Tech Tunnel

https://ashutoshtripathi.com

Bayes’ Theorem for Data Science Professionals

Bayes Theorem is the extension of Conditional probability. Conditional probability helps us to determine the probability of A given B, denoted by P(A|B). So Bayes’ theorem says if we know P(A|B) then we can determine P(B|A), given that P(A) and P(B) are known to us.

In this post I am concentrating on Bayes’ theorem assuming you have good understanding of Conditional probability. In case you want to revise your concepts, you may refer my previous post on [Conditional probability with examples](#).

Formula derivation:

From conditional probability we know that

- ⇨ P(A|B) = P(A and B)/P(B)
- ⇨ P(A and B) = P(B) * P(A|B) --------------[1]

Similarly

- ⇨ P(B|A) = P(B and A)/P(A) = P(A and B)/P(A) [In Joint Probability order does not matter]
- ⇨ P(A and B) = P(A) * P(B|A) --------------[2]

From equation [1] and [2],

- ⇨ P(B) * P(A|B) = P(A) * P(B|A)
- ⇨ **P(A|B) = P(A) * P(B|A) / P(B)**

Which mean if we know P(A|B) then we can easily determine P(B|A) and vice versa. Assuming we know the total probabilities P(A) and P(B).

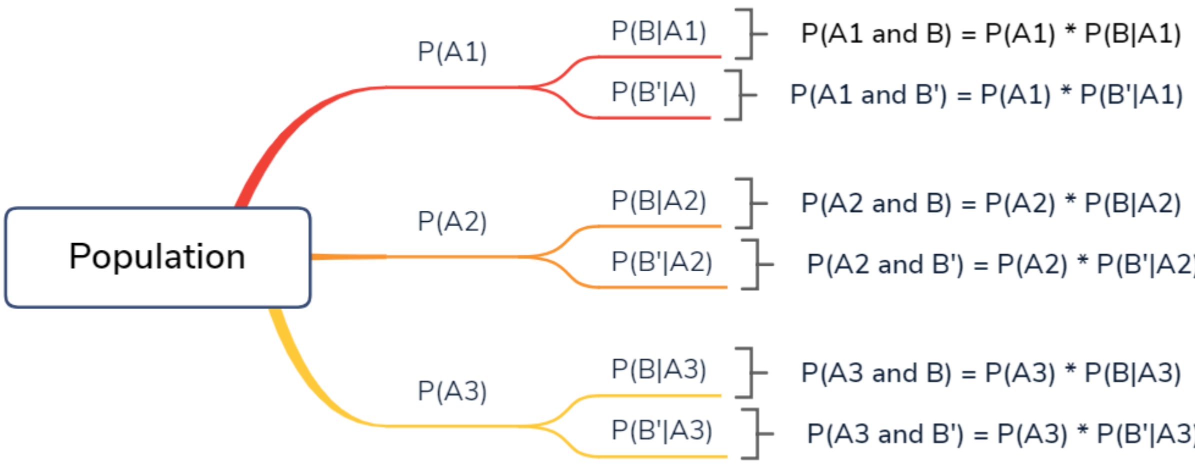
More Generalized Definition:

Let A1, A2, A3… Ak be a collection of mutually exclusive and exhaustive events with probability P(Ai), I = 1, 2, 3…k. Then for any event B for which P(B) > 0

P(Ai|B) = P(Ai and B)/P(B)

= P(B|Ai)*P(Ai)/∑[P(B|Ai)*P(Ai)

Below is the tree representation of the Bayes’ Theorem.



Bayes Theorem: To Find Reverse Probabilities

P(A1|B) = P(A1) * P(B|A1) / P(B)

P(A1) and P(B) are known as marginal probabilities.
P(B|A1) and P(A1) is given to us.
P(B) can be calculated as
P(B) = P(A1)*P(B|A1) + P(A2)*P(B|A2) + P(A3)*P(B|A3) and also known as Total Probability

Bayes Theorem: Find Reverse Probabilities

P(A1|B') = P(A1)*P(B'|A1 /P(B')

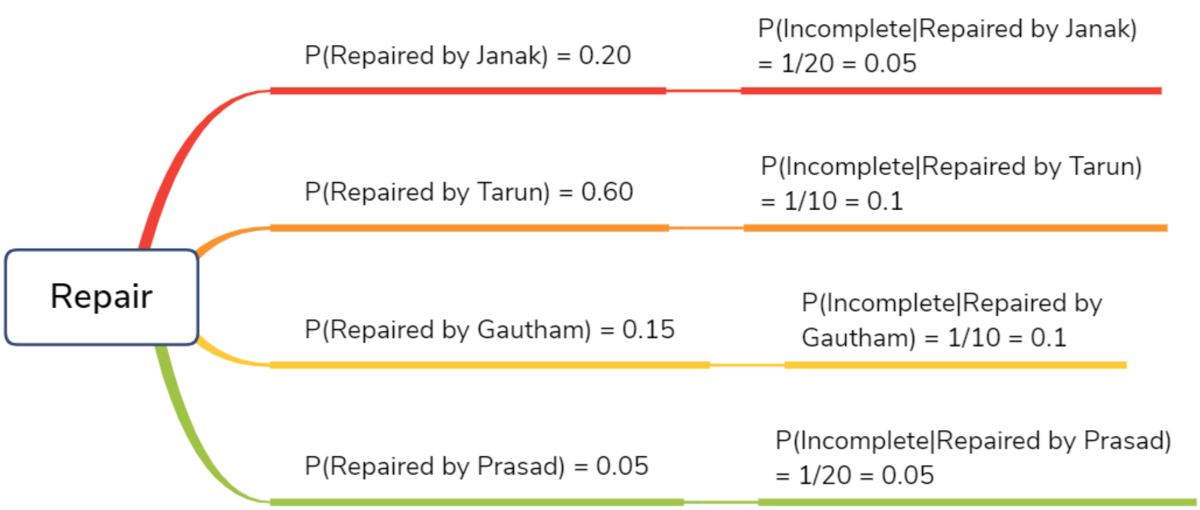
P(B') can be calculated as

P(B') = P(A1)*P(B'|A1) + P(A2)*P(B'|A2) + P(A3)*P(B'|A3)

Now let's solve some example to get a feeling of Bayes' theorem.

Example 1

Technicians regularly make repairs when breakdowns occur on an automated production line. Janak, who services 20% of the breakdowns, makes an incomplete repair 1 time in 20.Tarun, who services 60% of the breakdowns, makes an incomplete repair 1 time in 10. Gautham, who services 15% of the breakdowns, makes an incomplete repair 1 time in 10 and Prasad, who services 5% of the breakdowns, makes an incomplete repair 1 time in 20. For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janak?



Reverse Probability
P(Janak | Incomplete) = ?

Solution:
P(Janak | Incomplete) = P(Repaired by Janak)*P(Incomplete | Repaired by Janak)/P(Incomplete)
P(Janak | Incomplete) = 0.20*0.05/[0.20*0.05 + 0.60*0.1 + 0.15*0.1 + 0.05*0.05]

Example 2 Clinical trials

Epidemiologists claim that probability of breast cancer among Caucasian women in their mid-50s is 0.005. An established test identified people who had breast cancer and those that were healthy. A new mammography test in clinical trials has a probability of 0.85 for detecting cancer correctly. In women without breast cancer, it has a chance of 0.925 for a negative result. If a 55-year-old Caucasian woman tests positive for breast cancer, what is the probability that she in fact has breast cancer?

Solution:

Write what is given:

P(Cance) = 0.005

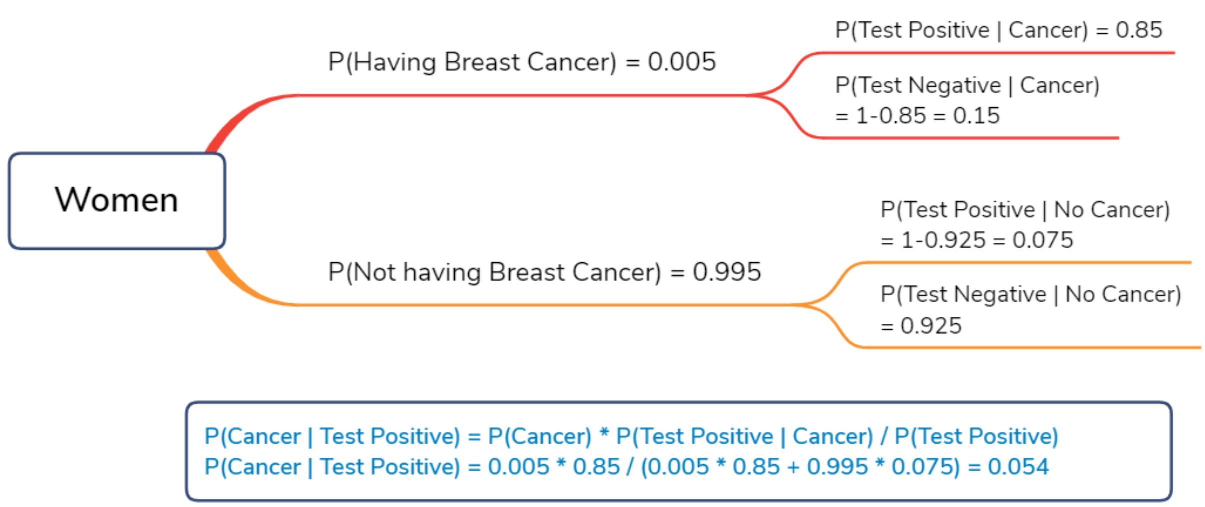
P(Test Positive | Cancer) = 0.85

P(Test Neg|No cancer) = 0.925

P(Cancer|Test Positive)= P(Cancer) * P(Test Positive | Cancer) / P(Test Positive)

Probability Table for the given question

	Probability of having Cancer or not	Test being Positive	Test being Negative
Cancer	0.005	0.005*0.85 = 0.00425	0.005*0.15 = 0.00075
No Cancer	0.995	0.995*0.075 = 0.074625	0.995*0.925 = 0.920375
Total	1.00	0.078875	0.921125



<u>Example 3</u>

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word “free” appears in 20% of the mails marked as spam. Assuming 0.1% of non-spam mail includes the word “free” and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word “free” appears in it.

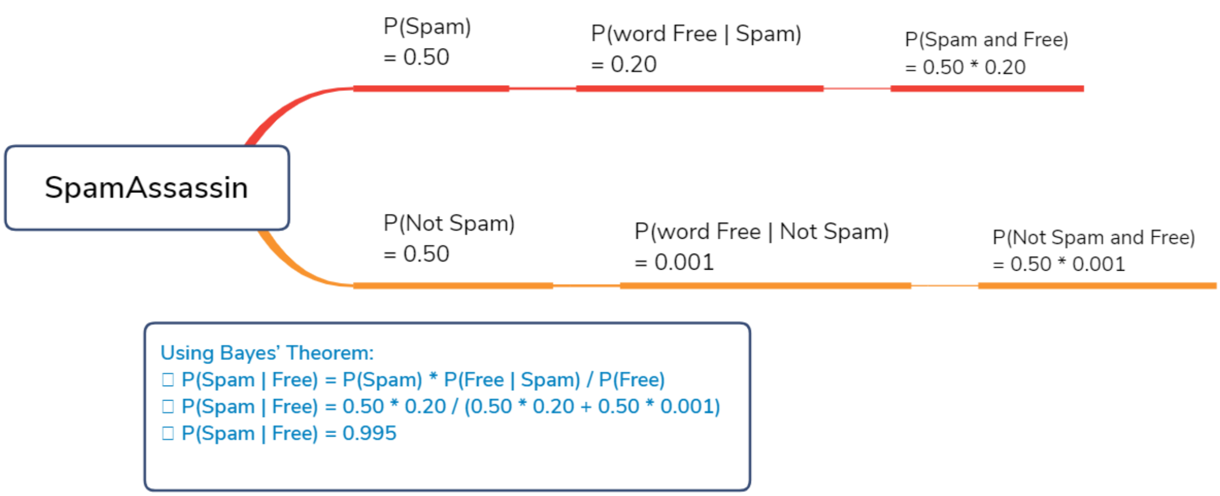
Data Given:

⇨ P(Free | Spam) = 0.20
⇨ P(Free | Non Spam) = 0.001
⇨ P(Spam) = 0.50 => P(Non Spam) = 0.50
⇨ P(Spam | Free) = ?

Using Bayes’ Theorem:

⇨ P(Spam | Free) = P(Spam) * P(Free | Spam) / P(Free)
⇨ P(Spam | Free) = 0.50 * 0.20 / (0.50 * 0.20 + 0.50 * 0.001)
⇨ P(Spam | Free) = 0.995

Now visualize using probability tree:



Thank You

For more Articles please visit -> https://ashutoshtripathi.com