Example of Bayes Theorem

Let's consider another example of the use of Bayes Theorem. This time we are given different information from the previous example.

In a certain clinic 0.15 of the patients have got the HIV virus. Suppose a blood test is carried out on a patient. If the patient has got the virus the test will turn out positive with probability 0.95. If the patient does not have the virus the test will turn out positive with probability 0.02.

If the test is positive what are the probabilities that the patient

a) has the virus

b) does not have the virus?

If the test is negative what are the probabilities that the patient

c) has the virus

d) does not have the virus?

Let's give the events the following labels

H = the patient has got the virus

P = the outcome of the test is positive

The question above gives us the following information.

$P(H) = 0.15$

$P(P \mid H) = 0.95$

$P(P \mid \overline{H}) = 0.02$

And we are asked to find the following

a) $P(H \mid P)$

b) $P(\overline{H} \mid P)$

c) $P(H \mid \overline{P})$

d) $P(\overline{H} \mid \overline{P})$

To find the first of these, $P(H|P)$, we can write down Bayes Theorem

$$P(H \mid P) = \frac{P(P \mid H)P(H)}{P(P)}$$

But we notice that we are given two of the values on the right-hand side $P(P|H)$ and $P(H)$ but we are not given $P(P)$ – the probability of a positive result. So we have to work this out ourselves.

Now there are two ways that a patient could have a positive result:

- he has the virus, and he gets a positive result – $H \wedge P$
- he does not have the virus, and he gets a positive result – $\overline{H} \wedge P$

We have to work out the probabilities of both these cases and add them together.

$$P(P) = P(H \wedge P) + P(\overline{H} \wedge P)$$

Now from the second axiom of probability we have

$P(H \wedge P) = P(P \mid H)P(H)$ and

$P(\overline{H} \wedge P) = P(P \mid \overline{H})P(\overline{H})$

And we are given these quantities in the question above. So, we can now work out $P(P)$

$$P(P) = P(P \mid H)P(H) + P(P \mid \overline{H})P(\overline{H}) = 0.95 \times 0.15 + 0.02 \times 0.85 = 0.1595$$

We can now substitute this into Bayes Theorem and obtain $P(H|P)$

$$P(H \mid P) = \frac{(P \mid H)P(H)}{(P \mid H)P(H) + P(P \mid \overline{H})P(\overline{H})} = 0.95 \times 0.15 / 0.1595 = 0.8934$$

The next part of the question above asks us to work out $P(\overline{H} \mid P)$. This is easy. It's just

$$P(\overline{H} \mid P) = 1 - P(H \mid P) \quad = 1 - 0.8934 = 0.1066$$

We are then asked to work out $P(H \mid \overline{P})$. Again we write down Bayes Theorem

$$P(H \mid \overline{P}) = \frac{P(\overline{P} \mid H)P(H)}{P(\overline{P})}$$

To work this out we need $P(\overline{P})$. But this is just *1-P(P)* and we worked out *P(P)* above. So, we can now work out $P(H \mid \overline{P})$

$P(H \mid \overline{P})$ = (0.05 × 0.15)/(1-0.1595) = 0.008923

The final part of the question asks for $P(\overline{H} \mid \overline{P})$. But again this is easy. It's just $1 - P(H \mid \overline{P})$

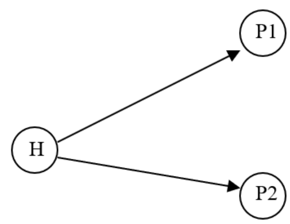
$P(\overline{H} \mid \overline{P})$ = 1- 0.008923 = 0.99107

Bayesian Networks

So far we have considered how Bayesian Probability Theory can relate two events – for example the probability that a man with a Porsche is a millionaire. But Bayesianism can be used to relate many events by tying them together in a network.

Let's consider the previous example again. Imagine that the patient is given a second test carried out independently of the first one. We can imagine the second test being carried out at a later date by a different experimenter using different equipment. So if there was an error on the first test then this does not affect the probability of an error on the second test. In other words the two tests are *independent*.

We can depict this using the following diagram



We consider event *H* to be the *cause* of the two events *P1* and *P2*. The arrows represent the fact that *H* is driving *P1* and *P2*. This diagram is a simple example of a Bayesian Network.

Suppose both *P1* and *P2* are positive. What then is the probability that the patient has the virus? In other words, we are asked to find $P(H \mid P1 \wedge P2)$

Just as before we can write down Bayes Theorem

$$P(H \mid P1 \wedge P2) = \frac{P(P1 \wedge P2 \mid H)P(H)}{P(P1 \wedge P2)}$$

Now there are two quantities here which we do not know immediately. The first is $P(P1 \wedge P2 \mid H)$ and the second is $P(P1 \wedge P2)$.

To work out $P(P1 \wedge P2 \mid H)$ we need to use the fact that the two tests are independent. Because they are independent we can write

P(P1^P2|H) = P(P1|H) P(P2|H)

We also need to find *P(P1^P2)*. Just as before when we were working out *P(P)* we need to break this down into two separate cases

- the patient has the virus and both tests are positive
- the patient does not have the virus and both tests are positive

Just as before we need to use the Second Axiom of Probability.

$$P(P1 \wedge P2) = P(P1 \wedge P2 \mid H)P(H) + P(P1 \wedge P2 \mid \overline{H})P(\overline{H})$$

But we have just seen that because the two tests are independent given H we can write

$$P(P1 \wedge P2) = P(P1 \mid H)P(P2 \mid H)P(H) + P(P1 \mid \overline{H})P(P2 \mid \overline{H})P(\overline{H})$$

$$= 0.95 \times 0.95 \times 0.15 + 0.02 \times 0.02 \times 0.85 = 0.135715$$

So we can substitute this into Bayes Theorem above and obtain

$$P(H \mid P1 \wedge P2) = \frac{0.95 \times 0.95 \times 0.15}{0.135715} = 0.99749$$

Previously we calculated the probability that the patient had HIV given one positive test as 0.8934. Now after two positive tests we see the probability has gone up to 0.99749. So after two positive tests we are much more certain that the patient does have the virus.

Now lets consider the case where one of the tests is postive and the other is negative. Clearly there must have been an error on one of the tests but we don't know which one. Can we conclude anything about whether the patient has the virus or not?

We need to calculate $P(H \mid P1 \wedge \overline{P2})$. We can do this using the same steps we went through above for the case of the two positive tests. We can write down Bayes Theorem

$$P(H \mid P1 \wedge \overline{P2}) = \frac{P(P1 \wedge \overline{P2} \mid H)P(H)}{P(P1 \wedge \overline{P2})}$$

Now we have to work out $P(P1 \wedge \overline{P2} \mid H)$ and $P(P1 \wedge \overline{P2})$. But as before we can use the fact that *P1* and *P2* are independent given *H*

$P(P1 \wedge \overline{P2} \mid H) = P(P1 \mid H)P(\overline{P2} \mid H)$ and

$$P(P1 \wedge \overline{P2}) = P(P1 \wedge \overline{P2} \mid H)P(H) + P(P1 \wedge \overline{P2} \mid \overline{H})P(\overline{H})$$
$$= P(P1 \mid H)P(\overline{P2} \mid H)P(H) + P(P1 \mid \overline{H})P(\overline{P2} \mid \overline{H})P(\overline{H})$$
$$= 0.95 \times 0.05 \times 0.15 + 0.02 \times 0.98 \times 0.85$$
$$= 0.023785$$

So we can substitute these values into Bayes Theorem and we obtain.

$$P(H \mid P1 \wedge \overline{P2}) = \frac{0.95 \times 0.05 \times 0.15}{0.023785} = 0.299$$

Notice that our belief in H has increased. Our prior belief was 0.15 but it has now gone up to 0.299. This may seem strange because we have been given two contradictory pieces of data. You might at first believe that the two pieces of data should cancel out and that our belief should not change. But if we look more closely we see that the probability of an error in each case is not equal.

The probability of a positive test when the patient is actually negative is 0.02. The probability of a negative test when the patient is actually positive is 0.05. Therefore we are more inclined to believe an error on the second test and this slightly increases our belief that the patient is positive.

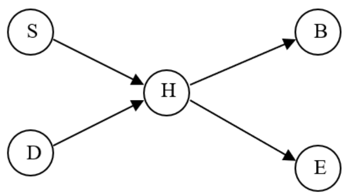
A More Complicated Bayesian Network

The above network contained three nodes. Let's look at a slightly more complicated one.

Suppose you are given the following facts about heart disease.

> Either smoking or bad diet or both can make heart disease more likely. Heart disease can produce either or both of the following two symptoms: high blood pressure and an abnormal electrocardiogram.

Here smoking and bad diet are regarded as causes of heart disease which in turn is a cause of high blood pressure and an abnormal electrocardiogram. So the appropriate network is



which uses the following symbols

S = smoking, D = bad diet, H = heart disease, B = high blood pressure,
E = abnormal electrocardiogram

Here H has two causes S and D. We need to know the probability of H given each of the four possible combinations of S and D. Let's suppose a medical survey gives us the following data.

$$P(S) = 0.3 \qquad P(D) = 0.4$$

$$P(H \mid S \wedge D) = 0.8$$
$$P(H \mid \overline{S} \wedge D) = 0.5$$
$$P(H \mid S \wedge \overline{D}) = 0.4$$
$$P(H \mid \overline{S} \wedge \overline{D}) = 0.1$$

$$P(B \mid H) = 0.7 \qquad P(B \mid \overline{H}) = 0.1$$
$$P(E \mid H) = 0.8 \qquad P(E \mid \overline{H}) = 0.1$$

Given this information we can now answer any question concerning this network. For example, suppose we want to know what the probability of heart disease is.

In a Bayesian Network if you wish to know the probability that node N, say, is true, you have to look at its parent nodes (i.e. its causes). You have to list all possible combinations of the values of the parent nodes and then consider the probability that N is true for each combination

For example, in the previous problem when we wished to calculate the probability of a positive test P we looked at its parent node H which had two possible values – either the patient had the virus or he didn't. This meant there were two ways in which P could be true: either the patient had the virus and tested positive or he did not have the virus and tested positive. We had to calculate the probability of each case and add them together.

In this problem we have to consider all the possible combinations of the two parent nodes S and D. In this case there are four of them, so there are four ways he can have heart disease:

> He smokes and has bad diet and he has heart disease
> He does not smoke and has bad diet and has heart disease
> He smokes and does not have bad diet and has heart disease
> He does not smoke and does not have bad diet and has heart disease

We have to work out the probabilities of all four situations and add them together. Just as in the previous problem we use the Second Axiom of Probability. For example the probability of each of these cases is

$$P(S \wedge D \wedge H) = P(H \mid S \wedge D)P(S)P(D)$$
$$P(\overline{S} \wedge D \wedge H) = P(H \mid \overline{S} \wedge D)P(\overline{S})P(D)$$
$$P(S \wedge \overline{D} \wedge H) = P(H \mid S \wedge \overline{D})P(S)P(\overline{D})$$
$$P(\overline{S} \wedge \overline{D} \wedge H) = P(H \mid \overline{S} \wedge \overline{D})P(\overline{S})P(\overline{D})$$

The values of all the quantities on the right-hand side are given above. So we can work out the probabilities of all four cases. $P(H)$ is just the sum of all four.

$P(H) = 0.8\times0.3\times0.4 + 0.5\times0.7\times0.4 + 0.4\times0.3\times0.6 + 0.1\times0.7\times0.6 = 0.35$

Suppose we wish to know the probability of heart disease and smoking, i.e. you are told that a patient smokes and has heart disease but you do not know whether he has bad diet or not. There are two ways in which a patient could have heart disease and smoke: they are

He smokes and has bad diet and he has heart disease
He smokes and does not have bad diet and has heart disease

But these are just the first and third case from the list above. So we have already worked out their probabilities. All we have to do is add them together.

$$P(H \wedge S) = 0.8{\times}0.3{\times}0.4 + 0.4{\times}0.3{\times}0.6 = 0.168$$

From this we can now work out the probability that a patient has heart disease given that he smokes $P(H \mid S)$. We use the Second Axiom

$$P(H \mid S) = \frac{P(H \wedge S)}{P(S)} = \frac{0.168}{0.3} = 0.56$$

Suppose we wish to know the probability of having an abnormal electrocardiogram (ECG) given that a patient smokes, i.e. you are told a patient smokes but you do not know whether he has bad diet or whether he has heart disease.

An abnormal ECG is caused by H heart disease which is in turn caused by smoking S. So we have a chain of cause and effect. But we do not know the value of the middle link H in this chain. So we have to consider each possible value of H as follows

He smokes and has heart disease and has abnormal ECG
He smokes and does not have heart disease and has abnormal ECG

We have to work out the probabilities of each case

$$P(E \mid S) = P(E \mid H)P(H \mid S) + P(E \mid \overline{H})P(\overline{H} \mid S)$$

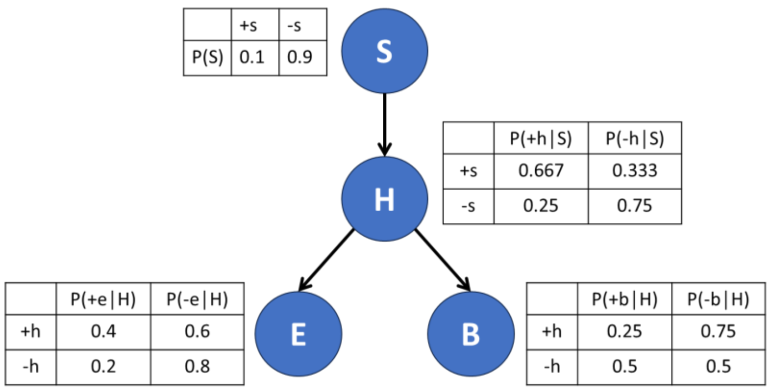
Now we are given $P(E \mid H)$ and $P(E \mid \overline{H})$ in the original data. And we have just worked out $P(H \mid S)$ immediately above. $P(\overline{H} \mid S)$ is just $1\text{-}P(H \mid S)$. So we have all the quantities on the right-hand side. So we can now work out $P(E \mid S)$

$P(E \mid S) = 0.8 \times 0.56 + 0.1 \times 0.44 = 0.492$

This shows that you can use the Bayesian Network to find the influence of one node on another. *S* influences *E* via *H*. So we first calculate the influence of *S* on *H* and then *H* on *E*. Because we do not know whether *H* is true or not in this case we must first calculate the probabilities of each value of *H* given *S* and then calculate the probability of *E* in each of those cases.

This is the reason Bayesian Networks are also called Influence Diagrams.

Another Example



	+s	-s
P(S)	0.1	0.9

	P(+h|S)	P(-h|S)
+s	0.667	0.333
-s	0.25	0.75

	P(+e|H)	P(-e|H)
+h	0.4	0.6
-h	0.2	0.8

	P(+b|H)	P(-b|H)
+h	0.25	0.75
-h	0.5	0.5

$$P(+h \mid +e, -b) = \frac{p(+e, -b, +h)}{p(+e, -b)}$$

$$= \frac{p(+e, -b, +h)}{p(+e, -b, +h) + p(+e, -b, -h)}$$

$$= \frac{p(+e, -b|+h) \times P(+h)}{p(+e, -b|+h) \times P(+h) + p(+e, -b|-h) \times P(-h)}$$

$$\begin{aligned} p(+h) &= p(+h, +s) + p(+h, -s) \\ &= p(+h| + s) \times p(+s) + p(+h| - s) \times p(-s) \\ &= 0.667 \times 0.1 + 0.25 \times 0.9 \\ &= 0.2917 \end{aligned}$$

$$\begin{aligned} p(-h) &= 1 - p(+h) \\ &= 0.7083 \end{aligned}$$

$$\begin{aligned} p(+e,-b|+h) &= p(+e|+h) \times p(-b|+h) \\ &= 0.4 \times 0.75 \\ &= 0.3 \end{aligned}$$

$$\begin{aligned} p(+e,-b|-h) &= p(+e|-h) \times p(-b|-h) \\ &= \mathbf{0.}2 \times \mathbf{0.5} \\ &= 0.1 \end{aligned}$$

$$P(\,+h \mid +e,-b\,) = \frac{p(+e,-b|+h) \times P(+h)}{p(+e,-b|+h) \times P(+h) + p(+e,-b|-h) \times P(-h)}$$

$$= \frac{0.3 \times 0.2917}{0.3 \times 0.2917 + 0.1 \times 0.7083}$$

$$= 0.55267$$