Derivation of Bayes Theorem

The following simple example shows how Bayes Theorem works. Let the rectangle in Figure 1 represent all motor vehicles purchased during the past year. Let "Event A" represent all of those vehicles that were red. Let "Event B" represent all of those vehicles that were made by Jeep. The space of all vehicles is divided into 4 regions.

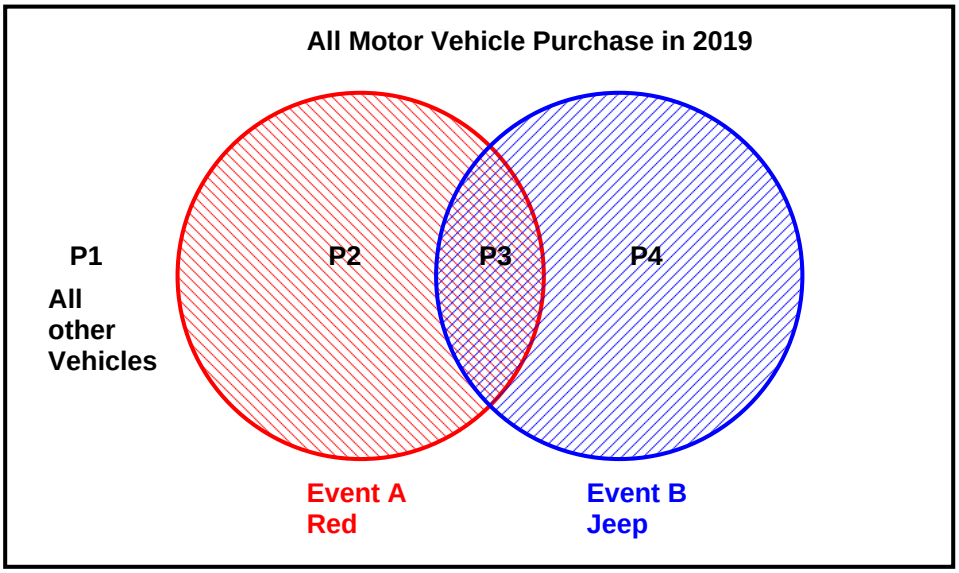

Figure 1. Pictorial representation of Bayes Theorem

The description of those regions along with the fraction (i.e., probability) of all such vehicles in each region is as follows:

$P1$: All vehicles that were not made by Jeep and were not red.
$P2$: All vehicles that were red but not made by Jeep
$P3$: All vehicles that were red and made by Jeep
$P4$: All vehicles that were made by Jeep but were not red.

The sum of all possibilities is unity:

$$P1 + P2 + P3 + P4 = 1$$

The fraction of all vehicles that were of type A (i.e. red) is given by $P(A) = P2 + P3$.

The fraction of all vehicles that were of type B (i.e., made by Jeep) is given by $P(B) = P3 + P4$.

The fraction of all vehicles that were of type A and of type B (i.e., red vehicles made by Jeep) is given by $P(AB) = P3$.

We define the conditional probability $P(A|B)$ as follows: Given that a vehicle was of type B (i.e., made by Jeep), what are the chances that it was also of type A (i.e., red)? In other words, for all vehicles in the blue circle, what fraction are also in the red circle? By inspection, we see that:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P3}{P3 + P4}.$$

This equation can be rewritten to have the form:

$$P(AB) = P(A|B) \times P(B).$$

Swapping A and B, this equation can also be written as

$$P(AB) = P(B|A) \times P(A)$$

Bayes theorem can finally be written as

$$P(AB) = P(A|B) \times P(B) = P(B|A) \times P(A).$$

With a bit of rearranging, the standard form for the theorem is given as:

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

Let us make up some numbers. Suppose 10% of all vehicles were red and 20% were made by Jeep. Suppose red vehicle buyers favored Jeeps more than other makes of vehicle such that 50% of all red vehicles were Jeeps.

We have the following:

$$\begin{aligned} P(A) =& P2 + P3 = 0.10 \\ P(B) =& P3 + P4 = 0.20 \\ P(B|A) =& \frac{P3}{P2 + P3} = 0.50 \end{aligned}$$

We can derive the following:

$$\begin{aligned} P3 =& P(AB) = P(B|A) \times P(A) = 0.50 \times 0.10 = 0.05 \\ P2 =& P(A) - P(AB) = 0.05 \\ P4 =& P(B) - P(AB) = 0.15 \\ P1 =& 1 - P2 - P3 - P4 = 0.75 \end{aligned}$$

Bayes Theorem expressed numerically is:

$$P(A|B) = P(B|A) \times \frac{P(A)}{P(B)}$$
$$\frac{P3}{P3+P4} = \frac{P3}{P2+P3} \times \frac{P2+P3}{P3+P4}$$

In other words, what this equation says is that the fraction of all Jeep buyers whose vehicle is red is given by the fraction of all red vehicle buyers whose vehicle is a Jeep times the ratio of the fraction of all red vehicle buyers divided by the fraction of all Jeep buyers.

$$\frac{0.05}{0.2} = \frac{0.05}{0.1} \times \frac{0.1}{0.2} = 0.25$$

Thus, we find that, of all Jeep buyers, 25% bought one that was red.

Often one applies Bayes theorem to cover all possible "hypotheses" - in the example, that means all possible colors of vehicles. Thus, one could designate each possible color as A_i, and then the sum of $P(A_i|B)$ over all values of i must be unity. The right side of the equation has the same property: the sum of $P(B|A_i) \times P(A_i)/P(B) = 1$. Note that, since the denominator does not depend on A_i, one can rewrite the right side as:

$$P(B) = \sum_i P(B|A_i) \times P(A_i).$$

This equation is sometimes useful in cases where the right side can be computed but $P(B)$ is not easily guessed otherwise.

Bayes Theorem is quite general, and the events A and B can correspond to any types of information. In the above example, events A and B are discrete, but one or the other can also be continuous variables. Commonly one has a set of data D and wants to determine the probability distribution for some parameter. In this case the hypothesis that the parameter has some particular value serves the role of event A above and the data serve the role of event B. One then writes:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}.$$

In the language of Bayes, $P(D|H)$ is called the likelihood, $P(H)$ is called the prior, $P(D)$ is called the evidence, and $P(H|D)$ is called the posterior. One again, H might be a single event or it might be an ensemble of events for which the sum of all probabilities is unity.

For example, suppose we have some variable x whose values (aside from a normalization factor) are given by a Gaussian distribution $f = \exp[-(x-x_0)^2/2]$, and we

want to estimate the mean of that distribution x_0, given a pair of measurements $D = [m_1, m_2]$. We write

$$P(x_0|D) = \frac{P(D|x_0) \times P(x_0)}{P(D)}.$$

We have $P(D|x_0) = \exp\{-[(m_1 - x_0)^2 + (m_2 - x_0)^2]/2\}$. If we have no a priori information about the mean, then $P(x_0) = 1$. (Sometimes, however, we do have prior information about x_0, e.g. as a result of previous measurements. In this case Bayes theorem can be used to update our knowledge as new information is collected.)

What is $P(D)$? It is the probability that one would have obtained a particular set of measurements. The most straighforward way to compute it is, once again, to make use of the fact that the integral of this equation over x_0 must be unity:

$$P(D) = \int \exp\{-[(m_1 - x_0)^2 + (m_2 - x_0)^2]/2\} dx_0$$

The integral can be evaluated analytically (it is, again aside from a normalization factor, $\exp\{-(m_1 - m_2)^2/2\}$), although in many cases it is simpler to compute numerically. (Note that the "evidence" depends on the actual measurements and the form of the equations that describe the hypothesis but not the parameters of the hypothesis themselves.) In this case one would find that the distribution function for x_0 is a Gaussian with a mean given by $\langle x_0 \rangle = (m_1 + m_2)/N$ (where $N = 2$ is the total number of points) and a standard deviation of $1/\sqrt{N}$.

A particular application of Bayes theorem involves deciding between two hypotheses, H_1 and H_2, one of which must be true. (This application is often called a Bayes Hypothesis Test.) One can write two equations:

$$\begin{aligned} P(H_1|D) &= \frac{P(D|H_1) \times P(H_1)}{P(D)} \\ P(H_2|D) &= \frac{P(D|H_2) \times P(H_2)}{P(D)} \end{aligned}$$

Taking the ratio, one get the "odds" of H_1 versus H_2:

$$\frac{P(H_1|D)}{P(H_2|D)} = \left[\frac{P(D|H_1)}{P(D|H_2)}\right] \times \left[\frac{P(H_1)}{P(H_2)}\right].$$

Here, the "evidence" $P(D)$ drops out. The first term in brackets on the right is sometimes called the "Bayes factor" or the "likelihood ratio."