Lecture 17 -- More probability, Bayesian Nets, Decision Theory

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:

A *random variable* is a function X(e) whose value may be any real number x, which may vary each time the experiment is done. An example is the date (in years, say) in the dental setting; others are the patient's age, number of past cavities, etc. (Here we would assume that all that information is kept in the patient's records and becomes part of what is meant by an outcome.) We define the *mean* or *average* or *expected* value <X> of a random variable X as follows, where X=x is the event of X having the value x, i.e., the set of all e's in S such that X(e)=x:

<X> = Sum x*P(X=x) where the sum is over all reals x.

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

Bayes' Theorem: $P(E \mid F) = P(F \mid E)\, P(E) / P(F)$

Now back to AI. One tool of considerable importance in various parts of AI is the Bayesian Network (or Bayes net). We will give a *very very* <u>*very*</u> brief introduction.

<u>Bayesian Networks.</u> Given an experiment and an associated sample space, a Bayesian network is an acyclic directed graph where each node represents a random variable X, and there are CPTs (conditional probability tables) giving P(X | ...) where the … are the parents of X (nodes with arrows to X). Intuitively, parent nodes values have an effect on the likelihood of a child node values.

Here is an example, based on work done by a member of my research team. The idea is that a "smart" system ought to know not only about how to do things, but also why it is doing them and how to tell if it is succeeding. That way it might be able to avoid compounding an error (making it worse) and instead take some sort of repair action. For this, we designed three ontologies: Indications (of anomalies), Failures (that may cause anomalies), and Responses (that might fix things). Then these in turn were connected with conditional probabilities. For example, an anomaly could be a mismatch between an action postcondition (the expected result) and the observed result; a cause (failure) could be a sensor-failure (the "observed result" was not what really happened), or a false precondition (the system's KB had it marked true); and responses could be to test the sensor or to check the precondition. For a given application, given an anomaly, some failures are more likely causes than others; and some responses are more likely to work than others.

One can assign initial conditional probability values more or less by guess, and then over time allow the system to adjust them as dictated by experience. This takes us in the direction of machine learning, next week's topic. But the basic architecture of the ontology module is that of a Bayesian network:

INDICATIONS ONT	FAILURE ONT	RESPONSE ONT
A	J [P(J|…)]	X [P(X|…)]
B	K [P(K|…)]	Y [P(Y|…)]
C	L [P(L|…)]	Z [P(Z|…)]
	M [P(M|…)]	

Where the [P(_|…)] are the CPTs showing the likelihood of the given item, based on conditional data from the ontology to the left.

Then is, say, an indication of type B occurs, the system can compute the probabilities of it being due to J, or K, or L, or M, and from that the probabilities of there being a successful result if it applies response X, or Y, or Z. Then it should choose the response with the greatest chance of success, unless that has drawbacks (cost, etc).

And this brings us to decision theory.

<u>Decision Theory.</u> Again we will content ourselves with a rather fanciful treatment of one example: to go to the dentist or not?

A patient has a toothache. Going to the dentist has a basic fee of $100.

If the probe catches (but not otherwise), the dentist will x-ray the patient's teeth, which will cost another $100 (no insurance!). And if the x-ray shows a cavity, it will have to be filled, another $100 cost.

But in *not* going, there is a risk of 0.95 that an undiscovered cavity will end up costing $1000 *more* than treating it now would cost ($300). (We'll assume this is independent of whether there is a toothache or not.)

So, going to the dentist seems to correspond to these costs and benefits:

-$100 basic fee
-$100 [x-ray] x P(C|T) = -$100 x 0.6 = -$60
-$100 [filling] x P(V|C&T) = -$100 x 0.879 = -$87.90
+$1000 [saved] x 0.95 x P(V|T) = $950 x 0.6 = $570

What shall we do with these numbers? It seems clear that we should add them, getting
-$100 -$60 -$87.90 + $570 = $322.10
as the net "value" of going to the dentist. (If course this ignores lots of other possible costs and benefits, but they can be treated in the same way, as long as we can translate all the costs and benefits into plausible numbers – which is controversial.)

And since the net value is positive, it seems that going is the wiser choice.

But what is it that we are really doing here?

We are really using the <u>principle of maximum expected utility</u>, in disguise (but making a mistake that will turn up below).

The question is whether to do an action (dental visit) or not. And it depends on what things that action might result in, the costs and benefits of those possible results, and their likelihoods.

If the dental is performed, an outcome involves a record concerning the following properties: Fee (F), Catch (C), and Cavity (V)

And we have relevant probabilities (where we know toothache (T)):
$P(F|T) = 1.0$
$P(C|T) = 0.62$
$P(V|T)=0.6$
$P(V|CT) = 0.879$
etc

For each possible outcome e, let U(e) be the net value (benefit – cost) of e.

Thus U(FCV) = -100 -100 -100 + 570 = $270, for example.

We then compute the “expected” value of U, over all outcomes, i.e., it’s average

<U> = Sum U(e) P(e|T)   [where the sum is over all elementary events e].

Computing this for the dental case (where there are four elem events), we get:

<U> = U(FCV) P(FCV|T)
+ U(F-CV) P(F-CV |T)
+ U(FC-V) P(FC-V |T)
+ U(F-C-V) P(F-C-V |T)

= -$300x0.54
-$1100x0.06
-$200x0.08
-$100x0.32

= -$162 – 66 – 16 – 32 = -$276

OK, now what do we do? We know what to expect if we go: on average, we’ll end up with a net loss of $276.

But what if we <u>don’t</u> go? Now the outcome properties are V and W (cavity gets worse). Also the costs are different. There is no $100 visit fee (F), no x-ray fee, no filling fee. Also there is no C or -C property.

So, we get

$$
\begin{aligned}
\langle U \rangle = \; & U(V{-}W)\, P(V{-}W|T) \\
& + U({-}V{-}W)\, P({-}V{-}W\, |T) \\
& \quad + U(VW)\, P(VW|T) \\
& \qquad U({-}VW)\, P({-}VW\, |T)
\end{aligned}
$$

$$
\begin{aligned}
= \; & -\$300 \text{x} P(V|T)P({-}W|V) + 0 - \$1300\text{x}\, P(V|T)P(W|V) + 0 \\
= \; & -\$300\text{x}0.60\ (0.05) \ - \ \$1300\text{x}0.60\text{x}0.95 \\
= \; & -\$6 - \$741 = -\$747
\end{aligned}
$$

Not-going is more expensive than going, by a difference of $471. This is less than the net difference we found at first, with our intuitive treatment. For instance, we are now taking into account that in not going, if there is a cavity there is still a (small) chance that it will not get worse and so when we eventually have it fixed it will still cost only $300.

The principle of maximum expected utility says that the best action is the one that gives the maximum value of <U>. While the above may make it look quite straightforward, many subtleties can arise in applying it in particular settings. Among these are ones of determining utilities, which are highly subjective, involving human judgments and preferences.

In fact, it is not always possible to determine utilities in a consistent manner, and when it is possible it may be that there are more than one way to assign values. Finally, it can take much effort to find the maximum <U>, and people tend to be content with a "good" <U> even if it not guaranteed to be the maximum. This is called "satisficing" – finding an action that gives a satisfactory <U>. Herbert Simon (one of the Dartmouth Ten) won the Nobel Prize in Economics for his work related to this idea. It is also an idea that has wide appeal in AI: perhaps an artificial agent intelligent would be better off taking an action that is good enough (for a given task) rather then putting effort into searching for the very best action possible.

While we will not go into it here, it is worth mentioning the BDI architecture for agents. This focuses on an agent's Beliefs-Desires-Intentions, thus broadening the notion of a KB to something more like a state of mind. The KB consists of the beliefs, the desires play a role much like goals (with possible preferences among them), and the intentions come into play when a plan is chosen (decided upon and queued for action).