Lecture 3: Bayesian Networks

- An example
- DAGs as representations of independence
- I-maps

Recall from last time: Conditional independence

Two variables X and Y are conditionally independent given Z if and only if

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z), \forall x, y, z$$

We denote this by $I(X, Y|Z)$.

In this lecture we discuss the use of graphical representations to capture independence properties.

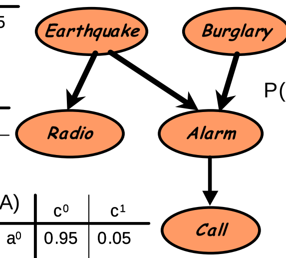
A Bayes net example

P(E)

e^0	e^1
0.995	0.005

P(B)

b^0	b^1
0.99	0.01



P(R|E)

	r^0	r^1
e^0	0.99999	0.00001
e^1	0.65	0.35

P(A|B,E)

	a^0	a^1
b^0,e^0	0.999	0.001
b^0,e^1	0.7	0.3
b^1,e^0	0.2	0.8
b^1,e^1	0.05	0.95

P(C|A)

	c^0	c^1
a^0	0.95	0.05
a^1	0.3	0.7

Using a Bayes net for reasoning (1)

Computing any entry in the joint probability table is easy:

$P(b = 1)P(e = 0)P(a = 1|b = 1, e = 0)P(c = 1|a = 1)P(r = 0|e = 0) \approx 0.0056$

What is the probability that a neighbor calls?

$$P(c = 1) = \sum_{e,b,r,a} P(c = 1, e, b, r, a) = 0.0568$$

What is the probability of a call in case of a burglary?

$$P(c = 1|b = 1) = \frac{P(c = 1, b = 1)}{P(b = 1)} = \frac{\sum_{e,r,a} P(c = 1, b = 1, e, r, a)}{\sum_{c,e,r,a} P(c, b = 1, e, r, a)}$$

This is **causal reasoning** or **prediction**

Using a Bayes net for reasoning (2)

Suppose we got a call. What is the probability of a burglary?

$$P(b=1|c=1) = \frac{P(c=1|b=1)P(b=1)}{P(c=1)} = 0.1034$$

What is the probability of an earthquake?

$$P(e=1|c=1) = \frac{P(e=1|b=1)P(b=1)}{P(c=1)} = 0.02688$$

This is **evidential reasoning** or **explanation**

What happens to the probabilities if the radio announces an earthquake?

$P(e=1|c=1, r=1) = 0.9993$ and $P(b=1|e=1, r=1) = 0.02688$

This is called **explaining away**. It is a special case of **inter-causal reasoning**

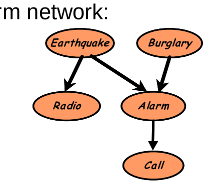
Using DAGs to represent independencies

- Graphs have been proposed as models of human memory and reasoning on many occasions (e.g. semantic nets, inference networks, conceptual dependencies)
- There are many efficient algorithms that work with graphs, and efficient data structures

Markov assumption

Given a graph G, what sort of independence assumptions does it imply?

E.g. Consider the alarm network:



We have $I(E,B)$, $I(R,\{B,A,C\}|E)$ and $I(C,\{E,B,R\}|A)$.

How about node A?

In general a variable is independent of its *non-descendents* given its parents.

Bayesian network structure

A Bayesian network structure is a directed acyclic graph (DAG) G whose nodes represent random variables $X_1, \ldots, X_n$. G encodes the following conditional independence assumptions:

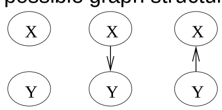
$$I(X_i, Nondescendents(X_i)|Parents(X_i)), \forall i = 1, \ldots n$$

We denote this set of independence assumption by $Markov(G)$.

I-Maps

A Bayesian network structure is an **I-map (independence map)** of a distribution P if P satisfies the independence assumptions $Markov(G)$.

Example: Consider all possible graph structures over 3 variables:



X	Y	$P_1(X,Y)$
x=0	y=0	0.08
x=0	y=1	0.32
x=1	y=0	0.32
x=1	y=1	0.48

X	Y	$P_2(X,Y)$
x=0	y=0	0.4
x=0	y=1	0.3
x=1	y=0	0.2
x=1	y=1	0.1

Which graph is an I-map for P_1? How about P_2?

Factorization

Given that G is an I-map for P, can we simplify the representation of P?

Example: If G contains two unconnected vertices X and Y, and G is an I-map for P, then we have $I(X,Y)$ and we can write $P(X,Y) = P(X)P(Y)$.

Let G be a Bayesian network structure over variables $X_1, \ldots, X_n$. We say that a distribution P **factorizes according to** G if P can be expressed as a product:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$$

The individual factors $P(X_i|Parents(X_i))$ are called **local probabilistic models** or **conditional probability distributions**

Bayesian network definition

A Bayesian network is a Bayesian network structure G together with a distribution P that factorizes over G, where P is specified as the set of conditional probability distributions associated with G's nodes.

Example: The Alarm network.

Factorization theorem

If G is an I-map of P, then P **factorizes according to** G:

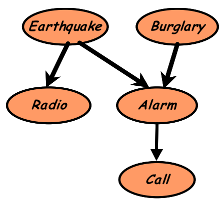
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$$

Proof: By the chain rule, $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$. Without loss of generality, we can order the variables X_i according to G. From this assumption, $Parents(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$. This means that $\{X_1, \ldots, X_{i-1}\} = Parents(X_i) \cup Z$, where $Z \subseteq Nondescendents(X_i)$. Since G is an I-map, we have $I(X_i, Nondescendents(X_i)|Parents(X_i))$, so:

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

and the conclusion follows.

Factorization example



The factorization theorem allows us to represent $P(C, A, R, E, B)$ as:

$$P(C, A, R, E, B) = P(B)P(E)P(R|E)P(A|E, B)P(C|A)$$

instead of:

$$P(C, A, R, E, B) = P(B)P(E|B)P(R|E, B)P(A|E, B, R)P(C|A, E, B, R)$$

Complexity of factorized representations

- If $|Parents(X_i)| \leq k, \forall i$, and we have binary variables, then every conditional probability distribution will require $\leq 2^k$ numbers to specify
- The whole joint distribution can then be specified with $\leq n \cdot 2^k$ numbers, instead of 2^n
- The savings are big if the graph is sparse ($k \ll n$).

Converse of the factorization theorem

If $P(X_1, \dots, X_n) = \prod_i P(X_i|Parents(X_i))$ the G is an I-map of P.

Proof: will be on the next homework

Minimal I-maps

- The fact that a DAG G is an I-map for P might not be very useful.
  E.g. Complete DAGs (where all arcs that do not create a cycle are present) are I-maps for *any distribution* (because they do not imply any independencies).
- A DAG G is a **minimal I-map** of P if G:
  1. G is an I-map of P
  2. If $G' \subseteq G$ then G' is not an I-map for P

Constructing minimal I-maps

The factorization theorem suggests an algorithm:

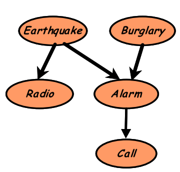
1. Fix an ordering of the variables: $X_1, \ldots, X_n$
2. For each X_i, select $Parents(X_i)$ to be the minimal subset of $\{X_1, \ldots, X_{i-1}\}$ such that $I(X_i, \{X_1, \ldots, X_{i-1}\} - Parents(X_i) | Parents(X_i))$.

This will yield a minimal I-map

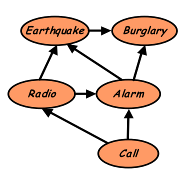
Non-uniqueness of the minimal I-map

- Unfortunately, a distribution can have **many minimal I-maps**, depending on the variable ordering we choose!
- The initial choice of variable ordering can have a big impact on the complexity of the minimal I-map:

  Example:



Ordering: E, B, A, R, C



Ordering:C, R, A, E, B

- A good heuristic is to use causality in order to generate an ordering.