Theological conundrums, lost bombs, and reverse conditioning: meet Bayes' Theorem

Federico Andreis
University of Xxxxxxxx
mail: federico.andreis@gmail.com

Probability (MATH08066)
5 March 2021

Course materials and resources

- Course main page [link]

- sensitive contents trigger warning [link]

- dedicated web forum - Piazza [link]

- (virtual) office hours: Tuesday 14-16 // please mail me [here] to book a slot!

A (very) diverse set of problems. . .

Consider the following:

- building a case for the abolition of capital punishment in late 1700s France

- recovering the daily wheel settings of the Enigma machine during WWII

- locating a hydrogen bomb on the Mediterranean seabed in the '60s

- invalidating the arguments that led to a life sentence for double infanticide in the 90s in the UK

- assessing the likelihood of having a disease, in light of a diagnostic test result

- creating an automatic filter for email spam.

. . . with something very special in common!

They can all be tackled using methods, whose core functioning is based on the same probabilistic result, universally known as **Bayes' theorem**.

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

where, following the usual notation, A and B denote events and $P(A) > 0$.

The result follows from the commutative property of the intersection operator, and the definition of conditional probability - more on this later.

Car insurance - encore

Let us get back to the car insurance example we have used in our previous lecture to play around with conditional probabilities.

	B: prone	B^c: not prone	
A: accident	20	30	50
A^c: no accident	30	120	150
	50	150	200

We have already learned how to derive joint and conditional probabilities starting from this contingency table.

A more interesting situation

Consider a more realistic context of incomplete information. Specifically, imagine we only have the conditional distributions of having a car accident, given proneness.

	B: prone	B^c: not prone	
A: accident	**0.4**	**0.2**	
A^c: no accident	**0.6**	**0.8**	
	1	1	

From historical data, we also happen to know that about 25% of all new policyholders tend to be accident-prone.

What about the probability of **being accident-prone, given that an accident occurred**?

A more interesting situation

Consider a more realistic context of incomplete information. Specifically, imagine we only have the conditional distributions of having a car accident, given proneness.

	B: prone	B^c: not prone	
A: accident	**0.4**	**0.2**	
A^c: no accident	**0.6**	**0.8**	
	1	1	

From historical data, we also happen to know that about 25% of all new policyholders tend to be accident-prone.

What about the probability of **being accident-prone, given that an accident occurred**?

Reversing the conditioning with Bayes' Theorem

Let's select a new policyholder at random from the pool who had an accident in the first year (A). What is the chance that the policyholder is prone to accidents (B)?

	B: prone	B^c: not prone	
A: accident	**?**	**?**	1
A^c: no accident	**?**	**?**	1

We want $P(B|A)$, and we know $P(B) = 0.25$, $P(A|B) = 0.4$. By Bayes' theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.4 \times 0.25}{P(A)}.$$

We still need the unconditional probability of having an accident $P(A)$.

A useful decomposition

Using the law of total probability from last lecture, we can write

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

Hence

$$
\begin{aligned}
P(\text{accident}) =& P(\text{accident}|\text{prone})P(\text{prone})+ \\
& +P(\text{accident}|\text{not prone})P(\text{not prone}) \\
=& 0.4 \times 0.25 + 0.2 \times (1 - 0.25) = 0.25.
\end{aligned}
$$

leading to

$$P(B|A) = \frac{0.4 \times 0.25}{P(A)} = \frac{0.4 \times 0.25}{0.25} = 0.4.$$

So, we believe that policyholders who had an accident during the first year, have a 40% chance of belonging to the accident-prone group.

A different angle...

We can express Bayes' theorem in an equivalent, yet surprisingly revealing way. Specifically, we can write

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

An initial probability $P(B)$ is **updated** into a conditional probability $P(B|A)$ after taking into account the event A occurring. In our insurance example, it means reviewing our belief on a policyholder being prone to accident, after having learned of an accident.

While straightforward to understand, this has profound implications, as it provides an interpretation of Bayes' theorem as describing a **mechanism of learning from experience**.

A medical example: diagnostic tests

A new home HIV test is claimed to have 95% sensitivity and 96% specificity, i.e. for people with HIV 95% will get a positive test, while for people without HIV, 96% will get a negative test.

The test is to be used in a population with an HIV prevalence of $1/100$.

Question: if someone gets tested and a positive result is obtained, what is the chance they actually have the disease?

Diagnostic tests: solution

We are interested in $P(\text{HIV}|\text{positive})$. Now,

- sensitivity: $P(\text{positive}|\text{HIV}) = 0.95$
- specificity: $P(\text{negative}|\text{no HIV}) = 0.96$
- prevalence: $P(\text{HIV}) = 0.01$.

$P(\text{HIV}|\text{positive}) = P(\text{positive}|\text{HIV})P(\text{HIV})/P(\text{positive})$ by Bayes' theorem, so all we need is $P(\text{positive})$:

$$
\begin{aligned}
P(\text{positive}) &= P(\text{positive}|\text{HIV})P(\text{HIV}) + \\
&\quad + P(\text{positive}|\text{no HIV})P(\text{no HIV}) \\
&= 0.95 \times 0.01 + (1 - 0.96) \times (1 - 0.01) = 0.0491
\end{aligned}
$$

hence $P(\text{HIV}|\text{positive}) = 0.95 \times 0.01/0.0491 \approx 0.19$.

Derivation of Bayes' theorem

By the commutative property of the intersection operator, it follows that

$$P(A \cap B) = P(B \cap A)$$

applying the chain rule to both sides yields

$$P(A|B)P(B) = P(B|A)P(A)$$

or, equivalently,

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

Generalisation to more than two events

In both theory and applications, it is often the case that we are interested in looking at more than two conditioning events.

Bayes' theorem can be extended to account for such situations. Consider a partition of the sample space $\{A_1, ..., A_k\}$, and let B denote an event of interest. Then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)} = \frac{P(B|A_i)}{P(B)}P(A_i), \quad P(B) > 0.$$

The denominator is once again derived using the law of total probability.

Who is responsible for Bayes' theorem?

While this might seem a trick question, Bayes' theorem has a long and interesting story. Did you know that:

- the result is named after the reverend Thomas Bayes (1702-1761) who, however, only proposed a solution for a very special case, and never published it

- Bayes' work containing the original seed of the theorem was polished and published, posthumously, by his friend, editor and American War of Independence's hero Richard Price (1723-1791)

- the formulation we know and use today is due to French scientist Pierre-Simon Laplace (1749-1827), who stripped the original problem of its religious connotation

- Bayes' theorem appears, therefore, to be yet another case of a mathematical result obeying Stigler's law of eponymy.

Interesting readings

- "The theory that would not die - how bayes' rule cracked the enigma code, hunted down russian submarines, emerged triumphant from two centuries of controversy" - McGrayne (2012) book review

- "An Essay towards solving a Problem in the Doctrine of Chances" - Bayes, Price (1763) original essay

- "Memoir on the Probability of the Causes of Events" - Stigler, Laplace (1986) translation of the original essay

- "Edward Simpson: Bayes at Bletchley Park" - Simpson (2010) article

- "Royal Statistical Society concerned by issues raised in Sally Clark case" - RSS Press Office (2001) article

Let's wrap up

In today's lecture we have learned about one of the most versatile and widely used probabilistic results: Bayes' theorem. The key takeaway points are:

1. it allows to express conditional probability statements in terms of the reverse conditioning

2. it has been (and keeps being) successfully applied to a wildly diverse set of problems, both applied and theoretical

3. while algebraically straightforward, it has profound implications if we choose to look at it as a learning paradigm.

In the next lecture, we will introduce and formalise the concept of random variable and its distribution.