i .

Since we are assuming that the phone must be in *exactly* one of the locations, we can find the probability of any of the event B_i given E by first applying Bayes' Theorem and then applying the general version of the law of total probability:

$$P(B_i|E) = \frac{P(E|B_i) \cdot P(B_i)}{P(E)} \qquad \text{Bayes Theorem. What to do about } P(E)?$$

$$= \frac{P(E|B_i) \cdot P(B_i)}{\sum_{j=1}^{n} P(E|B_j) \cdot P(B_j)} \qquad \text{Use General Law of Total Probability for } P(E)$$

Unknown Normalization Constant

There are times when we would like to use Bayes' Theorem to update a belief, but we don't know the probability of E, $\mathrm{P}(E)$. All hope is not lost. This term is called the "normalization constant" because it is the same regardless of whether or not the event B happens. The solution that we used above was the law of total probability: $\mathrm{P}(E) = \mathrm{P}(E|B)\,\mathrm{P}(B) + \mathrm{P}(E|B^\mathrm{C})\,\mathrm{P}(B^\mathrm{C})$. This allows us to calculate $\mathrm{P}(E)$.

Here is another strategy for dealing with an unknown $\mathrm{P}(E)$. We can make it cancel out by calculating the ratio of $\frac{\mathrm{P}(B|E)}{\mathrm{P}(B^\mathrm{C}|E)}$. This fraction tells you how many times more likely it is that B will happen given E than not B:

$$
\begin{aligned}
\frac{\mathrm{P}(B|E)}{\mathrm{P}(B^\mathrm{C}|E)} &= \frac{\frac{\mathrm{P}(E|B)\,\mathrm{P}(B)}{\mathrm{P}(E)}}{\frac{\mathrm{P}(E|B^\mathrm{C})\,\mathrm{P}(B^\mathrm{C})}{\mathrm{P}(E)}} \qquad && \text{Apply Bayes' Theorem to both terms} \\[2mm]
&= \frac{\mathrm{P}(E|B)\,\mathrm{P}(B)}{\mathrm{P}(E|B^\mathrm{C})\,\mathrm{P}(B^\mathrm{C})} \qquad && \text{The term } \mathrm{P}(E) \text{ cancels}
\end{aligned}
$$