6.825 Techniques in Artificial Intelligence

Probability

- Logic represents uncertainty by disjunction

OK, today's the day to switch gears into a whole new part of the class. So far, we've been reasoning about the world with models of the world that are deterministic. They talk about the world as being certainly in one state or another and as evolving from state to state deterministically. Within logic, we have a way of allowing ourselves uncertainty, which is to use disjunction. So you can say the box is either red or blue, and I don't know which. There **is** a method for articulating uncertainty.

6.825 Techniques in Artificial Intelligence

Probability

- Logic represents uncertainty by disjunction
- But, cannot tell us how likely the different conditions are

And you can imagine, if you're really confused, using really big disjunctions to say, well, I don't know whether this or this or this or that or that. But the problem with really big disjunctions is that you don't have a way of saying which of these outcomes is more likely than which other ones. So if I try to drive to Boston, it might take me ten minutes, but it might also take me twenty minutes, or an hour, or a further amount of time, and in thinking about how I want to drive or whether I want to go at this time of day, I really need some kind of quantitative understanding of the relative likelihood of these different things. It's not enough to know that there are a whole bunch of different possible outcomes. I want to know that some of them are more likely than others.

6.825 Techniques in Artificial Intelligence

Probability

- Logic represents uncertainty by disjunction
- But, cannot tell us how likely the different conditions are
- Probability theory provides a quantitative way of encoding likelihood

So in this second part of the course we're going to concentrate on probabilistic methods and models of the uncertainty in the world, and the big thing that they give us is the ability to attach numbers to the likelihood of various kinds of results.

Some experience with probability is a prerequisite for this course, so I assume that you've seen it once before, but today I'm going to just go over some of the basics, mostly just to establish a common vocabulary because I think we're going to be looking at it and using it in a way that may be different from the way that it's been introduced to you.

Foundations of Probability

First of all, I'd like to talk a little bit about the foundations of probability. People have been thinking about probability for a long time. Hume is probably the first philosopher who had a semi-modern but interesting view of probability, and he worried a lot about the problem of induction. Induction is going to be really important to us when we get to machine learning. How do I know that just because the sun has come up every other day before today, how do I know that it's going to come up tomorrow? That was a really big problem for Hume, and it often is for machine learning, as well.

Foundations of Probability

Is coin-flipping deterministic?

So let's think about coin-flipping for a minute. Do you think that coin-flipping is a deterministic or a non-deterministic process?

One answer is that it's deterministic; that if you knew everything about the coin flipping process, then you could predict the outcome. If you knew the initial conditions, if you knew the forces, if you knew the wind currents in the room, if you knew whatever there was to know, if you knew all that, then there would be an actual fact of the matter. It's going to land heads or it's going to land tails, and that's determinate, and there's no uncertainty. There's no deep uncertainty in the process. Any uncertainty we have about the outcome is really rooted in a lack of information.

Another view is that it's truly random. That no matter how much you know, you can't actually predict how the coin will come up. Maybe most modern physicists don't believe this about coins any more. But they do believe it about all sorts of other phenomena at the quantum level.

But even at the level of quantum mechanics, it's not absolutely clear that there's not really another story underneath there that would make it deterministic. We don't happen to know it, but maybe there is. There might be a story that would remove all the uncertainty. Or it would at least push the uncertainty down into yet another different level of the story. We're not going to get into the whole question of reductionism, but it's an important thing to at least keep in mind that we use this term uncertainty to really talk about two kinds of things, To talk about real randomness in a process and to talk about our uncertainty about a process.

Foundations of Probability

Is coin-flipping deterministic?

P(the sun comes up tomorrow) = 0.999

What if I said that the probability that the sun will come up tomorrow is 0.999? You'd probably disagree, saying that certainly, we have many more than 1000 examples of the sun coming up. 0.999 seems like too small a number.

Foundations of Probability

Is coin-flipping deterministic?

P(the sun comes up tomorrow) = 0.999

- Frequentist
  - Probability is inherent in the process
  - Probability is estimated from measurements

Probs can be wrong!

Probs can be inconsistent!

The standard view of probability, and the one that, if you took a statistics class is certainly the one that you were exposed to, is the frequentist view. And it says that probability is really statements of frequency. That is, in saying the probability of the sun coming up tomorrow is 0.999, you are saying that one out of a thousand times, it's not going to come up. And that the way that you can get that probability is by watching this event over and over and over lots of times, multiple trials, and measuring it.

In the frequentist view, that's what it means to be a probability. You estimate it by measuring it, and usually the idea is that the probability is something that's inherent in the process.

Foundations of Probability

Is coin-flipping deterministic?

P(the sun comes up tomorrow) = 0.999

- Frequentist
  - Probability is inherent in the process
  - Probability is estimated from measurements

If you want to take a frequentist approach to something like "the sun comes up tomorrow," it really matters what we mean by “tomorrow”. The question is, does it refer to actual tomorrow, or tomorrows in general? Because if you want to say, "Well, let me look and see how many days there have been in the past that we've seen the sun come up" and do some statistical estimation of the probability, that's somehow implying that today is like yesterday which is like the day before, which is like the day before that, which is like the actual tomorrow, and that therefore, that whole set can be taken together and thought of as samples of whether or not the sun comes up. So whether or not the sun comes up on actual tomorrow-- that particular question has never been tested. We have no data about that, so what can we measure? How could we gather frequentist information about whether the sun's going to come up tomorrow? Maybe we can't..

Foundations of Probability

Is coin-flipping deterministic?

P(the sun comes up tomorrow) = 0.999

- Frequentist
  - Probability is inherent in the process
  - Probability is estimated from measurements
- Subjectivist (Bayesian)
  - Probability is a model of your degree of belief

I think the frequentist view of probability is fraught with complications. It's very hard, I think, to get the story exactly right. We could go and try to do that, but I'm going to advocate a different approach, which is also, I think, much more useful for AI, which is the subjective approach, or sometimes called the Bayesian view, and that's that probability is a model of your degree of belief; your personal, private degree of belief. And so then it's no longer correct to say "the probability" that this coin comes up heads, or "the probability" that the sun comes up tomorrow. It's the degree to which I think this coin is going to come up heads, or the degree to which I think the sun is going to come up tomorrow.

Foundations of Probability

Is coin-flipping deterministic?

P(the sun comes up tomorrow) = 0.999

- Frequentist
  - Probability is inherent in the process
  - Probability is estimated from measurements
- Subjectivist (Bayesian)
  - Probability is a model of your degree of belief

Probs can be wrong!

Here's an interesting difference between the frequentist and the subjectivist views. In the frequentist view, you can be wrong. You could say, "I think the probability of this coin coming up heads is 0.6," and a frequentist could hit you on the head and say, "No, it's 0.4." There is a fact of the matter to argue about.

Foundations of Probability

Is coin-flipping deterministic?

P(the sun comes up tomorrow) = 0.999

- Frequentist
  - Probability is inherent in the process
  - Probability is estimated from measurements

Probs can be wrong!

- Subjectivist (Bayesian)
  - Probability is a model of your degree of belief

Probs can be inconsistent!

In the subjectivist view, you can't be wrong. It's like pain. You can't be wrong about whether you're in pain. You can't be wrong about your beliefs. Well, it turns out that you can be a little bit wrong. In the subjectivist probability view, you can't be wrong, per se, but you can be inconsistent. And we'll spend a little bit of time in this lecture exploring what it means to be inconsistent, and why you shouldn't be inconsistent.

Axioms of Probability

Let's talk about the axioms of probability theory, the basic ideas of probability, and then we'll think about whether that formal system is a good map or model for people's degree of belief about things in the world. It's easy to argue that the formal system does a good job of telling the frequentist story, which you probably already understand at a basic level, but the connection to the Bayesian story is more interesting.

Axioms of Probability

- Universe of atomic events (like interpretations in logic).

Probability theory is a logic. It's a language for taking about the likelihood of events. We start with a universe, and we'll just talk about the discrete case. Maybe when we get into learning we'll do a little bit of stuff in continuous probability, but the discrete case is enough for now. So you have some universe of atomic events, things that could happen or ways the world could be. You could almost think of atomic events as being like interpretations, back in the logical world.

Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events

An event, then is a set of atomic events, which is also a subset of the universe.

Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events $\rightarrow$ [0,1]

We're going to want to talk about the probability of events. How likely is it that some event will occur? Remember, again, back in logic, a formula in logic describes a set of interpretations. It talks about some set of ways the world could be. We're going to talk about sets of ways the world could be, also, but instead of assigning truth value one or zero to a set of ways the world could be, we're going to assign a probability between zero and one. So anything in between zero and one to a set of ways the world could be. You can think of a probability distribution as a function that maps events into the range zero and one.

Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events $\rightarrow$ [0,1]

In the discrete case, you can think of the probability of an event as being the proportion of the whole universe that is contained in that event. By definition, probability satisfies the following properties.

Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events $\rightarrow$ [0,1]
  - P(true) = 1 = P(U)

The probability of True is one. The probability of True is the probability of the universe, the probability that something in this realm of discussion that we have available to us is actually the case, is one. So when you say, "Here's my universe," and you say, "These are all the ways the world could be," well, the world's got to be in one of the ways that it can be.

Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events $\rightarrow$ [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P($\emptyset$)

The probability of False is zero. If you think of it in terms of atomic events, false is the empty set. So the probability that none of these events is happening is zero. There's a whole bunch of ways the world could be, and one of them is actually the case. So far, this just maps onto propositional logic directly.

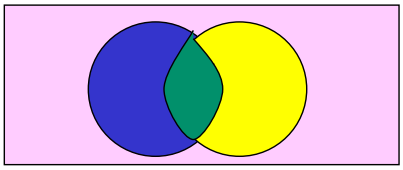
Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events $\rightarrow$ [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P(∅)
  - P(A v B) = P(A) + P(B) – P(A Æ B)

Then we really have just one more axiom, that the probability of A or B is the probability of A plus the probability of B minus the probability of A and B.

Axioms of Probability

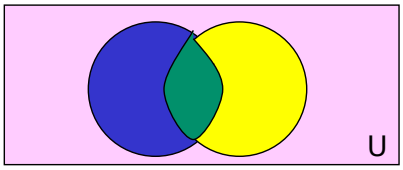
- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events $\rightarrow$ [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P(∅)
  - P(A v B) = P(A) + P(B) – P(A Æ B)



I'm sure you've all seen this argument from a Venn diagram, but we'll go over it again to be sure.

Axioms of Probability

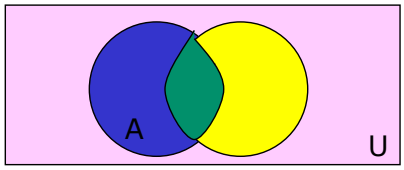
- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events $\rightarrow$ [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P(∅)
  - P(A v B) = P(A) + P(B) – P(A Æ B)



Let the pink box be the universe.

Axioms of Probability

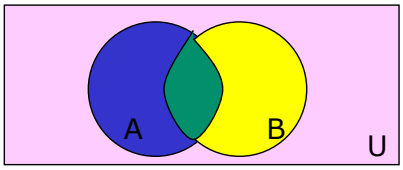
- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events → [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P(∅)
  - P(A v B) = P(A) + P(B) – P(A Æ B)



And let A be some event, some subset of the universe

Axioms of Probability

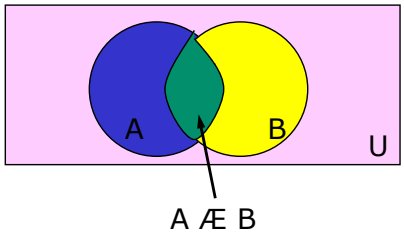
- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events → [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P(∅)
  - P(A v B) = P(A) + P(B) – P(A Æ B)



and let B be another event, another subset of the universe.

Axioms of Probability

- Universe of atomic events (like interpretations in logic).
- Events are sets of atomic events
- P: events → [0,1]
  - P(true) = 1 = P(U)
  - P(false) = 0 = P(∅)
  - P(A v B) = P(A) + P(B) – P(A Æ B)



Now, let's think about the probability that A or B is true. The probability that A is true is the proportion of atomic events that are in A. The probability that B is true is the proportion of atomic events in B. The probability that A or B is true is the proportion of atomic events that are in the union of A and B. So, how big is that union? We could say it's as big as A plus B, but that's not exactly right, because we'd be counting the atomic events in the overlapping section twice. So, we have to correct for that.

This argument leads us to the axiom that the probability of A or B is the probability of A plus the probability of B, minus the probability of A and B (that's the proportion of events in the overlap, which we don't want to count twice).

That's all you ever need to know about discrete probability. Everything else about probability is a consequence of these axioms (but some of the consequences are more obvious than others!).

Recitation Problem I

Prove that

- P(¬ A) = 1 – P(A)
- P(A v B v C) =
  P(A) + P(B) + P(C) –
  P(A ÆB) – P(A Æ C) – P(B Æ C) +
  P(A ÆB Æ C)

Here are a couple of simple probability exercises. Please do them before the next recitation.

A Question

Jane is from Berkeley. She was active in anti-war protests in the 60's. She lives in a commune.

- Which is more probable?
  1. Jane is a bank teller
  2. Jane is a feminist bank teller

Now I want to ask you a question. Try to answer it as if you're a normal human being (and not a hyperintellectual student ☺). Just read it on the slide and think about it.

A Question

Jane is from Berkeley. She was active in anti-war protests in the 60's. She lives in a commune.

- Which is more probable?
  1. Jane is a bank teller
  2. Jane is a feminist bank teller

1. A
2. A Æ B

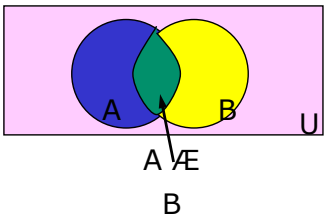
If you ask a group of regular people this question, the overwhelming answer is 2, that Jane is a feminist bank teller. But that answer is inconsistent with the theory of probability. No matter what the words mean, we can think of "Jane is a bank teller" as event A and "Jane is a feminist" as event B.

A Question

Jane is from Berkeley. She was active in anti-war protests in the 60's. She lives in a commune.

- Which is more probable?
  1. Jane is a bank teller
  2. Jane is a feminist bank teller

1. A
2. A Æ B



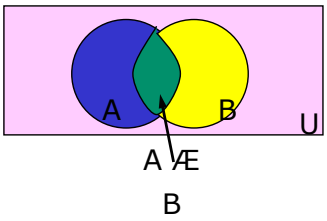
Then the question is, which is more probable: A, or A and B? But A and B has to be a subset of A. So it can never be the case that A and B is more probable than A by itself. A and B has more restrictions and conditions.

A Question

Jane is from Berkeley. She was active in anti-war protests in the 60's. She lives in a commune.

- Which is more probable?
  1. Jane is a bank teller
  2. Jane is a feminist bank teller

1. A
2. A Æ B



This example isn't mine. There's actually a fascinating field of psychology where they demonstrate all the ways in which people are clearly and systematically not correct probabilistic reasoners. It's interesting, and it leads you to cognitive theories of how it is that people may do their uncertain reasoning instead, and one big idea is that there is a notion of prototypes. So when you read this story, you get a prototypical Jane in mind. She probably has Birkenstocks on. And you think about that Jane and she doesn't seem like a bank teller. But she does seem like a feminist, and so answer number two just seems so much more attractive. But in terms of probability theory, it's not.

Dutch Book

So there's inconsistency in some people's beliefs, and now I can show that if you have this inconsistency, I ought to be able to cause you to make a set of bets with me that will allow me to win, no matter what happens. So let me do this. This is also sort of a parlor trick, but I like it. This way of arranging bets is called a "Dutch book". I'm not sure why, exactly. It's an old term from theory of probability and philosophy. The phrase “to make book” in English means to set the odds on a bunch of events, like the guys at the horse races do. The idea here is that if you have inconsistent beliefs, then I can set some odds on some bets such the bets are attractive to you and you will want to take them. But I can prove that no matter what happens in the world I win.

Dutch Book

- You believe
  - P(A) = 0.3
  - P(A Æ B) = 0.4 (and also that P(¬ (A Æ B)) = 0.6)

So imagine that you assign the probability of A to be 0.3 and the probability of A and B to be 0.4. I can show that I can get you into trouble that way.

Dutch Book

- You believe
  - P(A) = 0.3
  - P(A Æ B) = 0.4 (and also that P(¬ (A Æ B)) = 0.6)

You		Bet Stakes		A Æ B	¬ A ÆB	A Æ ¬B	¬ A Æ ¬ B
A	0.3						
A Æ B	0.4						

Here's a chart. Let's say you believe proposition A with probability 0.3, and proposition A and B with probability 0.4. There's an example like this in the book, but it's of a different problem.

Dutch Book

- You believe
  - P(A) = 0.3
  - P(A Æ B) = 0.4 (and also that P(¬ (A Æ B)) = 0.6)

You		Bet Stakes		A Æ B	¬ A ÆB	A Æ ¬B	¬ A Æ ¬ B
A	0.3	**A**	**3 to 7**				
A Æ B	0.4						

If you take a 3 to 7 bet on some condition C, then if C turns out to be true, you lose 7, but if it's false, you win 3.

Now, I offer you a bet. I say, if Jane turns out to be a bank teller (let that be our interpretation of A, just for intuition), then you pay me 7 dollars, but if she doesn't, I'll pay you 3 dollars.

Dutch Book

- You believe
  - P(A) = 0.3
  - P(A Æ B) = 0.4 (and also that P(¬ (A Æ B)) = 0.6)

You		Bet Stakes		A Æ B	¬ A ÆB	A Æ ¬B	¬ A Æ ¬ B
A	0.3	**A**	**3 to 7**				
A Æ B	0.4	**¬(A ÆB)**	**6 to 4**				

If you take a 3 to 7 bet on some condition C, then if C turns out to be true, you lose 7, but if it's false, you win 3.

And I also offer you another bet. If Jane turns out not to be a feminist bank teller, you'll pay me 4 dollars, but if she does, I'll pay you 6.

Since the odds on these bets match up with your beliefs, you should want to take the bets. In fact, people often **define** subjective probability as willingness to bet. So, in fact, if you're willing to take these bets, then we know something about your beliefs.

Dutch Book

- You believe
  - P(A) = 0.3
  - P(A Æ B) = 0.4 (and also that P(¬ (A Æ B)) = 0.6)

You		Bet Stakes		A Æ B	¬ A ÆB	A Æ ¬B	¬ A Æ ¬ B
A	0.3	**A**	**3 to 7**	-7	3 + ε	-7	3 + ε
A Æ B	0.4	**¬(A ÆB)**	**6 to 4**	6 + ε	-4	-4	-4

If you take a 3 to 7 bet on some condition C, then if C turns out to be true, you lose 7, but if it's false, you win 3.

To make these bets even a little bit more attractive (so that you think you'd actually make money, rather than break even), I can even add a bit to the positive outcomes for you.

So we can fill in the table, considering all the possible outcomes of events A and B, and writing in the cell how much money you will win or lose.

Dutch Book

- You believe
  - P(A) = 0.3
  - P(A Æ B) = 0.4 (and also that P(¬ (A Æ B)) = 0.6)

You		Bet Stakes		A Æ B	¬ A ÆB	A Æ ¬B	¬ A Æ ¬ B
A	0.3	**A**	**3 to 7**	-7	3 + ε	-7	3 + ε
A Æ B	0.4	**¬(A ÆB)**	**6 to 4**	6 + ε	-4	-4	-4

- No matter what the state of the world, you lose

If you take a 3 to 7 bet on some condition C, then if C turns out to be true, you lose 7, but if it's false, you win 3.

If you add up each of those columns, each of the ways the world could be, no matter what happens in the world, I win. And it's because you assigned a higher probability to A and B than you do to A. It doesn't matter what B is. You just shouldn't do that.

Dutch Book

- You believe
  - P(A) = 0.3
  - P(A Æ B) = 0.4 (and also that P(¬ (A Æ B)) = 0.6)

You		Bet Stakes		A Æ B	¬ A ÆB	A Æ ¬B	¬ A Æ ¬ B
A	0.3	**A**	**3 to 7**	-7	3 + ε	-7	3 + ε
A Æ B	0.4	**¬(A ÆB)**	**6 to 4**	6 + ε	-4	-4	-4

- No matter what the state of the world, you lose
- This is because your beliefs are inconsistent

If you take a 3 to 7 bet on some condition C, then if C turns out to be true, you lose 7, but if it's false, you win 3.

If your beliefs are consistent, I can't do this to you. If they follow the axioms of probability, I can't arrange to make money off of you no matter what happens. Obviously, even if your beliefs are consistent, you could lose money in some circumstances, but you'd also win money in some other circumstances. But if your beliefs are inconsistent with the laws of probability, I can guarantee that you always lose. So let that be our motivation for wanting to codify beliefs using laws of probability. It is obviously, patently not what people do, but it's probably at least a good foundation upon which to build computer systems that try to do a good job of solving problems in the world.

Random Variables

- Random variables

Now we're going to develop a set of tools for understanding and computing probabilities in complex domains. We'll start by talking about random variables.

Random Variables

- Random variables
  - Function: discrete domain $\rightarrow$ [0, 1]

The cliche about random variables is that they're neither random nor variables. You can think of a random variable as a function from some discrete domain, in our case, into zero and one.

Random Variables

- Random variables
  - Function: discrete domain $\rightarrow$ [0, 1]
  - Sums to 1 over the domain

It is also required that the probabilities assigned by the random variable to all values in the domain sum to 1. You can think of a random variable as describing a probability distribution in which the atomic events are the possible values that the variable could take on.

Random Variables

- Random variables
  - Function: discrete domain $\rightarrow$ [0, 1]
  - Sums to 1 over the domain
    - Raining is a propositional random variable

We'll mostly look at propositional random variables, things like “is it raining or is it not.”

Random Variables

- Random variables
  - Function: discrete domain $\rightarrow$ [0, 1]
  - Sums to 1 over the domain
    - Raining is a propositional random variable
    - Raining(true) = 0.2
      - P(Raining = true) = 0.2

To say that it's raining with probability 0.2, is to say that the raining random variable takes on value true with probability 0.2. We could write it like this to make the point about random variables being a function. But the more usual way to write it is P(raining = true) = 0.2

Random Variables

- Random variables
  - Function: discrete domain $\rightarrow$ [0, 1]
  - Sums to 1 over the domain
    - Raining is a propositional random variable
    - Raining(true) = 0.2
      - P(Raining = true) = 0.2
    - Raining(false) = 0.8
      - P(Raining = false) = 0.8

Now, because there are only two possible values for a propositional random variable, and they have to sum up to 1, then if the probability that raining is true is 0.2, then the probability that raining is false **has** to be 0.8.

Random Variables

- Random variables
  - Function: discrete domain $\rightarrow$ [0, 1]
  - Sums to 1 over the domain
    - Raining is a propositional random variable
    - Raining(true) = 0.2
      - P(Raining = true) = 0.2
    - Raining(false) = 0.8
      - P(Raining = false) = 0.8
- Joint distribution
  - Probability assignment to all combinations of values of random variables

Now, if you have multiple random variables, we can talk about their joint distribution. And that's really the probability assignment to all combinations of the values of the random variables. In general, the joint distribution cannot be computed from the individual distributions (which are typically called the "marginal" distributions).

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

Here's an example of the joint distribution. Imagine that we're embarking on a little dentistry and we want to understand the relationship between having toothaches and having cavities. Our domain has two random variables in it, two propositional random variables. Does the patient have a toothache or not? Does the patient have a cavity or not? So you can make a little table with spaces for the cross product of the values of the random variables. There's cavity, not cavity, toothache, not toothache, and then we have to fill in some probabilities.

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

- The sum of the entries in this table has to be 1

What do we know about the sum of the values in that table? That they have to add up to one. Remember, we said that the probability of the universe was one. Each cell in the table corresponds to an event. These events are mutually exclusive (it's impossible for any two of them to occur at once), so there's no overlap. And there are no other ways the world could possibly be. So, the probability of the universe is the sum of these probabilities.

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain

Given this table, given the joint distribution of all the variables in your domain, you can answer any probability question that anybody would ever ask you. That's, in some sense, all there is to know. And if you're in a domain that only has two variables, that's cool. You can just make up a table and then answer questions with it. We'll see how to answer some questions. But obviously, in a domain that's very big, you don't want to ever have to make that table, and so what we're going to do is spend the next couple of weeks looking at ways of doing probabilistic reasoning, taking advantage of some structural properties of your knowledge in the world to let you ask and answer probability questions without making the whole table. But today we'll use the table just to illustrate some points so that we know what the underlying definitions of various things are.

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]

What is the probability of cavities? The probability of cavity is 0.1. How do we compute it? We sum over all the different situations in which someone could have a cavity (in this case, with and without a toothache).

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]

Now, what's the probability of toothache? We get 0.05, by adding up the elements of the toothache column.

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

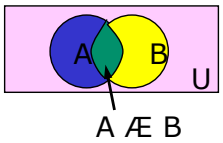
- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]
- P(A | B) = P(A Æ B)/P(B) [prob of A when U is limited to B]

Now let's look at the notion of conditional probability. We'll introduce the probability of A **given** B. What's the probability that A is true, if we already know that B is true? It's defined to be the probability of A and B divided by the probability of B.

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]
- P(A | B) = P(A Æ B)/P(B) [prob of A when U is limited to B]

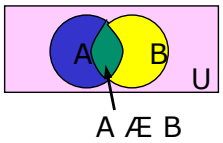


This is sort of like saying we are restricting our consideration just to the part of the world in which B is true; then what proportion of events in A? So if we look at the Venn diagram again, It's like we're going to compute the probability of A, but we're going to (temporarily) take the universe to be B, rather than U. So this conditional probability is the ratio of the green area to the green and yellow areas.

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]
- P(A | B) = P(A Æ B)/P(B) [prob of A when U is limited to B]
- P(cavity | toothache)

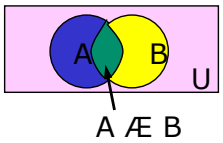


So in our example, we can ask, what is the probability that someone who comes into the dental office with a toothache has a cavity. That would be the probability of cavity given toothache.

Joint Distribution Example

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]
- P(A | B) = P(A Æ B)/P(B) [prob of A when U is limited to B]
- P(cavity | toothache) = 0.04/0.05 = 0.8



And how can we compute that probability? We start by figuring out the probability of cavity **and** toothache (which corresponds to the green area in the diagram), which is 0.04 (we can just look it up). Then we divide by the probability of toothache (corresponding to the green and yellow areas of the diagram), which is 0.05, yielding a result of 0.8. So, the probability that someone has a cavity, in general, is 0.1. But if we know they have a toothache, then the probability that they have a cavity goes up to 0.8. Ouch!

This is the structure of the kind of reasoning that we'll be doing a lot of. You start out with a prior probability of 0.1 that someone has a cavity. The idea here is that you walk into this problem having some belief about the likelihood that a patient has a cavity. Now, maybe this belief was derived from past experience, maybe it's derived from reading textbooks, maybe it's derived from talking to other people. Who knows what, but it's your personal, private belief that the next person walking through your door is going to have a cavity. Then you ask him, "How do you feel," and they say, "I have a toothache." Then if you know the probability of cavity and toothache, then you can figure out the probability of cavity in that case. We'll actually see a way of automating this process and see exactly what information you need to make it work out pretty efficiently. But that's how the story goes.

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)

There's another way to write down a conditional probability. You can say that the probability that A given B is equal to the probability of B given A times the probability of A, divided by the probability of B. This is just an algebraic massaging of the stuff that we already know. And on the face of it, it doesn't necessarily seem very useful. But what we'll see is that in fact it's a kind of reasoning that we naturally do all the time.

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
  - P(disease | symptom)
    = P(symptom | disease) P(disease)/ P(symptom)

Let me write this again with different kind names for the variables. You might, for instance, want to ask, what's the probability of some disease given this symptom. Well, that's the probability of the symptom given the disease, times the prior probability of the disease divided by the probability of the symptom.

So a common reasoning pattern is that start with a prior probability of the disease. This is our belief about the state of affairs when we don't have any evidence at all. Then, we get some information. Somebody tells me this person has a symptom. And then we would like to update our beliefs. We'd like to take what we used to believe and the evidence that we just got and combine them together and compute a new belief, a new degree of belief in whether this patient has this disease. In order to do that, we need, apparently, to know two things. We need to know the probability of this symptom, although we'll come back to that. It turns out that we can kind of finesse that. But what we really need to know is the probability of the symptom, given the disease. What's the probability, for instance, of having a toothache given that I have a cavity?

Why wouldn't I want to just learn the conditional probability of disease given symptom to begin with, rather than having to compute it from other things? The answer is that these conditional probabilities, of symptom given disease, tend to be more generally useful, more true across a broad range of situations.

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
  - P(disease | symptom)
    = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US

Imagine that the disease is bovine spongiform encephalopathy, or BSE. That is, mad cow disease. I have a cow that is paralyzed, and I want to know the probability that it has BSE. Here in the US, it's probably pretty low. But what if I were in England? Then, the probability might be higher.

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
  - P(disease | symptom)
    = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US
    - P(symptom | disease) should be the same
    - It is more useful to learn P(symptom | disease)

Why? Because the base rate is different in the two different places. But we expect the disease process to be essentially the same, so that the probability of the symptom of paralysis given the disease BSE remains the same on both sides of the Atlantic.

It turns out that in all kinds of domains it's easier and more useful and more generally applicable to learn these causal kinds of relationships and compute the diagnostic information when necessary using your base rate, than to try to learn these diagnostic probabilities directly. So this is how diagnostic systems normally get built. It also lets us chain evidence together in a way that I'm going to do next.

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
  - P(disease | symptom)
    = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US
    - P(symptom | disease) should be the same
    - It is more useful to learn P(symptom | disease)
- Conditioning
  - P(A) = P(A | B) P(B) + P(A | ¬B) P(¬B)

What's left to deal with is probability of the symptom. It turns out that there's an easy way you can deal with the probability of symptom. And we can do it using a process that's awfully useful to know about. There's a standard maneuver in probability called conditioning. Here's the general rule. We can say the probability of A is equal to the probability of A given B times the probability of B + probability or A given not B times the probability of not B.

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
  - P(disease | symptom)
    = P(symptom | disease) P(disease)/ P(symptom)
  - Imagine
    - disease = BSE
    - symptom = paralysis
    - P(disease | symptom) is different in England vs US
    - P(symptom | disease) should be the same
    - It is more useful to learn P(symptom | disease)
- Conditioning
  - P(A) = P(A | B) P(B) + P(A | ¬B) P(¬B)
    = P(A Æ B) + P(A Æ ¬B)

It's pretty easy to prove. We can see, by the definition of conditional probability, that the first term is equal to the probability that A and B, and the second term is the probability that A and not B. So the probability of A and B **or** A and not B is equal to the sum of the probabilities (since they don't overlap), which is equal to the probability that A.

We could use the conditioning rule to compute the probability of the symptoms. It's going to turn out to be the probability of symptom given disease times the probability of the disease, plus the probability of the symptoms given no disease times the probability of no disease.

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) · P(B)

I'm sure you've come across the idea of two events being independent. We'll say A and B are independent, if and only if the probability that A **and** B are true is the product of the individual probabilities of A and B being true.

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) · P(B)
  - P(A | B) = P(A)

There are two other ways to say the same thing, that give a different intution. One is that the probability of A given B is the same as the probability of A. That means that knowing that B is true doesn't give us any more information about the truth of A.

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) · P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)

We can, equivalently, turn this around in the other direction.

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) · P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)

Somebody's height and weight are not independent. Whether you get a good grade in this class and listen to the lectures, I hope are not independent. What you ate for breakfast and how you do in this class probably are independent, except that there are studies that show the students that don't eat breakfast don't do well in school. It may be that at some level everything is dependent, but we can at least make some kind of abstraction of independence just to get on with things, because it's going to turn out that independence relations are the key to doing probabilistic reasoning efficiently.

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) · P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)
- Independence is essential for efficient probabilistic reasoning

If every possible thing you could know bears on every possible other thing that could be, then in some sense there's nothing to but to consider completely specified world states all at once. But if these things can be kind of taken apart a little bit, if you can think about breakfast without thinking about the color of the car you drive, then that would be good, and so we're going to look at ways of using independence relations to make reasoning more efficient.

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) · P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)
- Independence is essential for efficient probabilistic reasoning
- A and B are conditionally independent given C iff
  - P(A | B, C) = P(A | C)

There's a more general notion, which is called conditional independence. We'll say that A and B are conditionally independent given C if and only if the probability of A given B and C is equal to the probability of A given C.

To understand this, assume someone told you C. Now, the question is, if they were also to tell you B, would that change the probability that A is true? And if A and B are conditionally independent given C, then if you already know C, B won't tell you anything about A.

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) · P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)
- Independence is essential for efficient probabilistic reasoning

- A and B are conditionally independent given C iff
  - P(A | B, C) = P(A | C)
  - P(B | A, C) = P(B | C)
  - P(A Æ B | C) = P(A | C) · P(B | C)

We can also write this in two other, equivalent ways.

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)

Imagine that now we have three propositions. We have whether you have a toothache, whether there's a dark spot in your X-ray -- we'll call it proposition X, a spot in your X-ray -- and whether you have a cavity.

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other

Again, in this way of thinking about primary causes and their symptoms, you might imagine that having a toothache is a symptom of having a cavity, and having a spot on your X-ray is a symptom of having a cavity, and certainly any pair of these variables is related, so if somebody walks in and says, "I have a toothache," then it's more likely that when you take the X-ray they're going to have a spot, and vice versa. So all these variables are related to each other.

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

But typically it will be the case that T and X are conditionally independent given C. The idea is that whether or not you have a toothache given that you have a cavity is not related to whether or not it's going to show up on the X-ray. All right? Whether it hurts and whether you can see it are independent given that it's there. This is a really crucial concept for the next two weeks, so I want to be sure that you've got the idea.

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)

Let me give you a car example, where it's more clear, maybe even than toothaches. We know more about cars than toothaches, most of us. Let's say something's wrong with your car, and you're considering the following proposition. One is that the battery is dead. One is that the radio works; that the radio plays when you turn it on, and the other one is that the starter turns over. This is an English idiom meaning that when you turn the key it makes noise.

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)
- None of these propositions are independent of one another

Are any of these propositions -- just two of them -- independent of one another? I don't think so. Because if you walk into a car and the radio plays, you're going to think it's more likely that the starter will turn over. But if the radio doesn't play, you might naturally think, hmm, maybe the battery's dead and so the starter will not turn over. So knowing this, here's the information about this. Any one of these gives you information about the other one.

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)
- None of these propositions are independent of one another
- R and S are conditionally independent given B

So, what if I tell you the battery is good? Definitely, I know it. I tested it. Now, does the radio playing tell you anything about the starter? No. Well, you know, it might, actually, if it shared a wire to some other place in the car, but naively, if the starter's connected to the battery and the radio's connected to the battery, and they don't share anything else, then given that I know the battery works or given that I know the battery's dead, either way, the radio doesn't give you information about the starter. So in this case, we would say that R and S are conditionally independent given B.

Combining evidence

- Bayesian updating given two pieces of information

So now let's think about how would you do Bayesian updating when you have multiple pieces of information to integrate. Imagine that a person walks in the door of my dentists office and they complain of a toothache, and I take an x-ray and it is positive. Then my problem is to update my belief in whether or not the patient has a cavity, given both of these pieces of information. Formally, I want to know the probability of C given T and X.

We're going to start writing big conditional probability formulas. They're just going to get bigger and have more things on the left and the right of the bar. If there are multiple variables on either side, assume they're conjoined. Assume that they’re connected with an and.

Combining evidence

- Bayesian updating given two pieces of information

$$P(C|T,X) = \frac{P(T,X|C)P(C)}{P(T,X)}$$

Here is the probability that the person has a cavity given that they have a toothache and there's a spot on their X-ray. How can we decompose it? Well, we can start by using Bayes' rule, which gives us the probability of T and X given C times the probability of C, divided by the probability of T and X.

Combining evidence

- Bayesian updating given two pieces of information

$$P(C|T,X) = \frac{P(T,X|C)P(C)}{P(T,X)}$$

- Assume that T and X are conditionally independent given C

Now, unless we make some assumptions, we can't really go any farther. Without making some assumptions, you can't know the probability of T and X, unless you've gone out and assessed it. Unless somebody's done studies of how many people have toothaches and spots on their X-rays, or how many people have both of those symptoms given they have a cavity. So what we typically do is make an assumption of conditional independence. We often assume that the manifestation of different symptoms is independent given that the disease is present.

Combining evidence

- Bayesian updating given two pieces of information

$$P(C|T,X) = \frac{P(T,X|C)P(C)}{P(T,X)}$$

- Assume that T and X are conditionally independent given C

$$P(C|T,X) = \frac{P(T|C)P(X|C)P(C)}{P(T,X)}$$

So in this case, let's assume that T and X are conditionally independent given C. That lets us take this expression apart. The probability of T and X given C becomes the probability of T given C times the probability of X given C.

Combining evidence

- Bayesian updating given two pieces of information

$$P(C|T,X) = \frac{P(T,X|C)P(C)}{P(T,X)}$$

- Assume that T and X are conditionally independent given C

$$P(C|T,X) = \frac{P(T|C)P(X|C)P(C)}{P(T,X)}$$

- We can do the evidence combination sequentially

Now, this is pretty cool. At least, the top looks really very nice. The top says, well, I started out with this prior probability of someone having a cavity, and I find out that there's a spot on their X-ray so I multiply in a factor that takes that evidence into account, and then I find out that they have a toothache and then I multiply in another factor that takes that into account, and you can probably see just from the structure, at least, of the top part of the formula that we could do this evidence combination sequentially if we wanted to. And we get the same answer, right? So if first you found out that the patient had a toothache, and then later you found out that you had the X-ray, you could just fold that in incrementally.

Normalizing Factor

$$P(C|T,X) + P(\neg C|T,X) = 1$$

So this part all looks pretty good, and the only problem we're left with is the normalizing constant P(T,X). The reason people call it a normalizing constant is because another way to look at it is that we know that the probability of C given T and X plus the probability of not C given T and X has to be one.

Normalizing Factor

$$P(C|T,X) + P(\neg C|T,X) = 1$$

$$\frac{P(T|C)P(X|C)P(C)}{P(T,X)} + \frac{P(T|\neg C)P(X|\neg C)P(\neg C)}{P(T,X)} = 1$$

Now, you can turn both of these things around using Bayes' rule (like we did on the last slide).

Normalizing Factor

$$P(C|T,X) + P(\neg C|T,X) = 1$$

$$\frac{P(T|C)P(X|C)P(C)}{P(T,X)} + \frac{P(T|\neg C)P(X|\neg C)P(\neg C)}{P(T,X)} = 1$$

$$P(T|C)P(X|C)P(C) + P(T|\neg C)P(X|\neg C)P(\neg C) = P(T,X)$$

And now multiply through by P(T,X), and you're done.

Recitation Problems II

- Show that P(A) >= P(A,B)
- Show that P(A|B) + P(~A|B) = 1
- Show that the different formulations of conditional independence are equivalent:
  - P(A | B, C) = P(A | C)
  - P(B | A, C) = P(B | C)
  - P(A Æ B | C) = P(A | C) · P(B | C)
- Conditional Bayes' rule. Write an expression for P(A | B,C) in terms of P(B | A,C).

Please do these problems for recitation.