Bayesian Networks

Two possible views of Bayesian Networks:

- A representation of a joint probability distribution
- A collection of conditional independence statements

That is, a Bayesian network will allow us to concisely represent a full joint probability distribution. It explicitly represents dependencies among random variables.

More formally,

A Bayesian network is a directed acyclic graph in which each node is annotated with probability information.

If there is an arrow (directed edge) from X to Y, then we say that X is a parent of Y.

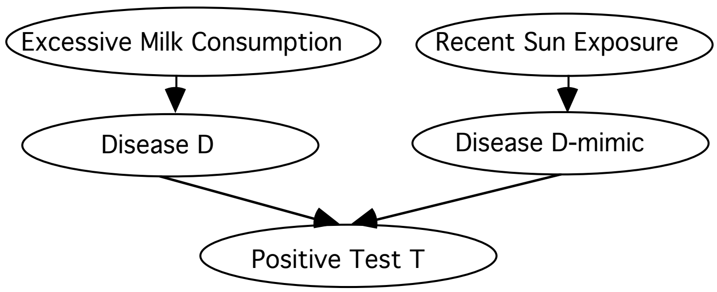
X_i has conditional probability distribution $\mathbf{P}(X_i \mid \text{Parents}(X_i))$.
This quantifies the effect of parents on a node. Intuitively, an arrow from X to Y means that X has a direct influence on Y.

To construct a Bayesian network, we need to:

1. Set out the topology
2. Give conditional probability tables at the nodes

This is all that's needed to implicitly give the full joint probability distribution.

An example of a Bayesian network

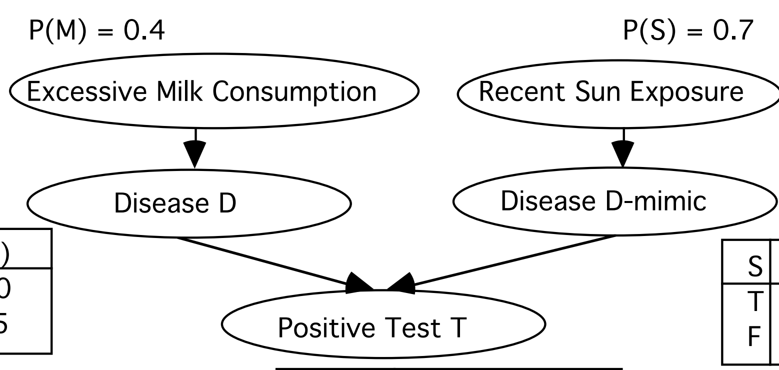


Once we've decided upon:

- Relevant variables,
- Topology,

We need to specify:
Conditional probabilities for nodes that participate in direct dependencies – i.e., we need to give a **conditional probability table** at each node. We will calculate any other probability values from these.

Conditional probability tables



M	P(D)
T	.90
F	.05

S	P(Dm)
T	.70
F	.01

D	Dm	P(T)	P(not T)
T	T	.95	.05
T	F	.80	.20
F	T	.10	.90
F	F	.05	.95

Each node of the Bayesian network has a conditional probability table (CPT) specified for it.

- Each row contains the conditional probability of each node value for a conditioning case.
- A conditioning case is a possible combination of values for the parent nodes.
- Each row must sum to 1.
- If the node represents a Boolean random variable, the conditional probabilities given in each row can be for one variable value only, since the other can easily be derived from it.

CPTs and the joint probability distribution

Let $P(X_1=x_1 \wedge X_2=x_2 \wedge \ldots \wedge X_n=x_n) = P(x_1, x_2, \ldots, x_n)$ be an entry in the joint probability distribution.

$$P(x_1, x_2, \ldots, x_n) =$$

$$\prod_{i=1}^{n} P(x_i \mid Parents(X_i))$$

That is, an entry in the joint probability distribution is a product of elements of the conditional probability tables.

Example. $P(\neg m, \neg s, \neg d, \neg d_m, t)$

$= P(\neg m)P(\neg s)P(\neg d|\neg m)P(\neg d_m|\neg m)P(t|\neg d \wedge \neg d_m)$
$= (0.6)(0.3)(0.95)(0.99)(0.05)$
$= 8.4 \times 10^{-3}$

So a Bayesian network allows us to represent the full joint probability distribution concisely and, therefore, more efficiently.

Constructing a Bayesian network

First, consider the following rule.

The Chain Rule:
$P(x_1, x_2, \ldots, x_n) = P(x_n|x_{n-1}, \ldots, x_1)\, P(x_{n-1}|x_{n-2}, \ldots, x_1) \ldots P(x_2|x_1)P(x_1)$

So the specification of the full joint probability distribution is equivalent to the assertion that for every X_i in the network

$P(X_i|X_{i-1} \ldots X_1) = P(X_i|Parents(X_i))$

Intuitively, this says that a Bayesian network is correct only if each node is conditionally independent of its predecessors in the node ordering, given its parents. I.e., the parents of node X_i should contain all those nodes in $X_1, \ldots, X_{i-1}$ that directly influence X_i.

So the general procedure for designing a Bayesian network is as follows:

Add "root **causes**" first; then add the variables they influence, and so on.
Leaves are nodes that have no direct **causal** influence on anything else.

Note that there is an important difference between diagnostic and causal models. Causal models tend to have fewer links and the numbers are easier to come up with.

Steps to follow in constructing a Bayesian network

1. Choose a set of relevant random variables that describe the domain.
2. Choose an ordering for the variables. Order the variables so that "root causes" are first, then the variables they influence, and so on.
3. For i = 1 to n, where n is the number of random variables:
    a. Add a node to the network for variable X_i.
    b. Set Parents(X_i) to a minimal set of nodes already in the network such that the conditional independence property is satisfied.
    c. Define the conditional probability table for X_i.

Filling in conditional probability tables

It can be difficult to fill in CPTs, but there are techniques that can help to make the process easier. The idea here is to recognize in our Bayesian network some standard pattern.

Example 1. Deterministic nodes: A deterministic node has its value specified exactly by the values of its parents, with no uncertainty.

Example 2. Noisy logical relationships, such as "noisy or": A noisy-or is a generalization of a logical or. For example, consider the following

$$\text{cavity} \vee \text{abscess} \Rightarrow \text{mouth-pain}$$

in a noisy-or, the causal relationship may be inhibited (i.e., you can have a cavity but have no mouth pain)

This model assumes that
1. all causes are listed;
2. the inhibition of each parent is independent of the inhibition of any other parent.

So mouth-pain is false iff all of its true parents are inhibited. The probability of this is the product of the inhibition probabilities for each parent.

Example.
$P(\neg\text{mouth-pain}|\text{cavity}, \neg\text{abscess}) = 0.6$
$P(\neg\text{mouth-pain}|\neg\text{cavity}, \text{abscess}) = 0.2$

This is enough information to compute the values in the CPT:

		¬mouth-pain	mouth-pain
cavity	abscess	0.6 x 0.2 = 0.12	0.88
cavity	¬abscess	0.6	0.4
l¬cavity	abscess	0.2	0.8
l¬cavity	¬abscess	1.0	0.0

Exact inference in Bayesian networks

<u>Given</u>: an observed event (i.e., an assignment of values to evidence variables)
<u>Compute</u>: posterior probability distribution for a set of query variables

Let
X be a query variable
$\mathbf{E} = E_1, \ldots, E_m$ be a set of evidence variables
e is a particular observed event
$\mathbf{Y} = Y_1, \ldots, Y_n$ are non-evidence variables (hidden variables)

The complete set of variables is:

$$\mathbf{X} = \{X\} \cup \mathbf{E} \cup \mathbf{Y}$$

A typical query asks for

$$\mathbf{P}(X|\mathbf{e})$$

So for each value of X we compute
$P(x|\mathbf{e}) = P(x,\mathbf{e})/P(\mathbf{e})$

We do this by computing $P(x,\mathbf{e}) =$
$\Sigma_{\square} \ldots \Sigma_{\square} P(x, \mathbf{e}, \mathbf{Y})$, where each summation is done over all the values of $Y_i \in \mathbf{Y}$.

Example. Consider the Bayesian network given earlier for Disease D and the Disease D-mimic. Say that we know that a patient has had a positive test result. The patient also reports that they drink a lot of milk. What's the probability that the patient has Disease D?

[Details provided in class.]

What good are Bayesian networks?

Bayesian nets are generally more compact than the full joint probability distribution:

If we have n Boolean random variables and the maximum number of parents of a node in our network is k, then this is how the Bayesian net compares with the full joint probability distribution (with respect to size)

Bayesian net – $n2^k$
Joint – 2^n