Jan. 16. 2014.
CSE 150.

Review.

* Alarm example.

  Binary random variables

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

* Probabilities captures commonsense patterns of reasoning.

  1) Explaining away. $P(B=1|A=1) = P(A=1)$.
     but $P(B=1|A=1, E=1) < P(B=1|A=1)$
  2) Conflicting "symptoms": $P(A=1|J=1) > P(A=1)$.
     but $P(A=1|J=1, M=0) < P(A=1)$
  3) Interveving events (later today)

Today – from probabilities to graphs.

Motivation.

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

Alarm Example.

* Joint distribution.

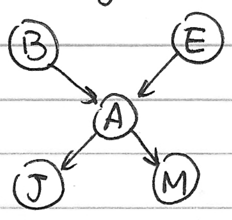
$$P(B, E, A, J, M) = P(B) \cdot P(E|B) \cdot P(A|B, E) \cdot P(J|B, E, A) \cdot P(M|B, E, A, J)$$

* Conditional independence.

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

Directed acyclic graph (DAG).

$P(B=1) = 0.001$



$P(E=1) = 0.002$

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

ex: $P(\cancel{B=1}\; B=1, E=0, A=1, J=1, M=0)$

$= P(B=1) \cdot P(E=0) \cdot P(A=1|B=1, E=0) \cdot P(J=1|A=1) \cdot P(M=0|A=1).$

$= 0.001 \times (1-0.002) \times 0.94 \times 0.9 \times (1-0.7).$

* Any query can be answered from joint distribution.

Ex. reasoning about intervening events.

compare. $P(A=1)$. $= 0.00252$

$P(A=1|J=1) = 0.0435 \uparrow$.

$P(A=1|J=1, B=1) = ?$

Brute force calculation.

- From product rule: $P(A=1|B=1, J=1) = \dfrac{P(A=1, B=1, J=1)}{P(B=1, J=1)}$

- From marginalization:

numerator : $P(A=1, B=1, J=1) = \sum_{e,m} P(A=1, B=1, J=1, E=e, M=m)$

denominator : $P(B=1, J=1) = \sum_{a,e,m} P(B=1, J=1, A=a, E=e, M=m)$

More efficient algorithm.

Exploit structure of DAG.

"conditionalized. Bayes rule"

$$P(A=1|B=1, J=1) = \frac{P(J=1|A=1, B=1)\cdot P(A=1|B=1)}{P(J=1|B=1)}$$

conditional independence

$$= \frac{\overset{0.9}{P(J=1|A=1)}\cdot P(A=1|B=1)}{P(J=1|B=1)}$$ from last lecture.

"conditionalized marginalization"

Denominator: $P(J=1|B=1) = \sum_a P(J=1, A=a|B=1)$

$= \sum_a P(A=a|B=1). P(J=1|A=a, B=1)$. "product rule"

$= \sum_a P(A=a|B=1)\cdot P(J=1|A=a)$

computed last time. CPT.

$= 0.849$.

Now we have. $P(A=1|J=1, B=1) = 0.9965 \uparrow\uparrow$. (see "compare" at the top of this page)

Belief Network. (BN).

A BN is a DAG.

(1) nodes represent random variables

(2) edges represent conditional dependencies.

(3) CPTs describe how each node depends on parents.

* Conditional Independence.

Generally true that. *product rule.*

$$P(X_1, X_2 \cdots X_n) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdots P(X_n|X_1, X_2 \cdots X_{n-1})$$
$$= \prod_{i=1}^{n} P(X_i | X_1, X_2 \cdots X_{i-1}) \qquad (\star)$$

In a given domain, suppose that.

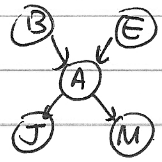
$$P(X_1, X_2 \cdots X_n) = \prod_i P(X_i | \text{parents}(X_i)) \qquad (\star\star)$$

where parents(X_i) is some subset of $\{X_1, X_2 \cdots X_{i-1}\}$

Big idea: represent dependence relations by a DAG.

- How to construct a BN?
  1. choose random variables.
  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. $(\star\star)$
     (c). define CPT $P(X_i | pa(X_i))$

Ex. alarm example.
{B, E, A, J, M}.



*, advantages.

1. complete, compact, consistent. representation. of joint distribution.

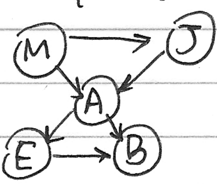
Ex: for binary variables, if k = max # of parents of graph, then.
$O(n \cdot 2^k)$ numbers will appear in CPTs v.s.
$O(2^n)$ for joint distribution.

2. clean seperation of qualitative and quantatative knowledge.
DAG encodes conditional independence.
CPTs encode numerical influences.

* Node Ordering.

- Best order is to add "root" causes. then the variables they influence, and so on ...

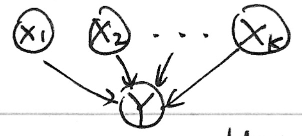
Ex: wrong ordering.
{M, J, A, E, B}.



- this DAG has two more extra edges. compared to DAG above

- from misordered graph
  - conditional independences in world not obvious.
  - more numbers (larger CPTs) to specify the same joint distribution.
  - less natural (more difficult) to assess the CPTs or learn CPT from data.

* Representing CPTs.



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

How to represent $P(Y=1 \mid X_1, X_2 \cdots X_k)$?