Discrete Math Lecture 5 (after midterm) and 6

1 Probability

We spent some time learning to count things. This is going to be very handy in today's class. We can count all possible ways a certain outcome can occur, which will allow us to reason about how likely it is that outcome occurs. How? Well, intuitively at least, if you have a bunch of equally likely outcomes, the chance a given kind of outcome occurs, is just

$$\frac{\text{number of outcomes of that kind}}{\text{total number of outcomes}}$$

This is where probability theory comes into play. In particular, we are interested in discrete probability, which deals with countably many outcomes and is attacked using many of the tools that we've been using. Another branch of probability theory (continuous probability) deals with uncountably many outcomes and is approached using methods from calculus and analysis. What that means for us is that we can get away by just summing things, while in continuous probability you will be doing lots of integration instead of summation.

Probability and randomness play an important role in computer science. Probabilistic methods are an important part of the analysis of algorithms and algorithm design. There are some problems for which we may not know of an efficient deterministic algorithm, but we can come up with a provably efficient randomized algorithm. To tackle these problems, we need a basic understanding of discrete probability. Often, if you are willing to accept an approximate answer, very simple randomized algorithms do the job. For example, the max cut problem asks you to divide the vertices of a graph into two sets so as to maximize the number of edges between the sets. However just picking two equal size sets of vertices at random allows one to get a partition such that, on the average, at least half of the graph edges are between the sets, so even the optimal solution cannot be more than twice as good!

Another important example is the power of two choices in load balancing. In essence this result says that if I have a bunch of servers to handle requests and I want to balance the load on them, then I can do much better than a random allocation, if I am allowed to pick two servers at random and send my request to the less loaded one. This counter-intuitive result, which can be proved using tools in probability theory, thus tells us that even a little bit of selection (choosing the better out of two random options) can lead to much better outcomes than a completely random load allocation.

The trouble with probability is that we, as humans, don't have good intuitions about probability. This is all the more reason to develop a formal system for reasoning about probability. Interestingly enough, a lot of the basic terms and formalism that we will study, were only developed in the early twentieth centuary, though people have been dealing with chance (for example games of chance) since much much earlier.

Let's consider rolling a standard 6 sided die. There are 6 possible numbers we can observe. These are the outcomes of the roll. We can pick any subset of the outcomes as an event we want to consider. For example, I might care about whether the roll is even or odd. Intuitively, each of these is equally likely. We can call any subset of the possible rolls as an event which we want to consider. This motivates the following definitions:

Definition 1 (Probability Space)**.** *A probability space* $(\Omega, \Pr)$ *is a non-empty finite set* Ω *and a function* $\Pr : \Omega \to \mathbb{R}^+$ *such that*

1. *For every* $w \in \Omega$ $\Pr(w) > 0$ *and*

2. $\sum_{w \in \Omega} \Pr(w) = 1$

An event is just a subset of the set Ω and the probability of an event $A \subseteq \Omega$ is the sum of the probabilities of the elements in A, i.e., we can define

$$\Pr(A) = \sum_{w \in A} \Pr(w)$$

If you are paying close attention, you will notice that this Pr on the left in the above equation is not the same as the Pr on the right. After all we defined Pr for elements of Ω, not subsets of Ω, but we will use the same name for this new function that we have just defined on the subsets of Ω.

If we have a fair die then we can say $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $Pr(w)$ is the same for every element of Ω. Since the probabilities must add up to 1, we obtain $\Pr(w) = 1/6$ for each possible $w \in \Omega$. A probability distribution specifies the probability of each event. Notice that it is sufficient to specify $\Pr(w)$ for each element w of Ω. In general this sort of probability distribution, where $\Pr(w) = \frac{1}{|\Omega|}$ for every $w \in \Omega$ is called the uniform distribution.

What if we want to consider two consecutive rolls of the dice. Then the sample space would be the set of ordered pairs of numbers (i, j) where each of i and j could be between 1 and 6. These are just the elements of the set $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$. If we throw both dice together though, and the dice are indistinguishable. Then the order doesn't matter and the sample space would be the set of unordered pairs of numbers, each number being between 1 and 6.

What about coins? It turns out that flipping a bunch of coins is probably the most important sample space for theoretical computer science. Why? Let's try to do the same exercise as with dice. What if we want to represent a series of three coin flips? We can do the same as with the dice: we have $\{H, T\} \times \{H, T\} \times \{H, T\}$ for our sample space. Now, if we were to generalize this to n coin flips, we would have something like $\{H, T\}^n$. This space looks a lot more familiar if we rename $T = 0$ and $H = 1$, giving us the space $\{0, 1\}^n$. That's right - the sample space of n coin flips can be represented as the set of length n binary strings!

Recall that a standard deck of playing cards consists of four suits, with cards numbered from 2 through 10, and a Jack, Queen, King, and Ace. This gives us 52 cards in total (4×13). Suppose we deal a hand of 13 cards from a standard 52 card deck to four players: North, East, South, and West. What is the probability that North holds all the aces?

First, we need to think about what our sample space should be. Since we are concerned with hands, this will be the result of the deal: all of the possible ways to distribute 13 card hands to four people. Combinatorially, this is distributing 52 distinguishable objects into four distinguishable boxes with 13 objects each. Then the size of our sample space is

$$|\Omega| = \binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13}.$$

Now, we want to count the number of such hands where North holds all the aces. In this case, North holds four cards that are fixed already, so we need to count how the other 48 cards are distributed. Let's call this event A. This gives us

$$|A| = \binom{48}{9}\binom{39}{13}\binom{26}{13}\binom{13}{13}.$$

Then the probability of our event is just

$$\begin{aligned}\Pr(A) &= \frac{|A|}{|\Omega|} \\ &= \frac{\binom{48}{9}\binom{39}{13}\binom{26}{13}\binom{13}{13}}{\binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13}} \\ &= \frac{13 \cdot 12 \cdot 11 \cdot 10}{52 \cdot 51 \cdot 50 \cdot 49} \\ &\approx 0.00264\end{aligned}$$

Of course we may not always be dealing with uniform probabilities. For example if we have a die such that $\Pr(1) = 1/2$ and all other probabilities are equal. i.e., $1/10$, then our definition of the probability of an event lets us calculate that the probability of rolling an odd number to be

$$\begin{aligned}\Pr(\text{odd}) &= \Pr(\{1,3,5\}) \\ &= \Pr(1) + \Pr(3) + \Pr(5) \\ &= \frac{1}{2} + \frac{1}{10} + \frac{1}{10} \\ &= 7/10\end{aligned}$$

Notice that when calculating probabilities for events we are dealing with sets, so we can use our old rules for counting the number of things in unions of sets. Thus we have, by the inclusion exclusion principle that

$$\begin{aligned}\Pr(A \cup B) &= \frac{|A \cup B|}{|\Omega|} \\ &= \frac{|A| + |B| - |A \cap B|}{|\Omega|} \\ &= \frac{|A|}{|\Omega|} + \frac{|B|}{|\Omega|} - \frac{|A \cap B|}{|\Omega|} \\ &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &\leq \Pr(A) + \Pr(B)\end{aligned}$$

This can be generalised to the union of more than two events. The last step gives us a rather useful inequality, called the **union bound**. If you have a bunch of events and want to upper bound the probability of their union, then we can just use the sum of their probabilities as an upper bound. This works even if the event sets are not disjoint. We will later study the notion of events depending on one another. But notice that whether the events are independent or not, we can still use the union bound. If the events are disjoint, the probability of the union is exactly the sum of the individual probabilities.

For example, if I am running an application with two known critical bugs that cause the system to crash and the chance of one bug is 2% and the other is 5%, then I can say for sure, that the overall chance of a crash is no more than 7%. This works even if the bugs are related in any sort of complex way, for example there could be cases where both bugs simultaneously occur and there could be cases where one of the bugs can cause the other.

Let's see another example use of the formula we just derived. If I have a fair dice and I want to know what the chance of rolling a multiple of 2 or 3 is, I can calculate as follows:

$$\begin{aligned}
\Pr(\text{multiple of 2 or 3}) &= \Pr(\text{multiple of 2} \cup \text{multiple of 3}) \\
&= \Pr(\text{multiple of 2} \cup \text{multiple of 3}) \\
&= \Pr(\text{multiple of 2}) + \Pr(\text{multiple of 3}) - \Pr(\text{multiple of 2 and 3}) \\
&= \Pr(\{2,4,6\}) + \Pr(\{3,6\}) - \Pr(\{6\}) \\
&= \frac{3}{6} + \frac{2}{6} - \frac{1}{6} \\
&= \frac{4}{6} = \frac{2}{3}
\end{aligned}$$

And just as for sets we have the same notion of disjointness of events. A and B are disjoint events if they have an empty intersection.

For any set of events A_i, where i goes from 1 to k, we have

$$\Pr(\bigcup_{i=1}^{k} A_i) \leq \sum_{i}^{k} \Pr(A_i)$$

with equality when the events are pairwise disjoint.

2 Conditional Probability

Something that does not have an immediate analog from set theory is the notion of conditional probability. Suppose we ask: What is the probability that a fair dice rolls a 2 given that it rolled an even number. Usually there are 6 possibilities: $\{1,2,3,4,5,6\}$ and all are equally likely, so the chance of a 2 is 1/6. However we have been told that the roll is even so it could not be 1, 3 or 5. Thus, really, there are only 3 possibilities: 2, 4 and 6. Once again each possibility is still equally likely and there are 3 possibilities, so the probability of rolling a 2 is 1/3. More formally, we can define the conditional probability of event A, given the event B to be:

Definition 2. *(Conditional Probability) The probability of A given B is denoted by*

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

.

Notice that the numerator represents both A and B being true while we are dividing by the probability that B is true. Recall our intuitive understanding of probability. The chance a given kind of outcome occurs, is just

$$\frac{\text{number of outcomes of that kind}}{\text{total number of outcomes}}$$

Intuitively, we know B happens, so among the possible outcomes in the numerator we should count only those where B also happened. In the denominator also we should only consider cases where B happened.

We can think the definition of conditional probability as follows. $\Pr(A)$ is just the size of the set A divided by the size of the set of all possible outcomes. Let's say S is the set of all possible outcomes.

$$\Pr(A) = \frac{|A|}{|S|}$$

Suppose we know are told B happened. So we must be inside the set B. Really, we are saying then that we want to see how many ways is it possible to have both A and B hold compared to the number of ways it is possible for just B to happen. How many ways can we have both A and B hold? It is $|A \cap B$. What about just B? That can happen in $|B|$ ways. So we should have

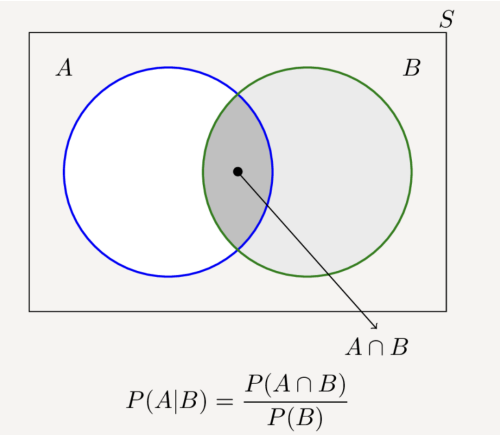
$$\Pr(A|B) = \frac{|A \cap B|}{|B|} = \frac{\frac{|A \cap B|}{|S|}}{\frac{|B|}{|S|}} = \frac{\Pr(A \cap B)}{\Pr(B)}$$

and indeed this matches the definition of $\Pr(A|B)$.

Another way of looking at this is if we restrict our attention to the event B, treat it like a sample space, and ask, what is the probability of A in this space? This is reflected in the fact that it turns out that

$$\Pr(A|B) = \frac{|A \cap B|}{|B|}$$

Representing this pictorially:



We can check that this definition matches our intuition in the dice example. Let A=2 was rolled and B = an even number was rolled. Then

$$\begin{aligned}
\Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\
&= \frac{|A \cap B|}{|B|} \\
&= \frac{|\{2\} \cap \{2,4,6\}|}{|\{2,4,6\}|} \\
&= \frac{\{2\}}{\{2,4,6\}} \\
&= \frac{1}{3}
\end{aligned}$$

Let's see some more examples

Suppose we roll two fair dice. First let's ask what is the chance that the sum of the dice is 7?

We already have studied some combinatorics which will be useful in solving this problem. We can call one die the first die and the other the second die. Then our sample space is the set of ordered pairs (i, j) where $i, j \in \{1, 2, 3, 4, 5, 6\}$. But how can the answer depend on whether we consider the dice distinguishable or not? After all if we roll two dice, then whether we roll them 1

by 1 or both together, the chance of the sum being 7 ought to be the same. And indeed this is true. But if we imagine rolling the dice together, and think of our sample space as the set of unordered pairs, then we do have to notice that every pair does not have the same probability! There are two ways to get $(5, 2)$; one die could be 2 and the other one 5, or it could be the other way around. But there is only one way to get $(6, 6)$, that is when both dice roll 6. So it will be convenient for us to assume that the sample space is the set of ordered pairs. This will make our calculations easier since every element in the sample space will have the same probability. Basically, we are saying $(2, 5)$ and $(5, 2)$ are different events, because we can imagine throwing the dice one after another.

Now let's count the total number of outcomes. We choose the roll of the first dice (6 possible ways) and then the second, leading to 6^2=36 possible outcomes. How many ways can we have the sum of the rolls be 7? Since we are treating the dice as distinguishable, we can call the number on dice 1 as x_1 and the number on dice 2 as x_2. Thus what we want to know is how many solutions are there to the equation

$$x_1 + x_2 = 7$$

Where x_1 and x_2 are each between 1 and 6. We can use what we learnt about combinatorics to count these solutions, and it turns out there are exactly 6 solutions. Thus, the probability of rolling a sum of 7 is $\frac{6}{36} = \frac{1}{6}$.

Notice that we could actually do this calculation, even while treating the dice as indistinguishable. Our sample space would now be the set of unordered pairs. We would just have to say that there are two kinds of pairs, ones where the numbers are distinct, and ones where the two numbers are the same. The first kind of pair has twice the probability as the second. So $\{2, 5\}$ has twice the probability as $\{6, 6\}$. A little counting shows us that there are 15 pairs of the first kind, and 6 of the second kind. If the second kind of pair appears with probability p, then the first kind has probability $2p$. This let's us calculate $6p + 15 \times 2p = 1$ so $p = 1/36$. Now the only way to get 7, since 7 is odd, involves pairs of the first kind How many ways are there: well it's just $\{1, 6\}$,$\{2, 5\}$ and $\{3, 4\}$ since we are proceeding as if teh order does not matter. So there are 3 different outcomes each resulting in a sum of 7, and each with probability 2/36, which yields the probability of rolling a 7 to be $3 \times \frac{2}{36} = \frac{6}{36} = \frac{1}{6}$. As expected, this is the same answer!

We should take care not to confuse distinguishable and indistinguishable objects while counting things. This is something you have already seen while doing problems in combinatorics! The answers to counting problems heavily depends on whether the objects are distinguishable or not.

Now let's ask a different question: what is the chance the sum of the two fair dice is 7, given that one of the dice is a 4?

We might as well use our initial approach, where we treated the dice as distinguishable. So there are only two possibilities, either the first dice is a 4, or the second one is. The number of ways this can happen, by the inclusion/exclusion principle is |dice 1 is a 4| + |dice 2 is a 4| − |both are 4| = $6 + 6 - 1$.So there are 11 ways. Now how many ways are there to get a 7? Only $(3, 4)$ and $(4, 3)$. Because each of these possibilities has the same probability of occurring, so we get the probability to be 2/11. Alternatively, we could use the definition of conditional probability to calculate. Let A = sum is 7 and B = one of the dice is 4 (the other could be anything, including 4). Remember we have 36 equally likely outcomes in total. Then

$$\begin{aligned}\Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\ &= \frac{\frac{|A \cap B|}{36}}{\frac{|B|}{36}} \\ &= \frac{|A \cap B|}{|B|} \\ &= \frac{2}{11}\end{aligned}$$

Notice how it is really the same calculation.

3 The Law of Total Probability

Suppose I know $\Pr(A|B)$ and also $\Pr(A|B^c)$ (B^c is just the complement of B). Then putting them together should tell me what $\Pr(A)$ is. But we should not just take the sum of these probabilities, it also matters how likely B is. If B is very likely, then we should weight $\Pr(A|B)$ more, since B is likely. In fact

$$\Pr(A) = \Pr(A|B) \cdot \Pr(B) + \Pr(A|B^c) \cdot \Pr(B^c)$$

The law of total probability generalizes this idea. Suppose $H = H_1 \cup H \cup \ldots \cup H_k$ is a partition of the underlying set in the probability space, Ω. A partition is a division of a set into disjoint subsets that together cover the whole set, i.e., $H_i \cap H_j = \phi$ for each pair of distinct sets in the partition and $\cup H_i = \Omega$. Then

$$\Pr(A) = \sum_{i=1}^{k} \Pr(A|H_i) \Pr(H_i)$$

Proof.

$$\begin{aligned}\Pr(A) &= \Pr\left(A \cap \bigcup_{i=1}^{m} H_i\right) \\ &= \Pr\left(\bigcup_{i=1}^{m} (A \cap H_i)\right) \\ &= \sum_{i=1}^{m} \Pr\left(A \cap H_i\right) \qquad \text{since } H_i \text{ are pairwise disjoint} \\ &= \sum_{i=1}^{m} \Pr\left(A \mid H_i\right) \Pr\left(H_i\right)\end{aligned}$$

□

Let's see an example using the law of total probability. Suppose there are only two pizzeries that deliver pizzas in Hyde Park. Pizzeria A delivers on time 80% of the time while Pizzeria B is on time 60% of the time. Thankfully UChicago is twice as likely to order from Pizzeria A than B for it's events. What is the chance that at a given event at the university, the pizza will arrive on time?

We can consider A and B to be the events that the pizza is ordered from Pizzeria A and B respectively, while T is the event that the Pizza is on time. Then

$$\begin{aligned}\Pr(T) &= \Pr(T|A)\Pr(A) + \Pr(T|B)\Pr(B) \\ &= 0.8 \times \frac{2}{3} + 0.6 \times \frac{1}{3} \\ &= \frac{2}{3}\end{aligned}$$

So there is a 66% chance that the pizza is on time. Notice that this calculation really is just taking the weighted average of the probabilities in the two cases.

4 Bayes Theorem

Sometimes don't know what $\Pr(A|B)$ is but we do know $\Pr(B|A)$. Bayes Theorem is a very useful and very important tool in such situations.

The theorem says that, for any events A and B,

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

Let's see a proof of why this is true. It follows directly from the definition of conditional probability.

Proof.

$$\begin{aligned}\Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\ &= \frac{\Pr(A \cap B)}{\Pr(A)} \cdot \frac{\Pr(A)}{\Pr(B)} \\ &= \frac{\Pr(B \cap A)}{\Pr(A)} \cdot \frac{\Pr(A)}{\Pr(B)} \\ &= \Pr(B|A) \cdot \frac{\Pr(A)}{\Pr(B)}\end{aligned}$$

□

Let's now see an example. Suppose we have two bags, Bag 1 with 1 red and 99 blue balls and another bag, bag 2, with 99 red and 1 red ball. Now I choose one of the bags uniformly at random. So each bag has a half chance of being chosen. After choosing the bag, I draw a ball at random out of the bag. What is the chance that I draw a red ball? Well, we can use the law of total probability

$$\begin{aligned}\Pr(\text{Red ball drawn}) &= \Pr(\text{Red ball drawn}|\text{Bag 1})\Pr(\text{Bag 1}) + \Pr(\text{Red ball drawn}|\text{Bag 2})\Pr(\text{Bag 2}) \\ &= \frac{1}{100} \cdot \frac{1}{2} + \frac{99}{100} \cdot \frac{1}{2} \\ &= \frac{1}{2}\end{aligned}$$

Well that isn't a surprise, since we choose the balls with the same probability and overall, between the bags, there are an equal number of red and blue balls. But supposing I draw a ball and tell you it is red. Then what is the chance that it came from Bag 2? Without knowing the color of the ball, the chance that a ball is drawn from either Bag is half. But intuitively, given that it is

red, it should have been much more likely to have come from Bag 2. Bayes theorem tells us what the updated probability, after knowing that the ball is red, should be:

$$\begin{aligned}\Pr(\text{Bag 2 used}|\text{Red ball drawn}) &= \frac{\Pr(\text{Red ball drawn}|\text{Bag 2 used}) \cdot \Pr(\text{Bag 2 used})}{\Pr(\text{Red ball drawn})} \\ &= \frac{\frac{99}{100} \cdot \frac{1}{2}}{\frac{1}{2}} \\ &= \frac{99}{100}\end{aligned}$$

Interestingly, in this case, $\Pr(\text{Bag 2 used}|\text{Red ball drawn}) = \Pr(\text{Red ball drawn}|\text{Bag 2 used})$. But in general, this is certainly not the case.

Excercise 1. *Modify the above example, when there is a* $2/3$ *chance of drawing from bag 1 and only a* $1/3$ *chance of drawing from Bag 2 and redo this calculation.*

Notice we used the total probability calculation we just did for Pr(Red ball is drawn). Also notice that Bayes theorem allowed us to figure out Pr(Bag 2 was used|Red ball drawn) in terms of Pr(Red ball drawn|Bag 2 was used), which is much easier to calculate.

Finally, it is good to remark that we might often have to calculate $\Pr(B)$ using the law of total probability. So for example you might write:

$$\Pr(B) = \Pr(B|A) \cdot \Pr(A) + \Pr(B|\neg A) \cdot \Pr(\neg A)$$

Remember that A not happening is the event exactly represented by the complement of A. So we will write $\neg A = A^c$. Now we have

$$\begin{aligned}Pr(A|B) &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \\ &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|A^c)\Pr(A^c)}\end{aligned}$$

Notice that in fact one of the terms in the sum in the denominator is exactly the numerator. You will see this in the following problem.

Excercise 2. *New Yorkers have only a* 30% *chance of liking deep dish pizza while* 70% *of Chicagoans like deep dish pizza. In a conference you are told that* 60% *of participants are New Yorkers and* 40% *are Chicagoans. You meet someone at dinner and they love deep dish pizza. What is the chance that they are from Chicago?*

Bayesian thinking can be very useful in many situations, we will see one real world application today. The key idea is that you start out with some prior belief about the probability of an event A. $\Pr(A)$ is called the **prior probability**. You then see some **evidence** E and you update the probability of A given the fact that you observed E to obtain the **posterior probability** $\Pr(A|E)$.

Suppose you are a doctor.Let's say that the chance of a patient having a certain rare disease is only $1/100$. This is the probability you would assign, for the event that the patient has the disease, prior to any testing. You are a doctor and you perform a test for the disease, and unfortunately the test comes up positive. The test was performed in a bunch of cases and experiments show that out of 10000 patients who had the disease, the test was positive in 9000 of them , so you reasonably estimate that the chance of the test yielding a positive result, given that the patient had the disease is $9/10$. Can you use this to conclude that there is a 90% chance that your patient has the disease? What is the probability that the patient actually has the disease? The intuitive idea is that you

have a prior probability in mind before the evidence from the test, and you must update your idea about the chance of the patient having the test, to take into account the evidence of the test. Bayes theorem tells us how to do this Bayesian update of our prior notion of the probability. Let D be the event that the patient had the disease, and T be the event that the test was positive.

Well, Bayes theorem tells us that

$$\Pr(D|T) = \frac{\Pr(T|D)\Pr(D)}{\Pr(T)}$$

I know $\Pr(T|D) = 9/10 = 0.9$ and $\Pr(D) = 1/100 = 0.01$, but still this is not enough! What about $\Pr(T)$? Let's say we believe the patient we walked into our office was some random member of the public, we know nothing else special about them. If we assume our test yields a positive result on one out of 50 members of the public drawn at random, i.e., $\Pr(T) = 1/50 = 0.02$, then we obtain

$$\Pr(D|T) = \frac{0.9 \times 0.01}{0.02} = 0.45$$

So despite the fact that at first site it might seem that the test is "90% accurate" we only have a 45% chance that our patient has the disease! Also, our calculation should reveal that to know what the chance of a patient having the disease, we can't just rely on the information about how often the test successfully detects the disease. That information is useless without

1. Information about how often the test returns a positive when used on a generic individual of the population we are interested in: $\Pr(T)$.

2. Information about how prevalent the disease is in the general population: $\Pr(D)$.

In fact you can see that what really mattered for us is the ratio $\Pr(D)/\Pr(T)$. If this is small, we are in trouble, since that means that while only a few people in the population have the disease, our test returns true much more often. This form of error is called a false positive.

In this example $\Pr(T|D)$ was quite high, 0.9. If $\Pr(T|D)$ is too low, it means that the test is quite likely to be negative, even when people have the disease. This type of error is called a false negative.

We need to know that there aren't too many false positives or false negatives for the test to be useful. The surprising insight provided by Bayes theorem is that to use this test reliably as a doctor, it isn't enough to do a study that shows that out of a bunch of people who had the disease, the test detected it in almost all of them. This could be true and yet the test could still be really unreliable!

This video provides a simple geometric description of what is going on in Bayes Theorem. I would highly recommend watching it.

Work on the exercise in the handout. We will discuss the solution in a few minutes. The exercise demonstrates how updating our probabilities based on evidence can be less than intuitive. We will see how the tools we have discussed allow us to reason about probabilities formally and get an answer to this (and many other) questions.

5 The Monty Hall problem

Let's start with the case of 1000 doors. Initially, if there is no option to switch, you have only a 1/1000 chance of winning. Let's call the door you initially chose as A and your other option after Monty opens 998 doors as B. Indeed the intuitive reasoning in the handout is correct. Switching

will result in a much higher chance of winning since really we can think as if Monty is offering you an option between your door and the set of 999 other doors; he has just already opened 998 of them. Thinking this way, we should guess that by switching we can win with 999/1000 probability.

Another way to see this is to reason as follows. Clearly without the switch we have only a 1/1000 chance to win since we had to pick 1 door out of 1000. But if we pre-commit to not switching, then whatever happens later on, the chance the car is behind our door can't change. I mean if 1 out of a 1000 times the car is behind door A and we decide not to switch then we get the car only 1/1000 times. It's not like the car will magically appear behind our door if it is not there initially.

Yet another way to think is as follows. Let's calculate our chance of winning if we do switch. How can we lose though, it's only if the car was behind door 1 initially, and since we couldn't see any difference between the doors, we just picked a door uniformly at random, so the chance of the car being behind our initial pick is just 1/1000. In every other case, we win by switching!

These intuitive calculations all confirm that we can win with 999/1000 probability. Let's do the calculation one final time using Bayes theorem. Let A be our initial choice of door and B be the other remaining unopened door that we can switch to after Monty opened the 998 doors, each with a goat behind it. Let I be the event that your initial choice had the car behind it and E be the event that the 998 doors Monty opened have goats behind them. Monty always will open 998 doors with goats behind them no matter what door we pick, so $\Pr(E|I) = 1$ and $\Pr(E) = 1$. What about $\Pr(I)$? With no information about anything else, the chance that we initially hit upon the right door is just 1/1000. So

$$\Pr(I|E) = \frac{\Pr(E|I)\Pr(I)}{\Pr(E)} = \frac{(1/1000) \times 1}{1} = \frac{1}{1000}$$

But wait, why can't we reason as follows. Let's say L was the event that door B has the car. Well, it is also a single door, so $\Pr(L)$, given nothing else, should be 1/1000. And then

$$\Pr(L|E) = \frac{\Pr(E|L)\Pr(L)}{\Pr(E)} = \frac{1 \times (1/1000)}{1} = \frac{1}{1000}$$

Where is the mistake? The point is quite subtle. The thing is that door B is not a fixed door before our and Monty's actions. Door B is defined in terms of the door that is not chosen by us and is left unopened by Monty after he shows 998 goats. So really $\Pr(L)$ itself is the probability that door B has the car after Monty's actions. But as we discussed, this door will have the car in every scenario where our original choice didn't have the car. And since our choice was made before the additional information supplied by Monty, this is 999/1000 of the time!

To make sense of $\Pr(L)$ we have to unpack the definition of door B and keep in mind that door B is an outcome of Monty's actions that are guaranteed to remove 998 goats from our set of options. But then why isn't the chance of the car being behind each of the two remaining doors just half? This is because we didn't get to use this information in our decision that defined door A, so the chance that the car is behind door A remains fixed at 1/1000. Whereas for door B, in every scenario the car was not behind A, it is guaranteed to be behind B. Thus, the event L is actually precisely the event $\neg I$ since if the car is not behind our initial choice, then by the way B is defined, it is guaranteed to be behind door B, and so L will occur. Take some time to appreciate this point, it is quite a subtle one.

6 Independence

Our discussion on conditional probability leads us to observe that sometimes two events may not affect each other. That is observing that B happened tells us nothing about whether A is more

or less likely to happen. Another way to say this would be to say that observing B leaves the probability of A unchanged. That is

$$\Pr(A|B) = \Pr(A)$$

Simplifying this a bit we get

$$\frac{\Pr(A \cap B)}{\Pr(B)} = \Pr(A)$$

Or

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

If this is the case, we say that the events A and B are independent. If $\Pr(A|B) > \Pr(A)$ we say that the events are positively correlated, that is observing B makes A more likely to happen. Similarly if $\Pr(A|B) < \Pr(A)$ we say that the events are negatively correlated.

Excercise 3. *Show that if* $\Pr(A|B) > \Pr(A)$ *then* $\Pr(B|A) > \Pr(B)$.

It is worth observing that events being disjoint and events being independent are two very different things. A and B are disjoint if A and B can never occur together. That is $\Pr(A \cap B) = 0$. Independence , on the other hand, is when $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$. Two events could very well be disjoint and highly negatively correlated, since knowing that one event occurred tells you that the other event must not have occurred.

Suppose you have a fair die. Let E be the event that an even number rolled and P be the event that a prime number rolled. We have $\Pr(E \cap P) = |\{2\}|/6 = 1/6$ while $\Pr(E) = |\{2,4,6\}|/6 = 3/6 = 1/2$ and $\Pr(P) = |\{2,3,5\}|/6 = 3/6 = 1/2$. Thus

$$\Pr(P) \cdot \Pr(E) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} > \Pr(P \cap E)$$

Thus the events are negatively correlated. This makes sense because out of the three even numbers on a die, only one is prime, but of the three odd numbers, two are prime. So a number that is rolled being even reduces the chance that it is prime.

It is an important fact that correlation is not the same thing as causation, a misconception that it is wise to be wary of while reading any sort of statistics about real world phenomena. When A and B are positively correlated, often there is some third event that may be contributing to both A and B. For example, suppose you observe that ice cream sales and deaths by drowning are positively correlated. That is, if I tell you the number of ice cream sales is high in a given month in Chicago, you can confidently assert that probably more people than average died due to drowning in that month in Chicago. But of course, it's not very reasonable to conclude that eating ice cream is a factor that causes people to die of drowning. A more plausible explanation is that both these events are positively correlated with the weather being warmer, which leads to both higher rates of ice cream consumption and higher numbers of people swimming, and hence higher numbers of people drowning.

Also, A and B might be positively correlated and B and C might be positively correlated, but A and C need not be positively correlated. That is to say correlation is not transitive. In the words of Terrence Tao: "it is generally true that good exam scores are correlated with a deep understanding of the course material, and memorising from flash cards are correlated with good exam scores, but this does not imply that memorising flash cards is correlated with deep understanding of the course material"

7 The Probabilistic Method

We will now see a truly remarkable application of probability. We have already seen various ways in which you can prove that an object of a certain kind exists. You could do a direct proof, by constructing the object. But you could also argue by contradiction, showing that if such an object didn't exist, then something false must result as a consequence. This is an indirect proof and it may not be constructive, i.e, it may tell you such an object exists, yet not provide an concrete example. Another indirect method is the probabilistic method. Here we consider a set of objects and pick one at random. If we manage to show that there is a non-zero probability of picking one of the kind we are searching for, then definitely some such object must exist! Another way to think of this is to say that we describe a randomized process to generate objects. Then if it has any chance at all of generating the kind of object we want, then some such object must exist.

Recall we saw an example of the pigeon-hole principle, where we showed that out of any 6 people, either 3 are friends or there are 3 out of whom none are friends. In fact if there are only 5 people in total, then this is not guaranteed. This is a simple example in a topic of study called Ramsey theory. Let us define the Ramsey number $R(k)$ to be the smallest number n such that if there are n people, then a group of at least k size exists where the members of the group either are all friends or none are friends with anyone else in the group. Calculating $R(k)$ is surprisingly hard. $R(3) = 6$ as we have already seen. $R(4) = 18$. But already we only know that $43 \leq R(5) \leq 48$. And for the later Ramsey numbers we only have even wider estimates!

> Imagine an alien force, vastly more powerful than us landing on Earth and demanding the value of $R(5)$ or they will destroy our planet. In that case, we should marshal all our computers and all our mathematicians and attempt to find the value. But suppose, instead, that they asked for $R(6)$, we should attempt to destroy the aliens.
>
> - Paul Erdos

Yet we will be able to prove some estimates that are not that much worse than the best known results about Ramsey numbers! We will do this by using the probabilistic method.

Theorem 1. *For any $k \geq 3, R(k) > 2^{\frac{k}{2}-1}$.*

Proof. We consider an arbitrary group of n people and assume that the probability that each pair is friendly or not is $\frac{1}{2}$, independent of any other pairs. One way to think about this is to take each pair of people and flip a coin and assign whether they are friendly or not based on the outcome.

Then whether there are k people who are all mutually friendly or unfriendly has a total probability of $1/2^{\binom{k}{2}}$. To see this, we consider an arbitrary group of k people. Since their relationship status is assigned independently of each other with $\frac{1}{2}$, this becomes $\left(\frac{1}{2}\right)^p$ for p pairs. But since there are k people, there are $\binom{k}{2}$ possible pairs, so we get $p = \binom{k}{2}$.

Now, we note that there are two distinct possible configurations of interest: all $\binom{k}{2}$ pairs are friendly or they are all unfriendly. This means that we can double our probability to $2 \cdot \frac{1}{2^{\binom{k}{2}}}$.

But this is for an arbitrary group of k people out of n. We must now consider all possible such groups. There are $\binom{n}{k}$ of these. If we enumerate these groups, we can let K_i be the event that the i-th group of k people are either mutually friendly or unfriendly. We have $\Pr(K_i) = 2 \cdot \frac{1}{2^{\binom{k}{2}}}$ for $1 \leq i \leq \binom{n}{k}$. Then, by the union bound, we have

$$\Pr(\bigcup_i K_i) \leq \sum_{i=1}^{\binom{n}{k}} \Pr(K_i) = \binom{n}{k} \cdot 2 \cdot \frac{1}{2^{\binom{k}{2}}}$$

Note that if this probability is 1 , then this says that for our choice of n, we are guaranteed to have a group of k mutually friendly or unfriendly people. But if it is less than 1 then there must exist configurations where this is not the case. If we can argue that for $n \leq 2^{\frac{k}{2}-1}$, the probability is less than 1 then for all these values of n, we know that there are some configurations such that no k people are mutual friends or mutual enemies, which would prove the theorem.

First, we observe that $\binom{n}{k} \leq n^k$, which is a fairly rough estimate, and recall that $\binom{k}{2} = \frac{1}{2}k(k-1)$. So if $n \leq 2^{\frac{k}{2}-1}$ then

$$\begin{aligned}
\binom{n}{k} \cdot 2 \cdot \frac{1}{2^{\binom{k}{2}}} &\leq n^k \cdot 2 \cdot \frac{1}{2^{\binom{k}{2}}} \\
&\leq \left(2^{\frac{k}{2}-1}\right)^k \cdot 2 \cdot \frac{1}{2^{\binom{k}{2}}} \\
&= 2^{k\cdot(\frac{k}{2}-1)+1-\frac{1}{2}k(k-1)} \\
&= 2^{-\frac{k}{2}+1} < 1 \qquad \text{since } k \geq 3
\end{aligned}$$

This says that when we have groups of size up to $2^{\frac{k}{2}-1}$, there is a nonzero probability that they do not contain k mutually friendly or unfriendly people. In other words, to guarantee that our group contains such a subgroup, we need strictly more than these many people.

□