CSE 150 - 1/13/10

Motivation

- Joint distribution $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ involves $O(2^n)$ numbers for n binary random variables
- More compact representations
- More efficient algorithms for inference?

Alarm Example

- Binary random variables $\{0, 1\}$

B = burglary J = John calls
E = earthquake M = Mary calls
A = alarm

* Joint distribution

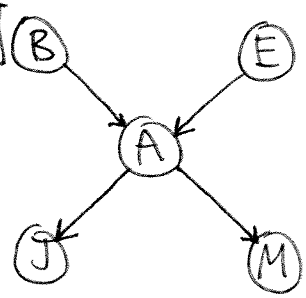
$$P(B, E, A, J, M) = P(B)\,P(E|B)\,P(A|B,E)\,P(J|B,E,A)\,P(M|B,E,A,J)$$

* Conditional Independence

$$P(B, E, A, J, M) = P(B)\,P(E)\,P(A|B,E)\,P(J|A)\,P(M|A)$$

* Directed Acyclic Graph (DAG)

$P(B=1) = 0.001$



$P(E=1) = 0.002$

B	E	$P(A=1 \mid B,E)$
0	1	0.29
1	0	0.94
0	0	0.001
1	1	0.95

A	$P(J=1 \mid A)$
0	0.05
1	0.9

A	$P(M=1 \mid A)$
0	0.01
1	0.7

* Conditional Probability Tables (CPTs)

* Joint Probabilities

ex: $P(B=1, E=0, A=1, J=1, M=1)$

$= P(B=1)\,P(E=0)\,P(A=1 \mid B=1, E=0)\,P(J=1 \mid A=1)\,P(M=1 \mid A=1)$

$= (0.001)(1-0.002) \ldots$

* Any query can be answered from joint distribution

Ex: $P(B=1, E=0 \mid M=1)$

From Product Rule: $P(B=1, E=0 \mid M=1) = \dfrac{P(B=1, E=0, M=1)}{P(M=1)}$

From Marginalization:

$$P(B=1, E=0, M=1) = \sum_{a,j \in \{0,1\}} P(B=1, E=0, M=1, A=a, J=j)$$

$$P(M=1) = \sum_{b,j,a,e \in \{0,1\}} P(M=1, B=b, E=e, A=a, J=j)$$

More efficient Algorithms? Yes.

Exploit Structure of DAG.

$\boxed{\text{Belief Network (BN)}}$

A BN is a DAG in which

(i) nodes represent random variables

(ii) edges represent conditional dependencies

(iii) CPTs describe how each node depends on its parents

* Conditional Independence

Generally true from the product rule that

$$P(X_1, X_2, \ldots, X_n) = P(X_1)\, P(X_2 \mid X_1) \ldots P(X_n \mid X_1, X_2, \ldots, X_{n-1})$$
$$= \prod_{i=1}^{n} P(X_i \mid X_1, X_2, \ldots, X_{i-1}) \qquad (*)$$

In a given domain, suppose that

$$P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid \text{parents}(X_i))$$

where parents(X_i) is some subset of $\{X_1, \ldots, X_{i-1}\}$

Big Idea: represent dependence relations by a DAG

* Constructing a BN

(1) choose random varibles

(2) choose ordering

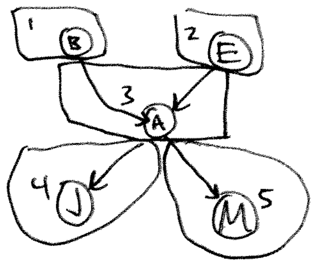
(3) while there are variables left:

(a) add node X_i

(b) set the parents of X_i to the minimal subset satisfying (*)

(c) define CPT $P(X_i \mid pa(X_i))$ parent configuration

Ex: $\{B, E, A, J, M\}$



* advantages:

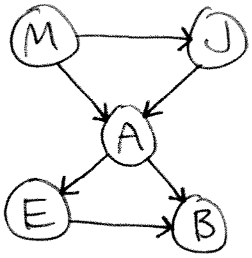
- complete, compact, consistant representation of joint distribution

Ex: for binary variables, if k = max # parents (in-degree) of graph, then $O(n2^k)$ numbers will be used versus $O(2^n)$ to represent joint distribution.

- clean separation of qualitative vs quantitative knowledge
  DAG encodes conditional independencies
  CPTs encode numerical influences

* Node ordering

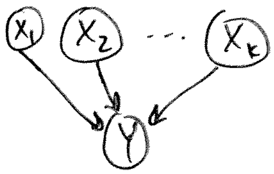
- Best order is to add "root causes", the the variables they influence, and so on...
- wrong order: $\{M, J, A, B, E\}$
  gives graph:



this DAG has two extra edges vs. original DAG.

- from "misordered" graph:
  - conditional independencies in world not obvious
  - more numbers in CPTs to specify the same joint distribution
  - less natural, more difficult to assess CPTs or learn CPTs from data

* Representing CPTs



for simplicity consider
$X_i \in \{0,1\}$
$Y \in \{0,1\}$

How to represent CPT $P(Y=1|X_1, X_2, \ldots, X_k)$?

(i) lookup table: $O(2^k)$ can store arbitrary CPT

(ii) "deterministic" node

"AND" $P(Y=1|X_1,\ldots,X_k) = \prod_{i=1}^{k} X_i$

"OR" $P(Y=0|X_1,\ldots,X_k) = \prod_{i=1}^{k} (1-X_i)$

(iii) noisy-or node

use k numbers $p_i \in [0,1]$ to parameterize $O(2^k)$ elements of CPT.

$$P(Y=0|X_1,\ldots,X_k) = \prod_{i=1}^{k} (1-p_i)^{X_i} \quad \text{for } X_i \in \{0,1\}$$

$$P(Y=1|X_1,\ldots,X_k) = 1 - \prod_{i=1}^{k} (1-p_i)^{X_i}$$

Why called "noisy-or"?

Look at $P(Y=1|X_1=X_2=\ldots=X_k=0) = 1 - \prod_{i=1}^{k} (1-p_i)^0 = 0$

Look at $P(Y=1 \mid X_1, \ldots, X_k)$ when one and only one parent $X_i = 1$; rest are zero.

$$P(Y=1 \mid X_1 = \cdots = X_{i-1} = 0, X_i = 1, X_{i+1} = \cdots = X_k = 0)$$

$$= P(Y=1 \mid X_i = 1, X_{j \neq i} = 0)$$

$$= 1 - (1-p_i)^1 \prod_{j \neq i}^{k} (1-p_j)^0 = p_i$$

Intuitively, p_i represents the probability that $X_i = 1$ by itself triggers $Y=1$.