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

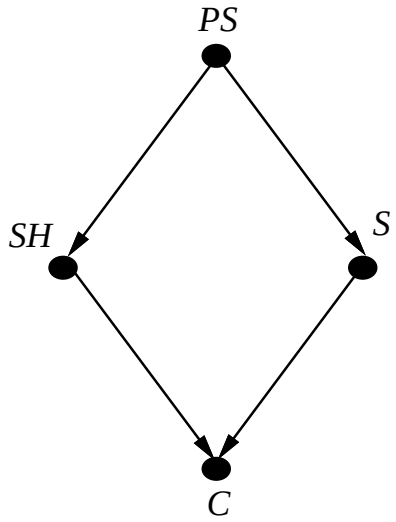
- a model that's simpler to think about
- (sometimes) requires far fewer numbers to represent the probability

Example

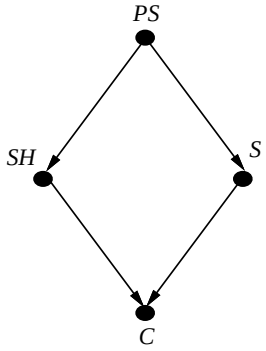
You want to reason about whether smoking causes cancer. Model consists of four random variables:

- 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

$$\Pr(A \mid B \cap C) = \Pr(A \mid C)$$

Equivalently,

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

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

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 Representation

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

$$I_{\Pr}(X, \text{NonDes}_G(X) \mid Par_G(X))$$

- $\text{NonDes}_G(X)$ consists of the nondescendants of X in the network
- 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

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

$$\text{Par}(X_i) \subseteq \{X_1, \ldots, X_{i_1}\} \subseteq \text{NonDes}(X_i)$$

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

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

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

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

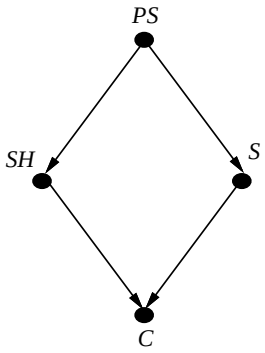
A quantitative BN *quantitatively represents* Pr if it qualitatively represents Pr and

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

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

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

Smoking Example Revisited



Here is a cpt for the smoking example:

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

PS	S
1	.4
0	.2

PS	SH
1	.8
0	.3

PS
.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.

Decision Trees

BNs have probabilities, but not utilities.

Decision trees are a first step to including both. They are trees with three kinds of nodes:

- *decision nodes*: usually denoted with a box
- *chance nodes*: usually denoted with a circle
- outcomes (consequences): usually denoted with a diamond
    - Can associate a utility with each consequence

Intuitively, the root of a decision tree represents an initial situation.

- Goal: devise an optimal plan
- For now, think of choosing a plan at time 0
    - Utilities represent your preference at time 0

Choosing a Used Car

You want to choose between two cars: $C \in \{c_1, c_2\}$:

- Quality Q is either good (q_1) or bad (q_2)
- Test T is t_0 (no test), t_1 (test c_1), or t_2 (test c_2)
  - t_1 costs \$50
  - t_2 costs \$20
  - You can't test both c_1 and c_2
- Test outcome O is $\emptyset$ if $T = t_0$ (no test); otherwise it's either pass (p) or fail (f)
- Value V depends on the kind of car and its quality
  - c_1 costs \$1,500; its market value is \$2,000
  - If it's bad, repairs will cost \$700
  - c_2 costs \$1,150; it's market value is \$1,400
  - If it's bad, repairs will cost \$150

Here's the qualitative decision tree (without probabilities). This is enough to compute the maximin plan: $t_0 + c_2$.

- Work backwards from leaves, marking each node with the best you can do there.

To compute the plan that maximizes expected utility, we need probabilities. Suppose they are:

- $\Pr(Q = q_1 \mid C = c_1) = 0.7$
  - so $\Pr(Q = q_2 \mid C = c_1) = 0.3$
- $\Pr(Q = q_1 \mid C = c_2) = 0.80$
- $\Pr(O = p \mid C = c_1, Q = q_1, T = t_1) = 0.90$
- $\Pr(O = f \mid C = c_1, Q = q_2, T = t_1) = 0.65$
  - A good car may fail the test (probability .1) and a bad car may pass (with probability .35).
- $\Pr(O = p \mid C = c_2, Q = q_1, T = t_2) = 0.75$
- $\Pr(O = p \mid C = c_2, Q = q_2, T = t_2) = 0.30$

Actually need $\Pr(Q = q_1 \mid C = c_1, T = t_1, O = p)$, etc. This can be computed using Bayes' rule:

$$\begin{aligned} & \Pr(q_1, p \mid c_1, t_1) \\ = \; & \Pr(p \mid c_1, q_1, t_1) \Pr(q_1 \mid c_1, t_1) = 0.9 \times 0.7 = 0.63 \end{aligned}$$

Similarly:

- $\Pr(q_2, p \mid c_1, t_1) = 0.35 \times 0.3 = 0.105$
- $\Pr(p \mid c_1, t_1) = 0.63 + 0.105 = 0.735$
- $\Pr(q_1 \mid c_1, t_1, p) = 0.63/0.735 = 0.86$
- ...

Now we can add probabilities to the decision diagram, and compute plan with best expected utility:

- Again, work backwards from leaves, marking each node with expected outcome.

Influence Diagrams

The decision tree formalism is very appealing but

- Even in very simple possible settings, decision trees can be huge
  - Imagine both tests were allowed in the previous example
- Lots of information is duplicated in subtrees

Influence diagrams attempt to capture all this is a simpler setting.

- Somewhat like a BN
  - missing edges represent conditional independence

The influence diagram for the car buying example:

- Decision is independent of C and Q given O

Associate with each node the set of values it can assume:

- $C \in \{c_1, c_2\}$, $T \in \{t_0, t_1, t_2\}$, $Q \in \{q_1, q_2\}$
- $O = \emptyset$ if $T = t_0$; otherwise, it's in $\{p, f\}$.
- Decision values are more complicated:
  - If $T = t_0$, then can choose c_1 or c_2
  - If $T = t_1$, then can choose c_1, c_2, [c_1 if $O = p$, c_2 if $O = f$], [c_2 if $O = p$, c_1 if $O = f$]
  - Similarly if $T = t_2$

Finally, must encode probabilities

- $\Pr(Q = q_1 \mid C = c_1) = 0.7$, etc.

Evaluating Influence Diagrams

We want to use an influence diagram to compute the choices that give the maximum expected utility.

- The naive way: convert the influence diagram to a decision tree
  - This loses all the advantages of influence diagrams!
- There are smarter algorithms that compute the value of each node in the influence diagram [Shachter 1986]
  - The running time can still be exponential in the number of nodes, although the space is linear in the number of nodes

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.