Probability for AI students who have seen some probability before but didn't understand any of it

"I am inflicting these proofs on you for two reasons:

1. These kind of manipulations will need to be second nature to you if you use probabilistic analytics in depth

2. Suffering is good for you" -Andrew W. Moore

1. Basic probabilty is straightforward, but mathematicians have made it incredibly complicated by starting with formal axioms, skipping some fundamental properties and using inconsistent notation. They do this because there are deep and beautiful connections to other areas of mathematics, such as set theory and logic, and this approach makes that clear. Unfortunately, it ignores the simple fact that if you don't already know these other ares, the connections are totally lost on you, and instead just confuse you. We'll try to fix this problem here.

2. Why probablity?

   (a) Genuine randomness - the results of quantum events.

   (b) Qualification - There are exceptions to rules we can't or don't want to list. "If I look both ways before crossing the street, I'll get to the other side safely..." is a fine rule, but it has lots of exceptions: "...if I don't get hit by a meteor, and I didn't miss that truck, and I didn't go even though I saw that truck, and I don't have a heart attack, and the world doesn't end...." Its easier to just say, "If I look both ways before crossing the street, I'll probably get to the other side safely."

   (c) Lack of Knowledge - We don't know enough information to reliable predict what's going to happen. This is what really is happening for dice and cards. Its not that the events are really random, but that we don't know the details enough to predict what will happen.

3. We're concerned here about lack of knowledge. Can we tell what state we're in without enough detail to know for sure.

   (a) A **random variable** is an event, such as an election.

   (b) Random variables take **values**, known as **outcomes**, such as So-and-so winning the election.

   (c) Each of the possible outcomes of an event gets a number p assigned to it, such that each number is $0 \leq p \leq 1$.

(d) The number is a prediction of the outcome of the event. These numbers reflect a measure of the liklihood that this outcome will be the outcome of the event. It's a prediction of the future. If it is a reapeating event, it says, "if this same situation comes up n times, then we expect this to be the outcome $p \times n$ times." Thus if the probablity of me having a good day is 0.6, then we would expect me to have 3 good days every 5 day work week.

(e) We'll write these numbers as: $P(E \leftarrow o) = p$. The probability that event E has outcome o is the number p.

4. If we look at the collection of all possible outcomes (the distribution of probabilities), we see something like the following table for the event of "what is the status of spare tire after we try to attach it to a car axle?":

$$P(\mathit{Spare} \leftarrow \mathit{tight}) = 0.6$$
$$P(\mathit{Spare} \leftarrow \mathit{loose}) = 0.3$$
$$P(\mathit{Spare} \leftarrow \mathit{off}) = 0.1$$

5. OK, where do the numbers come from? From data. We observe the world, make a table and count the number of times each outcome occurs dividing by the total number of samples.

data item number	1	2	3	4	5	6	7	8	9	10
status	tight	loose	tight	off	loose	tight	tight	tight	loose	tight

6. We will sometimes write the distribution as a vector (note the bold $\mathbf{P}$): $\mathbf{P}(S) = \langle 0.6, 0.3, 0.1 \rangle$ When doing this we have to know that the first number is for the outcome of tight, the second is for loose, and the third is for off. We'll also try to use single capital letters to represent events, and single lowercase letters for outcomes.

7. Sometimes we want to combine 2 events, "what is the probability of E_1 being o_1 AND E_2 being o_2?"

(a) Often, this is to combine *evidence* of something with the question we really want to know.

(b) For example, if we want to know if the spare in on tight, and we already know the wheel wobbles, we might ask what it the probability that the wheel is on tight and it still wobbles: $P(\mathit{Spare} \leftarrow \mathit{tight} \wedge \mathit{Connection} \leftarrow \mathit{wobbly}) =?$ This is still just a number.

(c) This comes from the data, just as the individual distributions do:

item number	1	2	3	4	5	6	7	8	9	10
Spare on Status	tight	loose	tight	off	tight	tight	loose	loose	loose	loose
Connection Firmness	solid	wobly	solid	wobly	solid	wobly	solid	wobly	wobly	wobly

item number	11	12	13	14	15	16	17	18	19	20
Spare on Status	tight	tight	tight	tight	off	tight	tight	loose	tight	tight
Connection Firmness	solid	solid	solid	solid	solid	wobly	solid	wobly	solid	solid

(d) We will sometimes indicate the whole table as $\mathbf{P}(CS)$

(e) We can also speak about the distribution on the conjunction. This is a two dimensional table known as the **joint distribution.**

		Spare		
		tight	loose	off
connection	wobly	0.1	0.25	0.05
	solid	0.5	0.05	0.05

(f) We just count all the times both outcomes occur in the data, divided by the number of data.

8. Marginalization

   (a) Marginalization is a fancy word for "adding up all the numbers in a row or column."

   (b) We do this to reconstuct the separate distributions from the joint table:

		Spare			
		tight	loose	off	margin
Connection	wobly	0.1	0.25	0.05	0.4
	solid	0.5	0.05	0.05	0.6
	margin	0.6	0.3	0.1	1

   (c) What do the margins mean? They remove the effects of one of the two events. So if we add up all the elements in the row where connection status is wobly, then what we get is $P(\textit{Connection} \leftarrow \textit{wobly})$. Go back and look at the original data and count up all the times you see 'wobly'. Its 8/20 or 0.4, just as the marginalization tells us. If you think about this for a minute, it should be clear from the data why this is so. Each of the columns is made up of the instances where connection status is wobly and some other property is true, divided by the number of data. When we add them all up, we get the number of times wobly occurs, divided by the number of data. If $N(x, y)$ is the number of times both x and y occur in the same datum, then

$$\begin{aligned} P(C \leftarrow w) =& \frac{N(w)}{n} \\ =& \frac{N(w,t) + N(w,l) + N(w,o)}{n} \\ =& \frac{N(w,t)}{n} + \frac{N(w,l)}{n} + \frac{N(w,o)}{n} \end{aligned}$$

   and that last line is the marginalization.

   (d) Marginalization seems obvious, but it turns out to be useful to reconstruct the individual probabilities from the joint.

9. You probably remember "*and* means multiply and *or* means add".

   (a) This is wrong, but just right enough to get you into serious trouble!

   (b) Look at the table. We know $P(C \leftarrow w \wedge S \leftarrow t)$ is 0.1.

   (c) but if we multiply the two we get $P(C \leftarrow w) \times P(S \leftarrow t) = 0.6 \times 0.4 = 0.24$, which we know is wrong.

   (d) See? *and* is **NOT** multiplying!

   (e) What about *or*?

   (f) Well, what about the question $P(C \leftarrow s \vee S \leftarrow t)$? Looking at the table and just adding them together, we get 1.2! That can't happen!!!!

10. So what's really going on?

(a) For *or* the rule is, "*or* means add when the or is across values of a *single* variable or a conjunction of variables"

(b) To get $P(S \leftarrow t \vee S \leftarrow l)$ we **do** add them up: $0.6 + 0.3 = 0.9$.

(c) If we add them all up, we have to get 1: $P(S \leftarrow t \vee S \leftarrow l \vee S \leftarrow o) = 1$. This has to be the case since all the data have one of those three outcomes, so when we count up the number of data in that table that have that property, it will be all of them. and since we divide by the number of data, we get 1. This is often given as an axiom of probability, which is why you never could understand it before– it was their fault, not yours.

(d) The question $P(C \leftarrow s \vee S \leftarrow t)$ is not really well-formed. What we're really interested in is $P((C \leftarrow s \wedge S \leftarrow t) \vee (C \leftarrow s \wedge S \leftarrow l) \vee (C \leftarrow s \wedge S \leftarrow o) \vee (C \leftarrow w \wedge S \leftarrow t))$

(e) Notice this is *all* of the cells in the table where either C is s or S is t, and if we add up all those, we get $0.5 + 0.1 + 0.05 + 0.05 = 0.7$, which is correct.

(f) When we incorrectly just added $P(C \leftarrow s) + P(S \leftarrow t)$, that could be interpreted as adding up the two margins, one for $P(C \leftarrow s)$ and one for $P(S \leftarrow t)$. Note that each margin is the sum of its corresponding row or column, su adding up the margins is like adding up the individual cells, but we add in 0.5 twice! So mathematically, we could just subtract out one of those 0.5s and get the right answer. What cell did the 0.5 come from? The $P(C \leftarrow s \wedge S \leftarrow t)$ cell. So we conclude that $P(C \leftarrow s \vee S \leftarrow t) = P(C \leftarrow s) + P(S \leftarrow t) - P(C \leftarrow s \wedge S \leftarrow t)$. Some like to make this an axiom of probability. I think it is bizarre to take a formula that contains two mistakes that cancel each other out and make *that* an axiom, but that's just me.

11. What about *and*?

(a) Well, that's more complicated, and we need a new idea to talk about it.

(b) This new idea is *conditional* probability. It's like conjuction, but different

(c) We say $P(E_1 \leftarrow o_1 | E_2 \leftarrow o_2)$ is read as "the probability that E_1 is o_1, given that E_2 is o_2".

(d) This is read straight form the data table, just like all our other numbers.

(e) BUT, we read only from those entries where E2 is o2.

(f) For example, if we want to know $P(C \leftarrow w | S \leftarrow l)$ we look at the 6 cases where the Spare is loose, and count the number of times C is wobly (5). Thus $P(C \leftarrow w | S \leftarrow l)$ is (5/6), or 0.83333....

(g) So,

$$\begin{aligned} P(C \leftarrow w \wedge S \leftarrow l) &= \frac{N(w,l)}{n} \\ P(S \leftarrow l) &= \frac{N(l)}{n} \\ P(C \leftarrow w | S \leftarrow l) &= \frac{N(w,l)}{l} \end{aligned}$$

thus,

$$\begin{aligned} P(S \leftarrow l)P(C \leftarrow w | S \leftarrow l) &= \frac{N(l)}{n}\frac{N(w,l)}{l} \\ &= \frac{N(w,l)}{n} \\ &= P(C \leftarrow w \wedge S \leftarrow l) \end{aligned}$$

(h) Thus *and* does mean multiply, but not in the way you thought at all.

(i) Returning to the example, we can find $P(C \leftarrow w \wedge S \leftarrow l) = P(S \leftarrow l)P(C \leftarrow w | S \leftarrow l)$ $0.3 \times 0.833333 = 0.25$ giving us exactly what was in the table.

(j) The reverse is also true: $P(CF \leftarrow wobly \wedge S \leftarrow loose) = P(CF \leftarrow wobly)P(S \leftarrow loose | CF \leftarrow wobly)$. This is another example of an "axiom" that makes sense if you look at the data.

(k) The way I like to think of conditional probability $P(X \leftarrow a | Y \leftarrow b)$ is that if we consider just the elements in the table where Y is b (or if we pretend that the only data we have is where Y=b) then what is the probability of X being a?

12. Conditional probability is interesting because it tells us the relationship between events and outcomes.

(a) Sometimes things are highly related. The probability that a student will have a programming project in a semester is pretty low, about 0.1. But if we know the student is a CS major $P(Assignment \leftarrow programming | Major \leftarrow cs)$ is pretty high, 0.999. There is a strong relationship between programming and major.

(b) Sometimes things are not related at all. The probability that Joe Student will pass Calculus II is about 0.98. The probability that Joe Student will pass Calculus II given that in rained in St. Petersburg on September 5, 1752 is going to be about the same, no?

(c) That's an interesting property, that knowing rain does not change the probability of passing. In terms of the numbers, this means $P(Grade \leftarrow passing | Weather \leftarrow rain) = P(Grade \leftarrow passing)$.

(d) This says that these events are independent of each other. There is no relationship between these two events. We will indicate that 2 events A and B are independent as: $A \perp B$

(e) There is an additional side effect of independence. We know:

$$P(Grade \leftarrow passing \wedge Weather \leftarrow rain) = P(Weather \leftarrow rain)P(Grade \leftarrow passing | Weather \leftarrow rain)$$

we know

$$P(Grade \leftarrow passing | Weather \leftarrow rain) = P(Grade \leftarrow passing), \text{ therefore}$$

$$P(Grade \leftarrow passing \wedge Weather \leftarrow rain) = P(Weather \leftarrow rain)P(Grade \leftarrow passing)$$

(f) When two events are entirely unrelated, conjunction is just multiplication. But ONLY THEN.

13. Another handy rule:

(a) $\mathbf{P}((A \wedge B) \vee (A \wedge \neg B)) = ?$

$$\begin{aligned}
\mathbf{P}((A \wedge B) \vee (A \wedge \neg B)) &= \mathbf{P}(A \wedge B) + \mathbf{P}(A \wedge \neg B) - \mathbf{P}((A \wedge B) \wedge (A \wedge \neg B)) \\
&= \mathbf{P}(A \wedge B) + \mathbf{P}(A \wedge \neg B) \\
&\text{it is also equal to:} \\
&= \mathbf{P}((A \vee A) \wedge (A \vee \neg B) \wedge (A \vee B) \wedge (B \vee \neg B) \\
&= \mathbf{P}(A \wedge (A \vee \neg B) \wedge (A \vee B)) \\
&\text{And general resolution tells the 2nd and third clause are just } A \\
&= \mathbf{P}(A) \\
&\text{setting the two equal} \\
\mathbf{P}(A) &= \mathbf{P}(A \wedge B) + \mathbf{P}(A \wedge \neg B)
\end{aligned}$$

14. really, that is just a restatement of marginalization: $P(A) = \sum_o P(A \wedge B \leftarrow o)$

15. Bayes Rule

(a) We're going to start using the distribution notation more, so pay attention. Instead of $P(\mathit{Grade} \leftarrow \mathit{passing})$ we'll say $\mathbf{P}(G)$ meaning the distribution across all the outcomes of G. And remember that $\mathbf{P}(G \wedge W)$ (or just $\mathbf{P}(GW)$) is the joint table across the combined outcomes of of the two events. And when we say $\mathbf{P}(G|W)$ we also mean a table: $P(\mathit{Grade} \leftarrow \mathit{passing}|\,\mathit{Weather} \leftarrow \mathit{rain})$, $P(\mathit{Grade} \leftarrow \mathit{passing}|\,\mathit{Weather} \leftarrow \mathit{sunny})$,$P(\mathit{Grade} \leftarrow \mathit{failing}|\,\mathit{Weather} \leftarrow \mathit{rain})$, and $P(\mathit{Grade\ is\ failing}|\,\mathit{Weather} \leftarrow \mathit{sunny})$

(b) Now, we know that, for any two events, $\mathbf{P}(GW) = \mathbf{P}(G)\mathbf{P}(W|G)$. We also know that $\mathbf{P}(GW) = \mathbf{P}(W)\mathbf{P}(G|W)$. We can set them equal to each other and get, $\mathbf{P}(G)\mathbf{P}(W|G) = \mathbf{P}(W)\mathbf{P}(G|W)$. Solve for $\mathbf{P}(W|G)$ to get:

$$\mathbf{P}(W|G) = \frac{\mathbf{P}(W)\mathbf{P}(G|W)}{\mathbf{P}(G)}$$

(c) This is known as Bayes' Rule. It is useful because we can discover the conditional probability of something by using the conditional probability going the other direction.

(d) Take this simple problem. You are a prosecutor who wants to know whether to charge someone with a crime. You know this person's fingerprints are at the crime scene.

i. What we need to figure out is what is the probability that the person is guilty given that his prints were at the crime scene: $P(\mathit{Guilty} \leftarrow \mathit{true}|\mathit{Fingerprints} \leftarrow \mathit{true})$ or just $P(G \leftarrow t|F \leftarrow t)$

ii. We know from Bayes' Rule that

$$P(G \leftarrow t|F \leftarrow t) = \frac{P(F \leftarrow t|G \leftarrow t)P(G \leftarrow t)}{P(F \leftarrow t)}$$

iii. If we assume that one person in the town of 100,000 comitted the crime, then $P(G \leftarrow t) = 0.00001$

iv. Our forensics experts tell us that if someone commits a crime, they leave behind fingerprints 99% of the time, so $P(F \leftarrow t | G \leftarrow t) = 0.99$

v. If there are typically 3 person's fingerprints at any particular locale, we estimate that the probability a particular person's fingerprints being at the crime scene as 3 in 100,000, or $P(F \leftarrow t) = 0.00003$

vi. So we can plug in our values:

$$\begin{aligned} P(G \leftarrow t | F \leftarrow t) &= \frac{P(F \leftarrow t | G \leftarrow t) P(G \leftarrow t)}{P(F \leftarrow t)} \\ &= \frac{0.99 \times 0.00001}{0.00003} \\ &= 0.33 \end{aligned}$$

vii. Perhaps not enough to arrest him, but definitely someone we want to talk to, and gather more evidence on (Combining evidence is something we'll talk about below).

viii. Later, the wise ass detective points out that the suspect lives in the crime scene. This changes our belief in the probability that the fingerprints would be found there, to $P(F \leftarrow t) = 0.99$.

ix. Replugging in, we get:

$$\begin{aligned} P(G \leftarrow t | F \leftarrow t) &= \frac{P(F \leftarrow t | G \leftarrow t) P(G \leftarrow t)}{P(F \leftarrow t)} \\ &= \frac{0.99 \times 0.00001}{0.99} \\ &= 0.00001 \end{aligned}$$

x. If the suspect lives there, there's no reason to be surprised that the fingerprints are there, and thus they fingerprints shouldn't make us believe guilt. We have to let him go.

(e) In that example, there was something that should be surprising, but you may not have noticed

i. We didn't know know one conditional probability $P(G \leftarrow t | F \leftarrow t)$, but we did know another $P(F \leftarrow t | G \leftarrow t)$.

ii. This is surprising because I said earlier that these numbers were just read from the data table, and it should be just as easy to read one conditional probability from the table as the other.

iii. But it turns out that in the real world it is often hard to get good data. Instead certain kinds of data is easier to gather than others. In this case, most of our data on criminals comes from cases we've solved, where we know who the guilty person is and whether or not his fingerprints were there at the crime scene. We can even generate our own data by hiring professionals to stage crimes and see if there were fingerprints left behind giving us $\mathbf{P}(F|G)$ directly.

iv. Another reason we may want to reverse the conditionality is because one direction is less stable than another.

A. Take a medical diagnosis task. We want to know the the probability of having a disease given that we have a symptom x: $P(D \leftarrow t|S \leftarrow x)$. We know that this is:

$$P(D \leftarrow t|S \leftarrow x) = \frac{P(S \leftarrow x|D \leftarrow t)P(D \leftarrow t)}{P(S \leftarrow x)}$$

B. A well informed doctor might know $P(D \leftarrow t|S \leftarrow x)$ directly, but what happens if there is an epidemic? $P(D \leftarrow t|S \leftarrow x)$ will change but we don't know how much.

C. A Bayesian doctor will be able to adjust $P(D \leftarrow t)$ based on information on the epidemic. None of the other numbers need to be adjusted, so the Bayesian doctor can adapt.

v. As both these examples show, it is often easier to find and keep track of *causal* conditional probabilities than it is for *evidential* conditional probabilities.

(f) Another example. You think you might have the horrible disease *Severnitis*. You know that Severnitis is very rare- the probability that someone has it is 1 in 10,000 (.0001). There is a test for it that is reasonably accurate, 99%. You go get the test, and it comes back positive. You think "oh no! I'm 99% likely to have the disease!" Is that correct?

i. Our events are the test T, and health H. T can take 2 outcomes, positive (p) and negative (n). H can take 2 outcomes, diseased (d) and well (w). We want to know $P(H \leftarrow d|\, T \leftarrow p)$

ii. We know that the disease is rare $P(H \leftarrow d) = 0.0001$. Thus $P(H \leftarrow w) = 0.9999$

iii. What does 99% accurate mean? Well typically, it means, if you have the disease, it says positive $P(T \leftarrow p|H \leftarrow d) = 0.99$ or, if you don't have the disease, it says negative: $P(T \leftarrow n|H \leftarrow w) = 0.99$

iv. Therefore $P(T \leftarrow n|H \leftarrow d) = 0.01$ and $P(T \leftarrow p|H \leftarrow w) = 0.01$

v. We start with Bayes' Rule:

$$P(H \leftarrow d|\, T \leftarrow p) = \frac{P(T \leftarrow p|H \leftarrow d)P(H \leftarrow d)}{P(T \leftarrow p)}$$

vi. We know the values of the numerator, but not the denomenator, but we can find them with some clever math:

$$P(H \leftarrow d|\, T \leftarrow p) = \frac{P(T \leftarrow p|H \leftarrow d)P(H \leftarrow d)}{P(T \leftarrow p)}$$
$$P(H \leftarrow w|\, T \leftarrow p) = \frac{P(T \leftarrow p|H \leftarrow w)P(H \leftarrow w)}{P(T \leftarrow p)}$$

and since,

$$P(H \leftarrow d|\, T \leftarrow p) + P(H \leftarrow w|\, T \leftarrow p) = 1$$
$$\frac{P(T \leftarrow p|H \leftarrow d)P(H \leftarrow d)}{P(T \leftarrow p)} + \frac{P(T \leftarrow p|H \leftarrow w)P(H \leftarrow w)}{P(T \leftarrow p)} = 1$$
$$P(T \leftarrow p|H \leftarrow d)P(H \leftarrow d) + P(T \leftarrow p|H \leftarrow w)P(H \leftarrow w) = P(T \leftarrow p)$$

vii. We can substitute this back into bayes rule above, to get:

$$P(H \leftarrow d|\, T \leftarrow p) = \frac{P(T \leftarrow p|H \leftarrow d)P(H \leftarrow d)}{P(T \leftarrow p|H \leftarrow d)P(H \leftarrow d) + P(T \leftarrow p|H \leftarrow w)P(H \leftarrow w)}$$

and we know all the values:

$$\begin{aligned} P(H \leftarrow d | T \leftarrow p) &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \\ &= 0.0098 \end{aligned}$$

viii. This is why doctors are hesitant to order expensive tests if its unlikely you have the disease. Even though the test is accurate, rare diseases are so rare that the very rarity dominates the accurracy of the test.

(g) In the Severnitis example, we generated a new denominator to get around not knowing some numbers. The process of generating that denominator is called normalization, which seems like a funny thing to call it, but there is a good reason.

i. Let's look again at calculating the value of $P(H \leftarrow w | T \leftarrow p)$

$$P(H \leftarrow w | T \leftarrow p) = \frac{P(T \leftarrow p | H \leftarrow w) P(H \leftarrow w)}{P(T \leftarrow p)}$$

ii. Note that it has the same denominator as $P(H \leftarrow d | T \leftarrow p)$, namely $P(T \leftarrow p)$. Let α be $1/P(T \leftarrow p)$. Then,

$$\begin{aligned} P(H \leftarrow w | T \leftarrow p) &= \alpha P(T \leftarrow p | H \leftarrow w) P(H \leftarrow w) \\ P(H \leftarrow d | T \leftarrow p) &= \alpha P(T \leftarrow p | H \leftarrow d) P(H \leftarrow d) \end{aligned}$$

Since the sum of these two must be one, then

$$\alpha \left[P(T \leftarrow p | H \leftarrow w) P(H \leftarrow w) + P(T \leftarrow p | H \leftarrow d) P(H \leftarrow d) \right] = 1$$

So α is sometimes thought of as the "constant needed to make the distribution of $\mathbf{P}(T|H)P(H)$ add up to 1." In math, the processing of making stuff add up to 1 is called normalization.

iii. Some authors skip over how to do normalization entirely, jumping stragit from bayes rule to using the α with the explanation that "α is just the normalization constant." I think they do this to make you feel stupid- don't stand for it!

16. Multiple sources of evidence

(a) In real world problems of conditional probability, there's often a lot of evidence, not just one piece.

(b) This means there are more variables we need to worry about. Take our crime investigation:

i. We get additional evidence that a new suspect not only has fingerprints at the scene, but was found with the stolen items. Now we want to know the probability that he is guilty, given that his fingerprints were left at the scene *and* he had the purloined loot: $P(G \leftarrow t | F \leftarrow t \wedge L \leftarrow t)$. We can still apply Bayes Rule to this:

$$P(G \leftarrow t | F \leftarrow t \wedge L \leftarrow t) = \frac{P(F \leftarrow t \wedge L \leftarrow t | G \leftarrow t) P(G \leftarrow t)}{P(F \leftarrow t \wedge L \leftarrow t)}$$

ii. In general:

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

iii. That's nice, but how likely are we to know $\mathbf{P}(BCD...|A)$? Pretty unlikely, actually. Think of our court case where we have 100 peices of evidence. Its almost certain that we we've never seen a case before with exactly that evedience, let alone enough cases to have meaningful data.

iv. And what about the table for $\mathbf{P}(E_1E_2E_3...E_{100}|G)$? (where each E_n is a separate piece of evidence) Assuming boolean events, that table would need to have 2^{100} entries, which is bigger than the memory of any computer.

v. So as the number of variables grows, this technique gets less useful.

17. Conditional independence - In order to deal with multiple sources of evidence, we'll need the idea of conditional independence

(a) Recall that if 2 events are independent, then their conjunction is just the product of the individual probabilities: if $A \perp B$ then $\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$

(b) That means that since A and B are unrelated, knowing something about one tells us nothing about the other.

(c) Sometimes two events are independent given some other event. $A \perp B|C$, meaning that if we know the value of C, then knowing something about A tells us nothing about B. (A note on notation: the $\perp$ symbol has higher precedence than the | symbol, so $A \perp B|C$ is $(A \perp B)|C$ *not* $A \perp (B|C)$. The second one makes no sense since only events can be independent of each other, and $B|C$ is not an event.

(d) How could that happen? well, imagine we have three events: whether the suspect touched the doorknob, whether the subject left fingerprints on the doorknob, and whether the suspect was at the crime scene. Obviously, knowledge about fingerprints affects our belief about whether the suspect was there, so they are not independent. But once we know for sure that the suspect touched the doorknob, then saying that there were no fingerprints on the doorknob does not affect our belief that the suspect was there, since we already know that he touched the doorknob. Thus we say that $F \perp P|T$ (F:fingerprints, P:present, T: touched).

(e) What does this mean? well, if two things are independent given a third, then their conjuction is the product of the two separate conditional probabilities: if $A \perp B|C$ then $\mathbf{P}(AB|C) =$

$\mathbf{P}(A|C)\mathbf{P}(B|C)$. This comes from:

$$\begin{aligned}
\mathbf{P}(ABC) &= \mathbf{P}(AB|C)\mathbf{P}(C) \\
\mathbf{P}(ABC) &= \mathbf{P}(B|AC)\mathbf{P}(AC) \\
&= \mathbf{P}(B|AC)\mathbf{P}(A|C)\mathbf{P}(C) \\
&\text{therefore} \\
\mathbf{P}(AB|C)\mathbf{P}(C) &= \mathbf{P}(B|AC)\mathbf{P}(A|C)\mathbf{P}(C) \\
\mathbf{P}(AB|C) &= \mathbf{P}(B|AC)\mathbf{P}(A|C)
\end{aligned}$$

But since once we know C, knowing A tells us nothing new about C, so

$$\begin{aligned}
\mathbf{P}(B|AC) &= \mathbf{P}(B|C) \\
\mathbf{P}(AB|C) &= \mathbf{P}(B|C)\mathbf{P}(A|C)
\end{aligned}$$

(f) This independence simplifies a 3 dimensional table into two 2 dimensional tables. The big win comes when you simplify a 100 dimensional table into ninety-nine 2 dimensional tables.

18. Naive Bayes

(a) If we know about some conditional independence, we can encode that into the equations. This is done in Bayesean Networks, which we'll cover later.

(b) When someone applies Naive Bayes to a problem, what they do is *assume* conditional independence of *all* the events: $\mathbf{P}(ABC...|Z) = \mathbf{P}(A|Z)\mathbf{P}(B|Z)\mathbf{P}(C|Z)...$

(c) This can then be plugged into Bayes Rule:

$$\begin{aligned}
\mathbf{P}(A|BCD...) &= \frac{\mathbf{P}(BCD...|A)\mathbf{P}(A)}{\mathbf{P}(BCD...)} \\
&= \frac{\mathbf{P}(B|A)\mathbf{P}(C|A)\mathbf{P}(D|A)...\mathbf{P}(A)}{\mathbf{P}(BCD...)} \\
&= \alpha\mathbf{P}(B|A)\mathbf{P}(C|A)\mathbf{P}(D|A)...\mathbf{P}(A)
\end{aligned}$$

(d) And thus we've managed to reduce a high demensional table into several low demensional table. In terms of the crime scene evidence above, the 2^{100} element table would be reduced to one hundred 4 element tables. The difference between 2^{100} and 400 is rather large.

(e) But if the naive bayes assumption isn't true, and it almost never is, then the naive bayes algorithm can't be correct, can it?

(f) Well, no, but if we're using this as a classifier, its not important that the probability is correct, just that the correct class have the highest probability. It turns out that in practice for many problems, this is the case.