Maximizing expected utility

Earlier we looked at many decision rules:

- maximin
- minimax regret
- principle of insufficient reason
- ...

The most commonly used rule (and the one taught in business schools!) is maximizing expected utility.

In this discussion, we assumed that we have a set S of states, a set O of outcomes, and are choosing among acts (functions from states to outcomes).

The good news: Savage showed that if a decision maker's preference relation on acts satisfies certain postulates, she is acting *as if* she has a probability on states and a utility on outcomes, and is maximizing expected utility.

- Moreover, Savage argues that his postulates are ones that reasonable/rational people should accept.

That was the basis for the dominance of this approach.

- We'll be covering Savage shortly.

Some subtleties

We've assumed that you are given the set of states and outcomes

- But decision problems don't usually come with a clearly prescribed set of states and outcomes.
  - The world is messy
  - Different people might model things different wawys

Even if you have a set of states and outcomes, even describing the probability and utility might not be so easy ...

- If the state space is described by 100 random variables, there are 2^{100} states!

Some issues for the rest of the course:

- Finding the right state space
- Representing probability and utility efficiently

Three-Prisoners Puzzle

- Two of three prisoners a, b, and c are chosen at random to be executed,
- a's prior that he will be executed is $2/3$.
- a asks the jailer whether b or c will be executed
- The jailer says b.

It seems that the jailer gives a no useful information about his own chances of being executed.

- a already knew that one of b or c was going to be executed

But conditioning seems to indicate that a's posterior probability of being executed should be $1/2$.

The Monty Hall Puzzle

- You're on a game show and given a choice of three doors.
  - Behind one is a car; behind the others are goats.
- You pick door 1.
- Monty Hall opens door 2, which has a goat.
- He then asks you if you still want to take what's behind door 1, or to take what's behind door 3 instead.

Should you switch?

The Second-Ace Puzzle

Alice gets two cards from a deck with four cards: A$\spadesuit$, 2$\spadesuit$, A$\heartsuit$, 2$\heartsuit$.

A$\spadesuit$ A$\heartsuit$	A$\spadesuit$ 2$\spadesuit$	A$\spadesuit$ 2$\heartsuit$
A$\heartsuit$ 2$\spadesuit$	A$\heartsuit$ 2$\heartsuit$	2$\spadesuit$ 2$\heartsuit$

Alice then tells Bob "I have an ace".

- Conditioning $\Rightarrow$ Pr(both aces | one ace) $= 1/5$.

She then says "I have the ace of spades".

- $\Pr_B(\text{both aces} \mid A\spadesuit) = 1/3$.

The situation is similar if if Alice says "I have the ace of hearts".

Puzzle: Why should finding out which particular ace it is raise the conditional probability of Alice having two aces?

Protocols

Claim 1: conditioning is always appropriate here, but you have to condition in the right space.

Claim 2: The right space has to take the *protocol* (*algorithm*, *strategy*) into account:

- a protocol is a description of each agent's actions as a function of their information.
  - **if** receive message
    then send acknowledgment

Protocols

What is the protocol in the second-ace puzzle?

- There are lots of possibilities!

Possibility 1:

1. Alice gets two cards
2. Alice tells Bob whether she has an ace
3. Alice tells Bob whether she has the ace of spades

There are six possible runs (one for each pair of cards that Alice could have gotten); the earlier analysis works:

- $\Pr_B(\text{two aces} \mid \text{one ace}) = 1/5$
- $\Pr_B(\text{two aces} \mid \text{A}\spadesuit) = 1/3$

With this protocol, we can't say "Bob would also think that the probability was 1/3 if Alice said she had the ace of hearts"

Possibility 2:

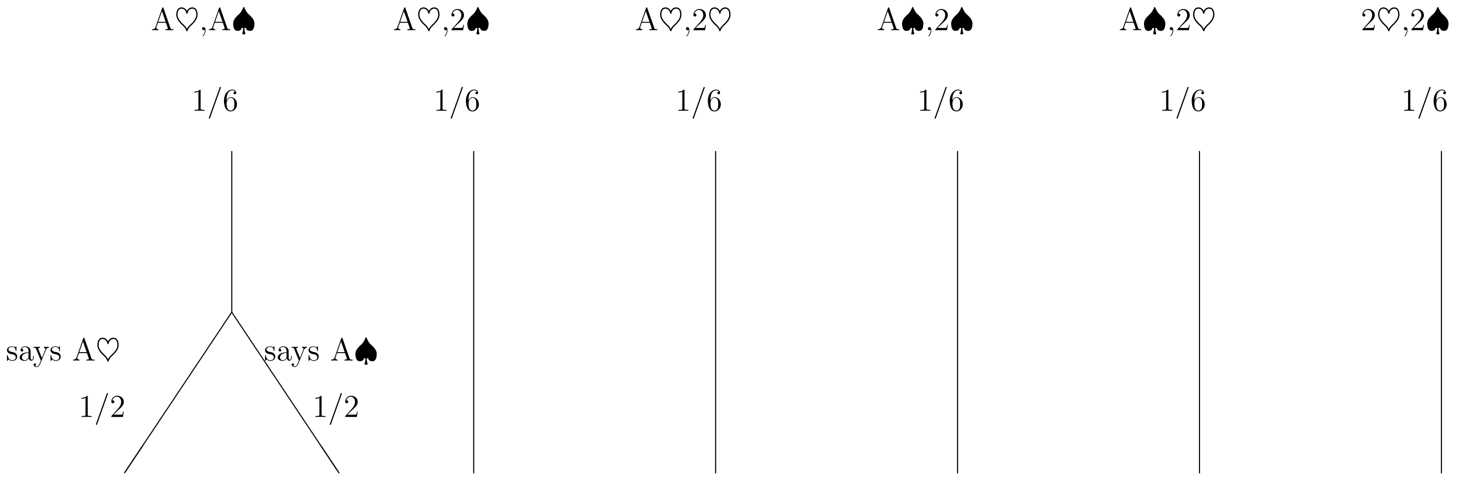
1. Alice gets two cards
2. Alice tells Bob she has an ace iff her leftmost card is an ace; otherwise she says nothing.
3. Alice tells Bob the kind of ace her leftmost card is, if it is an ace.

This protocol is not well specified. What does Alice do at step 3 if she has both aces?

Possibility 2(a):

- She chooses which ace to say at random:

Now there are seven possible runs.



- Each run has probability 1/6, except the two runs where Alice was dealt two aces, which each have probability 1/12.
- $\Pr_B(\text{two aces} \mid \text{one ace}) = 1/5$
- $\Pr_B(\text{two aces} \mid \text{A}\spadesuit) = \frac{1}{12}/(\frac{1}{6} + \frac{1}{6} + \frac{1}{12}) = 1/5$
- $\Pr_B(\text{two aces} \mid \text{A}\heartsuit) = 1/5$

More generally: Possibility 2(b):

- She says "I have the ace of spades" with probability α
  - Possibility 2(a) is a special case with $\alpha = 1/2$

Again, there are seven possible runs.

- $\Pr_B(\text{two aces} \mid \text{A}\spadesuit) = \alpha/(\alpha + 2)$
- if $\alpha = 1/2$, get $1/5$, as before
- if $\alpha = 0$, get 0
- if $\alpha = 1$, get $1/3$ (reduces to protocol 1)

Possibility 3:

1. Alice gets two cards
2. Alice tells Bob she has an ace iff her leftmost card is an ace; otherwise she says nothing.
3. Alice tells Bob the kind of ace her leftmost card is, if it is an ace.

What is the sample space in this case?

- has 12 points, not 6: the order matters
  - $(2\heartsuit, A\spadesuit)$ is not the same as $(A\spadesuit, 2\heartsuit)$

Now Pr(2 aces | Alice says she has an ace) = 1/3.

The Monty Hall puzzle

Again, what is the protocol?

1. Monty places a car behind one door and a goat behind the other two. (Assume Monty chooses at random.)
2. You choose a door.
3. Monty opens a door (with a goat behind it, other than the one you've chosen).

This protocol is not well specified.

- How does Monty choose which door to open if you choose the door with the car?
- Is this even the protocol? What if Monty does not have to open a door at Step 3?

Not to hard to show:

- If Monty necessarily opens a door at step 3, and chooses which one at random if Door 1 has the car, then switching wins with probability 2/3.

But ...

- if Monty does not have to open a door at step 3, then all bets are off!

Naive vs. Sophisticated Spaces

Working in the sophisticated space, which takes the protocol into account, gives the right answers, BUT ...

- the sophisticated space can be very large
- it is often not even clear what the sophisticated space is
  - What exactly is Alice's protocol?

When does conditioning in the naive space give the right answer?

- Hardly ever!

Formalization

Assume

- There is an underlying space W: the naive space
- Suppose, for simplicity, there is a one-round protocol, so you make a single observation. The sophisticated space S then consists of pairs (w, o) where
  - $w \in W$
  - o (the observation) is a subset of W
  - $w \in o$: the observation is always accurate.

Example: Three prisoners

- The naive space is $W = \{w_a, w_b, w_c\}$, where w_x is the world where x is not executed.
- There are two possible observations:
  - $\{w_a, w_b\}$: c is to be executed (i.e., one of a or b won't be executed)
  - $\{w_a, w_c\}$: b is to be executed

The sophisticated space consists of four elements of the form $(w_x, \{w_x, w_y\})$, where $x \neq y$ and $\{w_x, w_y\} \neq \{w_b, w_c\}$

- the jailer will not tell a that he won't be executed

Given a probability Pr on S (the sophisticated space), let $\Pr_W$ be the marginal on W:

$$\Pr_W(U) = \Pr(\{(w,o) : w \in U\}).$$

In the three-prisoners puzzle, $\Pr_W(w) = 1/3$ for all $w \in W$, but Pr is not specified.

Some notation:

- Let X_O and X_W be random variables describing the agent's observation and the actual world:

$$X_O = U \text{ is the event } \{(w,o) : o = U\}.$$

$$X_W \in U \text{ is the event } \{(w,o) : w \in U\}.$$

Question of interest:
When is conditioning on U the same as conditioning on the observation of U?

- When is $\Pr(\cdot \mid X_O = U) = \Pr(\cdot \mid X_W \in U)$?
- Equivalently, when is $\Pr(\cdot \mid X_O = U) = \Pr_W(\cdot \mid U)$?

When is conditioning on the jailer saying that b will be executed the same as conditioning on the event that b will be executed?

- The *CAR* (Conditioning at Random) condition characterizes when this happens.

The CAR Condition

Theorem: Fix a probability Pr on $\mathcal{R}$ and a set $U \subseteq W$. The following are equivalent:

(a) If $\Pr(X_O = U) > 0$, then for all $w \in U$

$$\Pr(X_W = w \mid X_O = U) = \Pr(X_W = w \mid X_W \in U).$$

(b) If $\Pr(X_W = w) > 0$ and $\Pr(X_W = w') > 0$, then

$$\Pr(X_O = U \mid X_W = w) = \Pr(X_O = U \mid X_W = w').$$

For the three-prisoners puzzle, this means that

- the probability of the jailer saying "b will be executed" must be the same if a is pardoned and if c is pardoned.
- Similarly, for "c will be executed".

This is impossible no matter what protocol the jailer uses.

- Thus, conditioning *must* give the wrong answers.

CAR also doesn't hold for Monty Hall or any of the other puzzles.

Why CAR is important

Consider drug testing:

- In a medical study to test a new drug, several patients drop out before the end of the experiment
  - for *compliers* (who don't drop out) you observe their actual response; for dropouts, you observe nothing at all.

You may be interested in the fraction of people who have a bad side effect as a result of taking the drug three times:

- You can observe the fraction of compliers who have bad side effects
- Are dropouts "missing at random"?
  - If someone drops out, you observe W.
  - Is $\Pr(X_W = w \mid X_O = W) =$
    $\Pr(X_W = w \mid X_W \in W) = \Pr(X_W = w)$?

Similar issues arise in questionnaires and polling:

- Are shoplifters really as likely as non-shoplifters to answer a question like "Have you ever shoplifted?"
- concerns of homeless under-represented in polls

Newcomb's Paradox

A highly superior being presents you with two boxes, one open and one closed:

- The open box contains a $1,000 bill
- Either $0 or $1,000,000 has just been placed in the closed box by the being.

You can take the closed box or both boxes.

- You get to keep what's in the boxes; no strings attached.

But there's a catch:

- The being can predict what humans will do
  - If he predicted you'll take both boxes, he put $0 in the second box.
  - If he predicted you'll just take the closed box, he put $1,000,000 in the second box.

The being has been right 999 of the the last 1000 times this was done.

What do you do?

The decision matrix:

- s_1: the being put \$0 in the second box
- s_2: the being put \$1,000,000 in the second box
- a_1: choose both boxes
- a_2: choose only the closed box

	s_1	s_2
a_1	\$1,000	\$1,001,000
a_2	\$0	\$1,000,000

Dominance suggests choosing a_1.

- But we've already seen that dominance is inappropriate if states and acts are not independent.

What does expected utility maximization say:

- If acts and states aren't independent, we need to compute $\Pr(s_i \mid a_j)$.
  - Suppose $\Pr(s_1 \mid a_1) = .999$ and $\Pr(s_2 \mid a_2) = .999$.
- Then take act a that maximizes

$$\Pr(s_1 \mid a)u(s_1, a) + \Pr(s_2 \mid a)u(s_2, a).$$

- That's a_2.

Is this really right?

- the money is either in the box, or it isn't ...

A More Concrete Version

The facts

- Smoking cigarettes is highly correlated with heart disease.
- Heart disease runs in families
- Heart disease more common in type A personalities

Suppose that type A personality is inherited and people with type A personalities are more likely to smoke.

- That's why smoking is correlated with heart disease.

Suppose you're a type A personality.

- Should you smoke?

Now you get a decision table similar to Newcomb's paradox.

- But the fact that Pr(heart disease | smoke) is high shouldn't deter you from smoking.

More Details

Consider two causal models:

1. Smoking causes heart disease:
   - $\Pr(\text{heart disease} \mid \text{smoke}) = .6$
   - $\Pr(\text{heart disease} \mid \neg\text{smoke}) = .2$
2. There is a gene that causes a type A personality, heart disease, and a desire to smoke.
   - $\Pr(\text{heart disease} \wedge \text{smoke} \mid \text{gene}) = .48$
   - $\Pr(\text{heart disease} \wedge \neg\text{smoke} \mid \text{gene}) = .04$
   - $\Pr(\text{smoke} \mid \text{gene}) = .8$
   - $\Pr(\text{heart disease} \wedge \text{smoke} \mid \neg\text{gene}) = .12$
   - $\Pr(\text{heart disease} \wedge \neg\text{smoke} \mid \neg\text{gene}) = .16$
   - $\Pr(\text{smoke} \mid \neg\text{gene}) = .2$
   - $\Pr(\text{gene}) = .3$

   Conclusion:

   - $\Pr(\text{heart disease} \mid \text{smoke}) = .6$
   - $\Pr(\text{heart disease} \mid \neg\text{smoke}) = .2$

Both causal models lead to the same statistics.

- Should the difference affect decisions?

Recall:

- $\Pr(\text{heart disease} \mid \text{smoke}) = .6$
- $\Pr(\text{heart disease} \mid \neg\text{smoke}) = .2$

Suppose that

- $u(\text{heart disease}) = -1,000,000$
- $u(\text{smoke}) = 1,000$

A naive use of expected utility suggests:

$$\begin{array}{rl} & EU(\text{smoke}) \\ = & -999,000 \Pr(\text{heart-disease} \mid \text{smoke}) \\ & \quad +1,000 \Pr(\neg\text{heart-disease} \mid \text{smoke}) \\ = & -999,000(.6) + 1,000(.4) \\ = & -599,800 \end{array}$$

$$\begin{array}{rl} & EU(\neg\text{smoke}) \\ = & -1,000,000 \Pr(\text{heart-disease} \mid \neg\text{smoke}) \\ = & -200,000 \end{array}$$

Conclusion: don't smoke.

- But if smoking doesn't cause heart disease (even though they're correlated) then you have nothing to lose by smoking!

Causal Decision Theory

In the previous example, we want to distinguish between the case where smoking causes heart disease and the case where they are correlated, but there is no causal relationship.

- the probabilities are the same in both cases

This is the goal of *causal decision theory*:

- Want to distinguish between $\Pr(s|a)$ and probability that *a causes s*.
  - What is the probability that smoking causes heart disease vs. probability that you get heart disease, given that you smoke.

Let $\Pr_C(s|a)$ denote the probability that a causes s.

- Causal decision theory recommends choosing the act a that maximizes

  $$\Sigma_s \Pr_C(s \mid a) u(s, a)$$

  as opposed to the act that maximizes

  $$\Sigma_s \Pr(s \mid a) u(s, a)$$

So how do you compute $\Pr_C(s \mid a)$?

- You need a good model of causality ...

Basic idea:

- include the causal model as part of the state, so state has form: (causal model, rest of state).
- put probability on causal models; the causal model tells you the probability of the rest of the state
- in the case of smoking, you need to know the probability that

In smoking example, need to know the probability that

- smoking is a cause of heart disease: α
- the probability of heart disease given that you smoke, if smoking is a cause: .6
- the probability of heart disease given that you don't smoke, if smoking is a cause: .2
- the probability that the gene is the cause: $1 - \alpha$
- the probability of heart disease if the gene is the cause (whether or not you smoke):
  $(.52 \times .3) + (.28 \times .7) = .352.$

$$
\begin{aligned}
EU(\text{smoke}) = \; & \alpha(.6(-999,000) + .4(1,000)) + \\
& (1-\alpha).352(-999,000) + .648(1,000))
\end{aligned}
$$

$$EU(\neg\text{smoke}) = \; (.2\alpha + .352(1-\alpha))(-1,000,000)$$

- If $\alpha = 1$ (smoking causes heart disease), then gets the same answer as standard decision theory: you shouldn't smoke.
- If $\alpha = 0$ (there's a gene that's a common cause for smoking and heart disease), you have nothing to lose by smoking.

So what about Newcomb?

- Choose both boxes unless you believe that choosing both boxes *causes* the second box to be empty!

A Medical Decision Problem

You want to build a system to help doctors make decisions, by maximizing expected utility.

- What are the states/acts/outcomes?

States:

- Assume a state is *characterized by n binary random variables*, $X_1, \ldots, X_n$:
  - A state is a tuple $(x_1, \ldots, x_n,\ x_i \in \{0, 1\})$.
  - The X_is describe symptoms and diseases.
    - $X_i = 0$: you haven't got it
    - $X_i = 1$: you have it
- For any one disease, relatively few symptoms may be relevant.
- But in a complete system, you need to keep track of all of them.

Acts:

- Ordering tests, performing operations, prescribing medication

Outcomes are also characterized by m random variables:

- Does patient die?
- If not, length of recovery time
- Quality of life after recovery
- Side-effects of medications

Some obvious problems:

1. Suppose $n = 100$ (certainly not unreasonable).
   - Then there are 2^{100} states
   - How do you get all the probabilities?
     - You don't have statistics for most combinations!
   - How do you even begin describe a probability distribution on 2^{100} states?
2. To compute expected utility, you have to attach a numerical utility to outcomes.
   - What the utility of dying? Living in pain for 5 years?
     - Different people have different utilities
     - Eliciting these utilities is very difficult
       - People often don't know their own utilities
     - Knowing these utilities is critical for making a decision.

Bayesian Networks

Let's focus on one problem: representing probability.

Key observation [Wright,Pearl]: many of these random variables are independent. Thinking in terms of (in)dependence

- helps structure a problem
- makes it easier to elicit information from experts

By representing the dependencies graphically, get

- a model that's simpler to think about
- (sometimes) requires far fewer numbers to represent the probability

Example

You want to reason about whether smoking causes cancer. Model consists of four random variables:

- C: "has cancer"
- SH: "exposed to second-hand smoke"
- PS: "at least one parent smokes"
- S: "smokes"

Here is a graphical representation:

Qualitative Bayesian Networks

This *qualitative Bayesian network (BN)* gives a qualitative representation of independencies.

- Whether or not a patient has cancer is directly influenced by whether he is exposed to second-hand smoke and whether he smokes.
- These random variables, in turn, are influenced by whether his parents smoke.
- Whether or not his parents smoke also influences whether he has cancer, but this influence is mediated through SH and S.
  - Once values of SH and S are known, finding out whether his parents smoke gives no additional information.
  - C is independent of PS given SH and S.

Background on Independence

Event A is independent of B given C (with respect to Pr) if

$$\Pr(A \,|\, B \cap C) = \Pr(A \,|\, C)$$

Equivalently,

$$\Pr(A \cap B \,|\, C) = \Pr(A \,|\, C) \times \Pr(B \,|\, C).$$

Random variable X is independent of Y given a set of variables $\{Z_1, \ldots, Z_k\}$ if for all values x, y, $z_1, \ldots, z_k$ of X, Y, and $Z_1, \ldots, Z_k$ respectively:

$$\begin{aligned} & \Pr(X = x \,| Y = y \cap Z_1 = z_1 \ldots \cap Z_k = z_k) \\ = \; & \Pr(X = x \,| Z_1 = z_1 \ldots \cap Z_k = z_k). \end{aligned}$$

Notation: $I_{\Pr}(X, Y \,|\, \{Z_1, \ldots, Z_k\})$

Why We Care About Independence

Our goal: to represent probability distributions compactly.

- Recall: we are interested in state spaces *characterized by random variables* $X_1, \ldots, X_n$
- States have form $(x_1, \ldots, x_n)$: $X_1 = x_1, \ldots, X_n = x_n$

Suppose $X_1, \ldots, X_5$ are independent binary variables

- Then can completely characterize a distribution by 5 numbers: $\Pr(X_i = 0)$, for $i = 1, \ldots, 5$
- If $\Pr(X_i = 0) = \alpha_1$, then $\Pr(X_i = 1) = 1 - \alpha_i$
- Because of independence,

$$\Pr(0, 1, 1, 0, 0) = \alpha_1(1 - \alpha_2)(1 - \alpha_3)\alpha_4\alpha_5.$$

- Once we know the probability of all states, can compute the probability of a set of states by adding.

More generally, if $X_1, \ldots, X_n$ are independent random variables, can describe the distribution using n numbers

- We just need $\Pr(X_i = 0)$
- n is much better than 2^n!

Situations where $X_1, \ldots, X_n$ are all independent are uninteresting

- If tests, symptoms, and diseases were all independent, we wouldn't bother doing any tests, or asking patients about their symptoms!

The intution behind Bayesian networks:

- A variable typically doesn't depend on too many other random variables
- If that's the case, we don't need too many numbers to describe the distribuiton

Qualitative Bayesian Networks: Definition

Formally, a *qualitative Bayesian network* (BN) is a *directed acyclic graph.*

- *directed* means that the edges of the graph have a direction (indicated by an arrow)
- *acyclic* means that there are no cycles (you can't follow a path back to where you started)

The nodes in the BN are labeled by random variables.

Given a node (labeled by) X in a BN G,

- the *parents* of X, denoted $\mathrm{Par}_G(X)$, are the nodes pointing to X
  - in the BN for cancer, the parents of C are S and SH; the only parent of S is PS.
- the descendants of X are all the nodes "below" X on the graph
  - the only descendants of S are S itself and C
- the *nondescendants of* X, denoted $\mathrm{NonDes}_G(X)$, are all the nodes that are *not* descendants.
  - the nondescendants of S and PS and SH

Qualitative Representation

A qualitative Bayesian network G *represents* a probability distribution Pr if, for every node X in the network

$$I_{\mathrm{Pr}}(X, \mathrm{NonDes}_G(X) \,|\, Par_G(X))$$

- X is independent of its nondescendants given its parents in G

Intuitively, G represents Pr if it captures certain (conditional) independencies of Pr.

- But why focus on these independencies?
- These are the ones that lead to a compact representation!

Topological Sort of Variables

$X_1, \ldots, X_n$ is a *topological sort* of the variables in a Bayesian network if, whenever X_i is an ancestor of X_j, then $i < j$.

Key Point: If $X_1, \ldots, X_n$ is a topological sort, then

$$\mathrm{Par}(X_i) \subseteq \{X_1, \ldots, X_{i-1}\} \subseteq \mathrm{NonDes}(X_i)$$

Thus, if G represents a probability distribution Pr and $X_1, \ldots, X_n$ are toplogically sorted, then

$$\Pr(X_i \mid \{X_1, \ldots, X_{i-1}\}) = \Pr(X_i \mid \mathrm{Par}(X_i))$$

This is because X_i is independent of its nondescendants given its parents.

The Chain Rule

From Bayes' Rule, we get

$\Pr(A_1 \cap \ldots \cap A_n) = \Pr(A_n \mid A_1 \cap \ldots \cap A_{n-1}) \times \Pr(A_1 \cap \ldots \cap A_{n-1})$

Iterating this (by induction), we get the *chain rule*:

$$
\begin{aligned}
& \Pr(A_1 \cap \ldots \cap A_n) \\
= \; & \Pr(A_n \mid A_1 \cap \ldots \cap A_{n-1}) \times \Pr(A_{n-1} \mid A_1 \cap \ldots \cap A_{n-2}) \\
& \times \cdots \times \Pr(A_2 \mid A_1) \times \Pr(A_1).
\end{aligned}
$$

In particular, if $X_1, \ldots, X_n$ are random variables, sorted topologically:

$$
\begin{aligned}
& \Pr(X_1 = x_1 \cap \ldots \cap X_n = x_n) \\
= \; & \Pr(X_n = x_n \mid X_1 = x_1 \cap \ldots \cap X_{n-1} = x_{n-1}) \times \\
& \quad \Pr(X_{n-1} = x_{n-1} \mid X_1 = x_1 \cap \ldots \cap X_{n-2} = x_{n-2}) \times \\
& \quad \ldots \times \Pr(X_2 = x_2 \mid X_1 = x_1) \times \Pr(X_1 = x_1).
\end{aligned}
$$

If G represents Pr, then

$$
\begin{aligned}
& \Pr(X_1 = x_1 \cap \ldots \cap X_n = x_n) \\
= \; & \Pr(X_n = x_n \mid \cap_{X_i \in \mathrm{Par}_G(X_n)} X_i = x_i) \times \\
& \quad \Pr(X_{n-1} = x_{n-1} \mid \cap_{X_i \in \mathrm{Par}_G(X_{n-1})} X_i = x_i) \times \\
& \quad \cdots \times \Pr(X_1 = x_1).
\end{aligned}
$$

Key point: if G represents Pr, then Pr is completely determined by conditional probabilities of the form

$$\Pr(X_j = x_j \mid \cap_{X_i \in \mathrm{Par}_G(X_j)} X_i = x_i).$$

Quantitative BNs

A *quantitative Bayesian network* G is a qualitative BN + a *conditional probability table (cpt)*:

For each node X, if $\mathrm{Par}_G(X) = \{Z_1, \ldots, Z_k\}$, for each value x of X and $z_1, \ldots, z_k$ of $Z_1, \ldots, Z_k$, gives a number $d_{x,z_1,\ldots,z_k}$. Intuitively

$$\Pr(X = x \mid Z_1 = z_1 \cap \ldots \cap Z_k = z_k) = d_{x,z_1,\ldots,z_k}.$$

A quantitative BN *quantitatively represents* Pr if it qualitatively represents Pr and

$$d_{x,z_1,\ldots,z_k} = \Pr(X = x \mid Z_1 = z_1 \cap \ldots \cap Z_k = z_k).$$

If G quantitatively represents Pr, then we can use G to compute $\Pr(E)$ for all events E. Remember:

$$\begin{array}{rl} & \Pr(X_1 = x_1 \cap \ldots \cap X_n = x_n) \\ = & \Pr(X_n = x_n \mid \cap_{X_i \in \mathrm{Par}_G(X_n)} X_i = x_i) \times \\ & \quad \Pr(X_{n-1} = x_{n-1} \mid \cap_{X_i \in \mathrm{Par}_G(X_{n-1})} X_i = x_i) \times \\ & \qquad \cdots \times \Pr(X_1 = x_1). \end{array}$$

Smoking Example Revisited

Here is a cpt for the smoking example:

S	SH	$C=1$
1	1	.6
1	0	.4
0	1	.1
0	0	.01

PS	$S=1$
1	.4
0	.2

PS	$SH=1$
1	.8
0	.3

$PS=1$
.3

- The table includes only values for $\Pr(C=1 \mid S, SH)$, $\Pr(S=1 \mid PS)$, $\Pr(SH=1 \mid PS)$, $\Pr(PS=1)$
  - $\Pr(C=0 \mid SH) = 1 - \Pr(C=1 \mid SH)$
  - Can similarly compute other entries

$$
\begin{aligned}
& \Pr(PS=0 \cap S=0 \cap SH=1 \cap C=1) \\
= \; & \Pr(C=1 \mid S=0 \cap SH=1) \times \Pr(S=0 \mid PS=0) \\
& \quad \times \Pr(SH=1 \mid PS=0) \times \Pr(PS=0) \\
= \; & .1 \times .8 \times .3 \times .7 \\
= \; & .0168
\end{aligned}
$$

What do BNs Buy Us?

If each node has $\leq k$ parents, need $\leq 2^k n$ numbers to represent the distribution.

- If k is not too large, then $2^k n \ll 2^n$.

May get a *much* smaller representation of Pr.

Other advantages:

- The information tends to be easier to elicit
  - Experts are more willing to give information about dependencies than to give numbers
- The graphical representation makes it easier to understand what's going on.

Many computational tools developed for Bayesian networks:

- Computing probability given some information
- Learning Bayesian networks

They've been used in practice:

- e.g., in Microsoft's help for printer problems.
- In modeling medical decision making

Commercial packages exist.

Can we always use BNs?

Theorem: Every probability measure Pr on space S characterized by random variables $X_1, \ldots, X_n$ can be represented by a BN.

Construction:
Given Pr, let $Y_1, \ldots, Y_n$ be any ordering of the random variables.

- For each k, find a minimal subset of $\{Y_1, \ldots, Y_{k-1}\}$, call it $\mathbf{P}_k$, such that $\mathcal{I}(\{Y_1, \ldots, Y_{k-1}\}, Y_k \mid \mathbf{P}_k)$.
- Add edges from each of the nodes in $\mathbf{P}_k$ to Y_k. Call the resulting graph G.

G qualitatively represents Pr. Use the obvious cpt to get a quantitative representation:

- Different order of variables gives (in general) a different Bayesian network representing Pr.
- Usually best to order variables causally: if Y is a possible cause of X, then Y precedes X in the order
  - This tends to give smaller Bayesian networks.

Representing Utilities

The same issues arise for utility as for probability.

Suppose that preference is represented in terms of a utility function. How hard is it to describe the function?

- If an outcome depends on n factors, each with at least k possible values, get at least k^n possible outcomes.
  - Describing the utility function can be hard!

Example: Consider buying a house. What matters?

- price of house (p)
- distance from school (ds)
- quality of local school (sq)
- distance from work (dw)
- condition of house (c)

Thus, utility is a function of these 5 parameters (and maybe others):

$$u(p, ds, sq, dw, c)$$

Suppose each parameter has three possible values.

- Describing the utility function seems to require $3^5 = 243$ numbers.

We can do better if the utility is additively separable:

$u(p, ds, sq, dw, c) = u_1(p) + u_2(ds) + u_3(sq) + u_4(dw) + u_5(c)$

There are only 15 numbers to worry about

- We compute $u_1, \ldots, u_5$ separately

With additive separability, can consider each attribute independently.

- Seems reasonable in the case of the house.

Additive separability doesn't always hold. We want

- General conditions that allow for simpler descriptions of utilities
- Graphical representations that allow for easier representation
- Techniques to make utility elicitation easier

We won't cover the first two topics (no time ...), but this is currently a hot topic in AI.

Eliciting Utilities

For medical decision making, we need to elicit patients' utilities. There are *lots* of techniques for doing so. They all have the following flavor:

- [vNM] *standard gamble* approach: Suppose o_1 is the the worst outcome, o_2 is the best outcome, and o is another outcome:
  - Find p such that $o \sim (1-p)o_1 + po_2$.
  - Note that $(1-p)o_1 + po_2$ is a lottery.
- In this way, associate with each outcome a number $p_o \in [0,1]$.
- o_1 is associated with 0
- o_2 is associated with 1
- the higher p_o, the better the outcome

How do you find p_o?

- binary search?
- *ping-pong:* (alternating between high and low values)
- *titration:* keep reducing p by small amounts until you hit p_o

The choice matters!

Other approaches

Other approaches are possible if there is an obvious linear order on outcomes.

- e.g., amount of money won

Then if o_1 is worst outcome, o_2 is best, then, for each p, find o such that

$$o \sim (1-p)o_1 + po_2.$$

- Now p is fixed, o varies; before, o was fixed, p varied
- This makes sense only if you can go continuously from o_1 to o_2
- o is the *certainty equivalent* of $(1-p)o_1 + po_2$
- This can be used to measure risk aversion

Can also fix o_1, o, and p and find o' such that

$$(1-p)o_1 + po \sim o'.$$

Lots of other variants possible.

Problems

- People's responses often not consistent
- They find it hard to answer utility elicitation questions
- They want to modify previous responses over time
- They get bored/annoyed with lots of questions
- Different elicitation methods get different answers.
- Subtle changes in problem structure, question format, or response mode can sometimes dramatically change preference responses
  - Suppose one outcome is getting $100
    - Did you win it in a lottery?
    - Get it as a gift?
    - Get it as payment for something
    - Save it in a sale?
  - This makes a big difference!
  - Gains and losses *not* treated symmetrically

My conclusion: people don't "have" utilities.

- They have "partial" utilities, and fill in the rest in response to questions.