Lecture 10 (Part 1): Conditional probability III: conditional independence, the general chain rule, the birthday paradox

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 19 of the Math for CS book.

One of the things that we will be using in this lecture is the idea of conditioning on more than one events. Say we have three events A, B and C in a sample space Ω equipped with a probability distribution P. Then by $P(A|B, C)$ we basically mean $P(A|B \cap C)$, i.e. the probability of A happening given that *both* B and C have happened. It's not hard to verify the following identity (convince yourself!):

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

(You can verify it by plugging in the expressions for each of the terms on the LHS and RHS).

Obviously, we can condition on more than 2 events, for example given $A_1, \ldots, A_n$, we can talk about $P(A_i|A_1, \ldots, A_{i-1})$, i.e. the probability of A_i happening *given* that all of $A_1, \ldots, A_{i-1}$ have taken place.

1 Conditional independence

Definition 1 (Conditional independence)**.** *Given three events A, B and C in a sample space Ω with probability function P defined on it, A and B are said to be conditionally independent with respect to C if*

$$P(A \cap B|C) = P(A|C)P(B|C).$$

What conditional independence means is that once we are given the information that C has happened (and our sample space now "shrinks down" to C), A and B behave like independent events.

We will now see that independence and conditional independence are two different things which typically have nothing to do with each other.

Question . Consider the experiment of tossing a coin twice. Let A be the event of observing heads in the first toss and B be the event of observing tails in the second toss. If C is the event that an even number of heads are observed, prove that A and B are independent but not conditionally independent with respect to C.

Proof. It is clear that A and B are independent events (why?). Let's show that $P(A \cap B|C) \neq P(A|C)P(B|C)$ to show that A and B are not conditionally independent with respect to C. Note that the event C corresponds to the outcomes $\{HH, TT\}$, A corresponds to $\{HT, HH\}$, and B corresponds to $\{HT, TT\}$. From this we can see that $P(A|C) = P(B|C) = \frac{1}{2}$ (convince yourself!), however $P(A \cap B|C)$ is zero. □

This means that independence does not imply conditional independence. Let's see an example where we see two events that are not independent but are conditionally independent with respect to another event.

Question . Consider the following experiment. We have a box with two coins, one which is fair, and another which has two heads (we shall call it a two-headed coin). We first randomly pick one of the two coins from the box, and then toss the coin we twice. Let A be the event of seeing a heads in the first toss, B be the event of seeing a heads in the second toss, and let C be the event that the fair coin is picked in step one. Show that A and B are not independent but are conditionally independent with respect to C.

Proof. Using the tree method or Bayes theorem (left as an exercise), you can show that $P(A) = 3/4$, $P(B) = 3/4$, but $P(A \cap B) = 5/8$, and hence A and B are not independent.

Now let us condition on C, i.e. assume that C has happened and a fair coin was picked in step one. Then $P(A|C) = 1/2$, $P(B|C) = 1/2$, and $P(A \cap B|C) = 1/4$ (why?), and so A and B are conditionally independent with respect to C. □

2 General chain rule

We saw the chain rule for independent events: if $A_1, \ldots, A_n$ are mutually independent events then $P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n)$. But what can we say if $A_1, \ldots, A_n$ are not mutually independent?

One way to understand this case is the following. We are interested in the probability that all of $A_1, \ldots, A_n$ happen. Let's try to compute it step-by-step or in stages, such that

- First A_1 happens with probability $P(A_1)$.
- Then *given* that A_1 has taken place, A_2 happens with probability $P(A_2|A_1)$. So the probability of A_1 and A_2 happening in this order is $P(A_1)P(A_2|A_1)$ (This can also be proved using the definition of conditional probability).
- Next, given that A_1 and A_2 have happened, A_3 happens with probability $P(A_3|A_1, A_2)$. So the probability of A_1, A_2, and A_3 happening in that order is $P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)$.

$$\vdots$$

$$\vdots$$

- And finally, A_n happens given that $A_1, \ldots, A_{n-1}$ have happened, and the probability of this is $P(A_n|A_1, \ldots, A_{n-1})$, and so the probability of $A_1, \ldots, A_n$ happening, one-by-one, in that order is

  $$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)P(A_4|A_1, A_2, A_3)\ldots P(A_n|A_1, \ldots, A_{n-1}).$$

  This is called the (general) chain rule.

It's not that hard to formally prove the general chain rule using induction and the definition of conditional probability. We will skip the proof here and move straight to an important example that employs the chain rule.

3 Birthday paradox

Let us make some simplifying assumptions before state the problems:

1. There are always 365 days in an year, i.e. assume that there are no leap years. Furthermore, let's number the days as 1 to 365, where day 1 is January $1st$ and day 365 is December 31st.
2. Everyone is/was born in a non-leap year.
3. If we pick a random person, then they are equally likely to have been born on any of the 365 days (let's forget about the year of birth).

Question . Assume the above conditions. Now suppose we choose n random persons to form a group. What is the smallest value of n for which we will *definitely* have at least two people in the group who share a birthday (disregarding the year of course)? What is the smallest value of n so that with *probability* 99% we have at least two people in the group who share a birthday?

Proof. If we wanted to be a 100% sure that two people in the grop have the same birthday, we would have to have a group of at least 366 people (why?), and so $n = 366$ is the smallest value we can go for.

Here is a paradox: if you are willing to go from 100% sure to 99% sure, a much much smaller value of n can do the job, i.e. for a significantly smaller value of n (as compared to 366), we can be 99% sure that two people in the randomly formed group have the same birthday.

Let A denote the event that at least two people in the group have the same birthday. We will find $P(A^c)$. Let us assume that we are choosing the n people for the group, one at a time from the population, at random. Let A_1 be the event that the first person has a birthday on one of the 365 days (obviously, $P(A) = 1$ based on our assumptions). Let A_2 be the event that the first two persons that are chosen have distinct birthdays, let A_3 be the event that the first three people that are chosen have distinct birthdays, $\ldots$, A_i be the event that the first i people that are chosen have all distinct birthadys, $\ldots$, and let A_n be the event that the all n people that were chosen have distinct birthdays. Then

$$A^c = A_1 \cap A_2 \cap \ldots \cap A_n.$$

So basically, we have to find $P(A_1 \cap \ldots \cap A_n)$. Are $A_1, \ldots, A_n$ independent events? They are not (why?), and so we will use the general chain rule to find this probability. Recall that the general chain rule says that

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\ldots P(A_n|A_1, A_2, \ldots, A_{n-1}).$$

Let's try to compute each of the probabilities involved in the above expression. What's $P(A_1)$? Well that's just 1 and we discussed this above. What about $P(A_2|A_1)$? We want that when we choose the second person, their birthay should be distinct from that of the first person, and the probability of that is $1 - \frac{1}{365}$ (why?). Next, given that the first two people have distinct birthdays, what's the probablity that the third person has a birthday distinct from the first two? That's $1 - \frac{2}{365}$, and so $P(A_3|A_1, A_2) = 1 - \frac{2}{365}$. Following the same idea, we can conclude that the probability that the i$^{\text{th}}$ person has a birthday distinct from the first $i-1$ people given that the first $i-1$ people have all distinct birthdays is $1 - \frac{(i-1)}{365}$, i.e. $P(A_i|A_1, \ldots, A_{i-1}) = \frac{365-(i-1)}{365}$. Thus,

$$P(A_1 \cap \ldots \cap A_n) = 1(1 - \frac{1}{365})(1 - \frac{2}{365}) \ldots (1 - \frac{n-1}{365}).$$

We will now use a basic fact from calculus:

Fact 2. *For all $x \geq 0$,*

$$(1-x) \leq e^{-x}.$$

Let's apply this inequality to each of the terms, i.e. use the fact that for every $2 \leq i \leq n$,

$$1 - \frac{(i-1)}{365} \leq e^{-\frac{(i-1)}{365}}.$$

So we get that

$$P(A_1 \cap \ldots \cap A_n) \leq e^{-\frac{1}{365}} e^{-\frac{2}{365}} \ldots e^{-\frac{(n-1)}{365}} = e^{-\frac{(1+2+3+\ldots+(n-1))}{365}} = e^{-\frac{n(n-1)}{730}},$$

where the last equality uses the fact that $1 + 2 + \ldots + (n-1) = \frac{n(n-1)}{2}$.

All this means that the probability that all n people have distinct birthdays is

$$P(A^c) = P(A_1 \cap \ldots \cap A_n) \leq e^{-\frac{n(n-1)}{730}}.$$

We want the probability that at least two people have the same birthday to be ≤ 0.99, i.e. $P(A) \geq 0.99$, and so we want that $P(A^c) \leq 0.01$. This basically means we want to choose a n that's large enough so that $P(A^c) \leq 0.01$. To find such an n, recall that

$$P(A^c) \leq e^{-\frac{n(n-1)}{730}}.$$

So if we can choose n so that the RHS becomes less than 0.01, we should be good. So we want

$$e^{-\frac{n(n-1)}{730}} \leq 0.01.$$

Inverting both sides, and flipping the direction of the inequality, we get

$$e^{\frac{n(n-1)}{730}} \geq 100$$

We can now take natural logarithm on both sides, to get

$$\frac{n(n-1)}{730} \geq \ln(100)$$

or

$$n(n-1) \geq 730 \times \ln(100).$$

If we can ensure that $(n-1)(n-1)$ is larger than the RHS than surely $n(n-1)$ is also larger than the right hand side, since $n(n-1) \geq (n-1)(n-1)$. So let's try to find an n such that

$$(n-1)(n-1) = (n-1)^2 \geq \ln(100) \times 730.$$

or

$$n \geq \sqrt{730 \times \ln(100)} + 1.$$

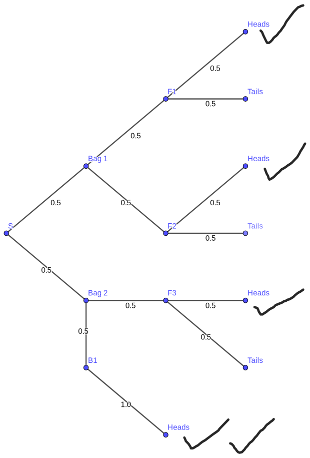
The RHS works out to be about 58. This means that if we choose an n that's at least 58, we can ensure that $P(A^c) \leq 0.01$. Let's choose $n = 60$. This is incredible!! If we willing to settle for 99% sureity instead of being a 100% sure, we need only 60 randomly chosen people to guarantee that two people have the same birthday! This is a sixth of 365, the number of people needed to ensure that you always have two people with the same birthday. □

4 The general chain rule and the tree method

We saw informally in the previous lecture that the "engine" behind the tree method is basically the law of total probability. Well, I lied a bit. While the version of the law of total probability we saw in the last lecture does explain why the tree method works for cases when the *depth* of the tree is two, it doesn't fully explain the case when the depth of the tree is more than two. By depth here I mean the length of the longest root to leaf path. One needs the general chain rule to address the case when the depth is more than two. Let's consider an example from Lecture 8 to demonstrate this.

Question . I have two bags. Bag 1 contains two fair coins and Bag 2 contains a fair coin and a coin with two heads. I toss a fair coin. If the outcome is heads, I pick Bag 1, and if the outcome is tails I pick Bag 2. Then from the bag in picked I the previous step, I randomly pick a coin such that both coins in the bag are equally likely to be picked. Finally I toss the coin I picked from the bag. What is the probability that the final outcome is heads?

Proof. Here is the tree (F_1 and F_2 are the two fairs coin in Bag 1, and F_3 and B_1 are the fair and two-headed coins respectively in Bag 2):



You will notice that the tree has depth/height 3: the first node branches into two possibilities depending on which box is picked, then the next level of nodes branch into two possibilities each depending which coin is picked, and finally there is another branching based on what the coin toss results in.

Consider the leaf that corresponds to the event of picking the box with a fair coin and a two-headed coin, then picking the fair coin, and then seeing a heads (this is the third leaf from the top with a check mark). The way we calculate the probability of this event is by multiplying all the probabilities along the root to leaf path for this event. Why does the multiplication work? Let's try to understand that now:

Let A_1 be the event of picking the box with a fair coin and a two-headed coin, let A_2 be the event of picking the fair coin, and let A_3 be the event of seeing a heads in the final toss. The leaf we talked about in the previous paragraph basically corresponds to $A_1 \cap A_2 \cap A_3$, and as discussed above the way we find the $P(A_1 \cap A_2 \cap A_3)$ is by multiplying the root-to-leaf probabilities. To see why the multiplication works, note that the $P(A_1)$ is 0.5. Given that A_1 has happened, i.e. the box with a fair coin and a two-headed coin is picked, the probability of picking a fair coin is 0.5, i.e. $P(A_2|A_1) = 0.5$. Finally, given that Box 2 was picked *and* the fair coin was picked from it, it's clear that the probability of seeing a heads is 0.5, i.e. $P(A_3|A_1, A_2) = 0.5$. Using the general chain rule, we know that

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) = 0.5 \times 0.5 \times 0.5.$$

So the reason why we multiply is because we are implicitly using the general chain rule! Each of the branches is basically labelled with a conditional probability!

Now we proceed in a similar way: we multiply all the probabilities on the root-to-leaf path for every leaf with a check mark, and the reason we do this is the same as we discussed above: we are using the general chain rule! Once we compute all the needed probabilities, we add them up to get the total probabilty of seeing heads. The last step of adding these probabilities is similar to the law of total probability.

Basically, what we are doing is, to *partition* A_3, the probability of seeing a heads, into various disjoint cases: "Box 1 and fair coin and heads", "Box 1 and fair coin and tails", "Box 2 and fair coin and heads",..., etc. More formally,

$$P(A_3) = P(A_1 \cap A_2 \cap A_3) + P(A_1 \cap A_2^c \cap A_3) + P(A_1^c \cap A_2 \cap A_3) + P(A_1^c \cap A_2^c \cap A_3).$$

(Convince yourself that all these sets form a partition of A_3). The four probabilities in the above expression correspond to the four leaves with a check mark. This explains why we add them up. Each of these probabilities can then be computed using the general chain rule as discussed above. □