Probabilistic reasoning

When an agent X is faced with an imminent decision about what to do, there might not be time to reason carefully and gather relevant data. In such a case, X might need to draw a conclusion about what to do based on easily available data and using quick reasoning methods. This is one kind of situation in which nonmonotonic reasoning can be useful.

Another such situation is one in which – although there might not be a time-crunch – some of the relevant data will be hard to get, or likely to change before X actually needs it. Then X might decide to go ahead to decide on a course of action anyway, trusting that the missing parts can be filled in later. Planning a trip from DC to San Francisco is like that: X cannot easily know details about ground transportation at the other end, but can infer nonmonotonically that it will work out somehow.

And in these cases, it might not be feasible – or even possible – to build into X's KB the full exact data that would allow X to draw conclusions that are more certain, less liable to error. An agent-designer is very unlikely to have obtained all the relevant data in advance, in cases that the designer cannot have anticipated.

However, there are situations in which it is possible to have a great deal of relevant data in advance. For instance, medical diagnosis of a current patient's condition can be based on very extensive data on previous cases; and machine translation of a current document can be based on extensive work on past documents. In such situations, it seems reasonable to assume that the current problem is just another instance of a vast number of others that have been previously encountered and that are all governed by the same underlying principles. If so, then it makes sense to base conclusions about the current problem in terms of how it matches up with the past ones.

However, the current case might easily fail to be an exact match with any previous case. It will resemble some cases in some ways, others in another, and so on. That is, in real-world domains there tend to be a vast array of individual details that are practically unknowable (variables that we cannot identify), so even if two cases were perfect matches we would not know.

This is where probabilities naturally fit in. Let U be the current problem, and let T, V, and C be three properties or characteristics or parameters of the general class of problem that U and many past ones belong to. For instance, the problem class might be that of deciding whether a cavity (V) is present, given current info concerning: reported toothache (T), and feeling a 'catch' with a dental probe (C). [This example is based on one in the book by Russell and Norvig.]

Then a robotic dental hygienist might consult a dental-data resource showing , say, that in 10,000 previous cases, it was found that a certain number N of reported

toothaches in fact corresponded to cavities; and a certain number M of positive probes (it catches) corresponded to cavities.

What does this mean for the current case (patient) U? Suppose U has a toothache and also there is a positive probe result. How do we use the resource data to determine the likelihood of a cavity in the case that T and C are both true?

We'll need to back up and start at the beginning, with elementary probability concepts, then conditional probability. Eventually (next lecture) we'll get to action decisions, such as whether to take an x-ray or not, based on so-called utility (risk and reward).

Elements of probability theory.

We suppose some sort of experiment (or activity) EXP is to be done, in the very general sense of some sort of data is to be gathered, somehow, and recorded. In advance, we do not know what the actual outcome will be in this case, but we may know what the possibilities are – all the ways the outcome might be.

We call the set of all possible outcomes of EXP the sample space, S. Each element e of S is called an *elementary event* (or "possible world"). Based on past experiments – or on assumptions we are willing to make – we might write P(e) =… for each e in S, where the … is filled in with whatever numerical value the past data or our assumptions indicate.

In the dental scenario, the experiment could be taken to be the full result after (say) asking about a toothache, probing and then taking an x-ray; and possible outcomes e could be these eight:

0.576 no toothache, no catch, no cavity (perfect teeth!)
0.008 no toothache, no catch, cavity (oops!)
0.144 no toothache, catch, no cavity
0.072 no toothache, catch, cavity
0.064 toothache, no catch, no cavity
0.012 toothache, no catch, cavity
0.016 toothache, catch, no cavity
0.108 toothache, catch, cavity

To each one of these, P would assign a number, P(e) – example values are show above on the left – based on past data or assumptions. Note that this is largely a truth-table in disguise, given all eight rows for three letters (with some extra information provided by the numbers).

Of course, if one is going to take an x-ray, then there is no need to worry about all this: just do it and know! But there may be reasons to estimate the likelihood of a cavity first: x-rays cost money; they present health risks; they take time. And that brings in utility theory, which we'll get to later on.

So, we are considering outcome for "problem" (or patient) U that is not yet known, so we don't know just which elementary event it is. That is, we have not yet done that experiment, so we rely on past data – or assumptions – to inform us about what to expect. That is what the probability function P is supposed to do.

We want to know the probability of something (a cavity, say) when there we know extra stuff, details about patient U: there is a toothache, or the probe catches, or both. How do we get this from the data? At first, the hygienist may learn T: patient U has a toothache. But there are lots of elementary events e in which T holds; four in fact. At this point, what is the hygienist to think? Each of the four has a certain probability, so maybe adding them up gives the probability of any one of them being the true outcome. And that is in fact how we define the probability in such a case.

Let E be a subset of S. Then E is called an event. Having a toothache (T being true) is an example. We define P(E) to be the sum of the P(e) values for all the e's in E. Note that E can be described by saying :the outcome is one of …" where the … are the elements of E; so we can regard E as a wff in some language. Then –E means the outcome is not any of the e's in E, etc.

Now we can give the axioms for elementary probability theory:

(i) 0 <= P(e) <= 1 for all e in S

(ii) Sum P(e) = 1 (sum over all e in S

The following are consequences of the axioms and the definition of P applied to events:

- If E & F is false (they have no elements in common), then P(E v F) = P(E) + P(F)
- P(-E) = 1-P(E)
- P(EvF) = P(e) + P(F) – P(E&F)
- P(false) = P(emptyset) = 0.
- P(true) = P(S) = 1

So now we have a way to express the probability of, say, having a toothache: P(T). In fact we can calculate it:

P(T) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

But that is not what we wanted! We wanted the probability of having a cavity, if we already know there is a toothache. Let's use V to mean there is a cavity. Then using the notation we introduced far above, this is written P(V | T) . Here we call T "conditional or posterior information". But how to we calculate P in such cases? We want somehow to exclude those elementary events e for which T is false. But if we

exclude them, the rest won't add up to 1. We settle this easily, by scaling things so they do add to 1:

Definition: P(E | F) = P(E & F) / P(F). [It is perhaps easier to remember in this form: P(E&F) = P(E | F) * P(F)]

We then find: P(V | T) = P(V&T) / P(T) = 0.12 / 0.2 = 0.6

So far, so good. Now what if U has a toothache and also the probe is positive? Then we want to find P(V | T&C). Again we calculate:

P(V | T&C) = P(V & T & C) / P(T & C) = 0.108 / 0.124 $\sim\sim$ 0.879

We are off to a good start. But one must be cautious in too quickly applying such ideas. Let's look at another example: *The Monty Hall problem*.

A prize is hidden behind one of three doors
You are allowed to initially pick one door (say door A, out of A B C)
Then you see one other door has no prize (say door B).
Decide whether to switch your choice to the remaining door (door C).

It seems that we want to find P(A | -B), where I now use each door letter to mean it has the prize. We compute: P(A|-B) = P(A&-B) / P(-B) = P(A) / P(-B) = (1/3) / (2/3) = 1/2. But this is wrong. We can go back to basics here and write out all elementary events (there are three), and for each consider switching vs not switching:

	switch	stay
A -B -C	lose	win
-A B -C	win	lose
-A -B C	win	win

One (of many) intuitive ways to think about it is this: there is a 2/3 chance that the prize is NOT behind door A [P(-A) = 2/3, same as for any one door not having the prize]. So that means there is a 2/3 chance it IS behind one of the other doors, B or C. And when we are shown that one of those has no prize, then the 2/3 chance has to be for the other of those two.

So what is wrong with the conditional result of 1/2? That result is correct if we have not yet chosen A, and are initially told –B, and need to decide whether to choose A (or C): if we start out knowing one door does not have the prize, then the other two share the probability 50-50. Why does choosing first matter? Well, it is a bit complicated, but our choice – along with knowledge of where the prize is – is part of what determines which other door is opened. And that is a lot of special information about choices and times and decisions – such as who picks which door to show us as having no prize, and why – that does not get included in the expression P(A|-B).

One more example: B is a biased coin, that ALWAYS lands head's up. U is unbiased, with a 50-50 chance of H or T. Here is the experiment:

1. Randomly pick B or U (without looking to see which it is); have an assistant write down which coin it is.
2. Toss the unknown coin twice, recording the two results as one of HH HT TH TT.

So the elementary events (outcomes) are these (with probabilities shown):

B H H 1/2
U H H 1/8
U H T 1/8
U T H 1/8
U T T 1/8

Suppose the two tossed give HH. We still do not know which coin it is, but now it seems much more likely that it is B. What is the exact probability that it is B, given the conditional HH information? That is, we want to find P(B | HH).

But P(B | HH) = P(B & HH) / P(HH)
= 1/2 / (1/2 + 1/8)
= ½ / (5/8) = ½ x 8/5 = 4/5

Note that we do not get 4/5 simply by adding the numbers for BHH and UHH above. This is because for conditional probability, the condition (HH in this case) removes some outcomes from consideration. So the probabilities of what is left (BHH and UHH) no longer sum to 1. Dividing by P(HH), as we did above in computing P(B | HH), amounts to renormalizing the numbers so that they do add to 1:

BHH 1/2 -> 1/2 x 8/5 = 4/5 = 80%
UHH 1/8 -> 1/8 x 8/5 = 1/5 = 20%

Let's get a few more standard notions out of the way:

Independent events E and F are ones satisfying P(E | F) = P(E); that is, whether or not F is given has no bearing on P(E). Equivalently, P(E&F) = P(E)P(F).

Bayes' Theorem: P(E | F) = P(F | E) P(E) / P(F)

The general direction all this leads to involves statistical notions such as expected value and the notion of utility, in order to perform automated decisions about actions to take. We will just barely have time to mention these things (and a few others) near the end of the semester. What we will turn to now is the general theme of actions: what they are, how to represent them, reason about them, form them into plans, decide which plans to enact, and carry them out. This too is far more than we

can go into in any great detail. But we will look at a few aspects a little more closely, while gesturing toward the rest.