3 Bayes' theorem

Independent events and some laws of probability

Two events A and B are said to be **independent** if

$$P(A \cap B) = P(A)P(B) \tag{1}$$

This is equivalent to says that two events are independent if

$$P(A|B) = P(A) \tag{2}$$

or

$$P(B|A) = P(B) \tag{3}$$

What it says is that the probability A and B happen together is just the probability that A happens multiplied by the probability B happens; so, for example, A happening doesn't change the probability that B happens.

Here is an example a slightly complicated example. Imagine a very boring game of like snakes and ladders with no snakes and no ladders. Every time it is your go you flip a coin, if you get a harp you go forward one step, if you get the other side you stay put. Hence for example, three rounds into the game your locations have probabilities

	0	1	2	3
P	1/8	3/8	3/8	1/8

Let S_3^3 be the event you are on the third square after three rounds, so $P(S_3^3) = 1/8$ and, so, if S_3^2 is the event you are at the third square after three rounds the $P(S_3^2) = 3/8$ because it corresponds to three possible sequences (H, H, T), (H, T, H) and (T, H, H), out of eight possible sequences in all. Let S_2^2 be the event that you are at square two after two goes, clearly $P(S_2^2) = 1/4$. Now

$$P(S_3^3|S_2^2) = 1/2 \tag{4}$$

which is different from $P(S_3^3)$ so your position after two moves is not independent from your position after three. It is similarly true that your position after three moves is not independent of your position after one move.

Before considering Bayes' theorem it is useful to note two laws of probability. The multiplicative law is

$$P(A \cap B) = P(A)P(B|A) \tag{5}$$

which follows directly from the definition of conditional probability. The additive law is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{6}$$

Finally, a corollary of the addative law tells us

$$P(A) = 1 - P(\bar{A}) \tag{7}$$

where, recall, $\bar{A}$ is the complement set to A, it is all the outcomes in the sample space that are not in A.

Bayes' theorem

Consider the formula for the conditional probability:

$$P(A|B) = \frac{P(A \cup B)}{P(B)} \tag{8}$$

which is equivalent to

$$P(A|B)P(B) = P(A \cup B) \tag{9}$$

which tells us that the probability of A and B is the probability of B multiplied by the probability of A given B. This makes lots of sense, but it is also notable that the left hand side doesn't look symmetric in A and B while the right hand side clearly is. Obviously this means we can write

$$P(A|B)P(B) = P(B|A)P(A) \tag{10}$$

or

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{11}$$

This formula is called the Bayes' rule and is surprisingly useful because there are lots of interesting problems where we are told $P(B|A)$ but would like to know $P(A|B)$.

Often the example given is related to testing. Lets say 5% of steaks sold as beef steak are actually made of horse and imagine we have a horsiness test which is positive 90% of the time when tested on horse and 10% of the time when tested on beef. If a piece of steak tests positive for horse, what is the chance it is horse? Let H be the event of being horse and Y the event of testing positive for horsiness. Now we know $P(H) = 0.05$ and $P(Y|H) = 0.9$; what we want is $P(H|Y)$ and this is what Bayes' rule is useful for:

$$P(H|Y) = \frac{P(Y|H)P(H)}{P(Y)} \tag{12}$$

We don't have $P(Y)$ but we can work it out:

$$P(Y) = P(Y|H)P(H) + P(Y|\bar{H})P(\bar{H}) \tag{13}$$

since $P(Y \cup H) = P(Y|H)P(H)$ and so on. Hence

$$P(Y) = 0.9 \times 0.05 + 0.1 \times 0.95 = 0.14 \tag{14}$$

Thus

$$P(H|Y) = \frac{0.9 \times 0.05}{0.14} = 0.32 \tag{15}$$

Hence, surprisingly, if a steak tests positive for horsiness it is still more likely to be beef. Basically, because there are so many more beef steaks than horse steaks, the relatively small false positive rate for beef still leads to a reasonably high chance a piece of steak that tests positive for horse is nonetheless beef.

There is a particular terminology associated with Bayes' rule; it is sometimes written:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{16}$$

The *posterior* is the probability estimated after the evidence is gathered, for example, the chance of horsiness after we have found the test is positive. The *likelihood* is how likely the evidence is given the event, in the example above, it is $P(Y|H)$; the *prior* is the probability estimated before the evidence is gathered, that is $P(H)$, finally *evidence* measure the probability of the evidence, $P(Y)$.

Naïve Bayes estimator

Many learning algorithms can be thought of have machines for estimating probabilities, often in the face of insufficient data to estimate the probabilities required. A common example used to illustrate this is a spam filter. Let W represent an ordered list of words that may be in a email, say:

$$W = (\text{enlargement}, \text{xxx}, \text{cheapest}, \text{pharmaceuticals}, \text{satisfied}, \text{leeds}) \tag{17}$$

and say $\mathbf{w}$ is a vector of zeros and ones indicating the presence or absence of different potential spam words in an email. Thus, an email that includes the words ‘enlargement‘, ‘xxx’ and ‘leeds’ but not ‘cheapest’, ‘pharmaceuticals’ and ‘satisfied’ would be represented by

$$\mathbf{w} = (1, 1, 0, 0, 0, 1) \tag{18}$$

Now let S represent the event of an email being spam. The objective with a spam filter is to estimate $P(S|\mathbf{w})$ for every possible vector $\mathbf{w}$ and then use a cut-off to label any email with a high probability of being spam as ‘spam’.

Obviously if you have a truly huge amount of data you could estimate this probability by counting:

$$P(S|(1, 1, 0, 0, 0, 1)) = \frac{\#\{\text{spam emails with the words enlargement, xxx and leeds}\}}{\#\{\text{all emails with the words enlargement, xxx and leeds}\}} \tag{19}$$

However there are $2^6 = 64$ possible $\mathbf{w}$ vectors, and of course in a real example you’d need many more than six words, thus, for anything but an infeasibly large data set, the amount of emails with the precise combination of words represented by a given $\mathbf{w}$ will be tiny, leading to a poor estimate of the probabilities.

An alternative approach is to use Bayes’ rule to get

$$P(S|\mathbf{w}) = \frac{P(\mathbf{w}|S)P(S)}{P(\mathbf{w})} \tag{20}$$

This doesn’t look any better, $P(\mathbf{w}|S)$ is no easier to estimate than $P(S|\mathbf{w})$. However, in the naïve Bayes estimator it is additionally assumed that the different words are independent so that

$$\begin{aligned} P((1, 1, 0, 0, 0, 1)|S) &= P(\text{enlargement}|S)P(\text{xxx}|S)[1 - P(\text{cheapest}|S)]\times \\ &\quad [1 - P(\text{pharmaceuticals}|S)][1 - P(\text{satisfied}|S)]P(\text{leeds}|S) \end{aligned} \tag{21}$$

This is clearly inaccurate, a spam email is ‘enlargement’ is more likely to contain ‘satisfied’ than one that doesn’t, that is why it is a ‘naïve’ classifier. The advantage though is that the individual probabilities are much easier to estimate, there will be more emails with ‘leeds’ in that there will be emails with the exact combination of words represented by $(1, 1, 0, 0, 0, 1)$ and so counting occurrences will be much more accurate.

Conditional probability

Lets return to our slightly odd example of independence involving the boring version of snakes and ladders. What is $P(S_1^1|S_2^1)$? In other words, if you observe that someone is on the first square after two moves, what is the probability that they were on the first square after one move. Well

$$P(S_1^1|S_2^1) = \frac{P(S_2^1|S_1^1)P(S_1^1)}{P(S_2^1)} \tag{22}$$

and we can calculate all these quantities: $P(S_2^1|S_1^1) = 1/2$, and $P(S_1^1) = 1/2$ as well, as is $P(S_2^1)$, so

$$P(S_1^1|S_2^1) = 1/2 \tag{23}$$

Similarly $P(S_3^2|S_2^1) = 1/2$. Finally consider the probability $P(S_3^2 \cap S_1^1|S_2^1)$. This is the probability of the sequence (H, T, H) when there are four sequence with S_2^1: (T, H, H), (T, H, T), (H, T, H) and (H, T, T). Thus

$$P(S_3^2 \cap S_1^1|S_2^1) = 1/4 \tag{24}$$

and hence

$$P(S_3^2 \cap S_1^1|S_2^1) = P(S_1^1|S_2^1)P(S_3^2|S_2^1) \tag{25}$$

This is an example of conditional independence: two events A and B are **conditionally independent** conditional on a third event C is

$$P(A \cap B|C) = P(A|C)P(B|C) \tag{26}$$

In this case A and B might be related to each other, but only through C, so if we know C then A and B are independent.

This applies to the snakes and ladders example; S_1^1 and S_3^2 are dependent. To work out the probability $P(S_1^1 \cap S_3^2)$ we can count the number of sequence of heads and tails this holds for: (H,T,H) and (H,H,T), two out of eight possible sequences so

$$P(S_1^1 \cap S_3^2) = 1/4 \tag{27}$$

However, since $P(S_3^2) = 3/8$ we see that

$$P(S_1^1 \cap S_3^2) \neq P(S_1^1)P(S_3^2) \tag{28}$$

Thus S_1^1 and S_3^2 are not independence events. It is clear why this is, the result after the first round affects the results after two rounds, and that in turn affects the result after three. However if we specify the result after two rounds, there is no further dependence of the result after the first round and the result after the third, they are related to each only through the result after the second round.

The reason to mention this here is that it is part of the idea behind a Markov chain. Markov chains go beyond this unit but roughly speaking a Markov chain encodes a set of conditional independence relationships between events. They are used to model many statistical processes. A Markov chain model of language, for example, might assume the probability of the next word in the sentence depends only on the previous one. For example, if you look at the sentence ‘Don’t forget you are going to Aunt Alicia’s’ a Markov model asked to predict the sixth word would look at the word ‘going’ and use the probability table for how often different words come after ‘going’; it wouldn’t consider the words coming earlier in the sentence, the ‘Don’t’ and ‘forget’ and so on. This restriction is clearly useful from a computational point of view; by assuming the word after ‘going’ only depends on the earlier words through the word ’going‘ the problem of language predictions is reduced to calculating a large, but not impossibly large, probability table. This model doesn’t work well, it is a very primitive model of language.

More sophisticated models, and until a few years ago the best models of language, use a Hidden Markov Model, which nonetheless is designed around assumptions about conditional independence.