Artificial Intelligence

Lesson 11
(From Russell & Norvig)

Conditional probability

- Conditional or posterior probabilities
  e.g., P(*cavity* | *toothache*) = 0.8
  i.e., given that *toothache* is all I know
- Notation for conditional distributions:
  P(*Cavity* | *Toothache*) = 2-element vector of 2-element vectors)
- If we know more, e.g., *cavity* is also given, then we have
  P(*cavity* | *toothache,cavity*) = 1
- New evidence may be irrelevant, allowing simplification, e.g.,
  P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8
- This kind of inference, sanctioned by domain knowledge, is crucial

Inference by enumeration

- Start with the joint probability distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

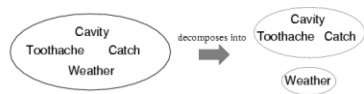
- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108+0.012+0.016+0.064} = 0.4$$

Independence

- *A* and *B* are independent iff
  P(*A*|*B*) = **P**(*A*) or **P**(*B*|*A*) = **P**(*B*) or **P**(A, B) = **P**(*A*) **P**(*B*)



P(*Toothache, Catch, Cavity, Weather*)
= **P**(*Toothache, Catch, Cavity*) **P**(*Weather*)

- 32 entries reduced to 12; for *n* independent biased coins, $O(2^n) \rightarrow O(n)$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  (1) **P**(*catch* | *toothache, cavity*) = **P**(*catch* | *cavity*)
- The same independence holds if I haven't got a cavity:
  (2) **P**(*catch* | *toothache,¬cavity*) = **P**(*catch* | *¬cavity*)
- *Catch* is conditionally independent of *Toothache* given *Cavity*:
  P(*Catch* | *Toothache,Cavity*) = **P**(*Catch* | *Cavity*)
- Equivalent statements:
  P(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)

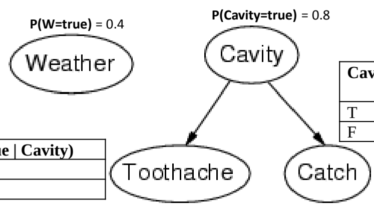
  P(*Toothache, Catch* | *Cavity*) = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*)

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- It describes how variables interact locally
- Local interactions chain together to give global, indirect interactions
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:
    P (X_i | Parents (X_i))- conditional probability table (CPT)

Example 1

- Topology of network encodes conditional independence assertions:



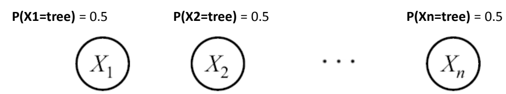
Cavity	P(T=true | Cavity)
T	.8
F	.4

Cavity	P(C=true | Cavity)
T	.9
F	.05

- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*
- It is usually easy for a domain expert to decide what direct influences exist

Example 2

- N independent coin flips :



- No interactions between variables: absolute independence
- Does every Bayes Net can represent every full joint?
- No. For example, Only distributions whose variables are absolutely independent can be represented by a Bayes' net with no arcs.

Calculation of Joint Probability

- Given its parents, each node is conditionally independent of everything except its descendants
- Thus,

$$P(x_1 \wedge x_2 \wedge \ldots \wedge x_n) = \prod_{i=1,\ldots,n} P(x_i|\text{parents}(X_i))$$

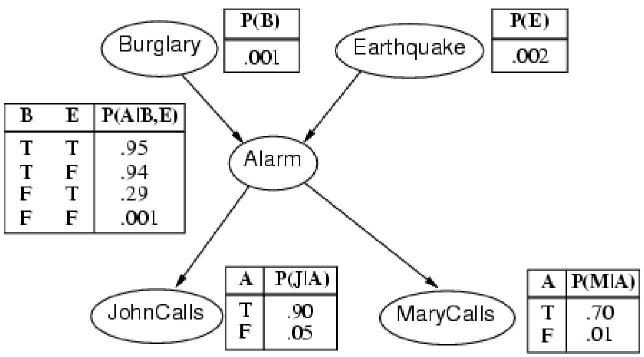
→ full joint distribution table

- Every BN over a domain implicitly represents some joint distribution over that domain

Example 3

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



Answering queries

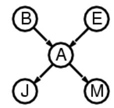
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
  - $P(b|j,\neg m) = P(b,j,\neg m)/P(j,\neg m)$
  - $P(b,j\ \neg m) = P(b,e,a,j,\neg m) + P(b,\neg e,a,j,\neg m) + P(b,e,\neg a,j,\neg m) + P(b,\neg e,\neg a,j,\neg m) =$
    $P(b)P(e)P(a|b,e)P(j|a)P(\neg m|a) +$
    $P(b)P(e)P(\neg a|b,e)P(j|\neg a)P(\neg m|\neg a) +$
    $P(b)P(\neg e)P(a|b, \neg e)P(j|a)P(\neg m|a) +$
    $P(b)P(\neg e)P(\neg a|b, \neg e)P(j|\neg a)P(\neg m|\neg a)$
  - Do the same to calculate $P(\neg b,j\ \neg m)$ and normalize
  - Worst case, for a network with n Boolean variables, $O(n2^n)$.

Laziness and Ignorance

- The probabilities actually summarize a potentially infinite set of circumstances in which the alarm might fail to go off
  - high humidity
  - power failure
  - dead battery
  - cut wires
  - a dead mouse stuck inside the bell
- John or Mary might fail to call and report it
  - out to lunch
  - on vacation
  - temporarily deaf
  - passing helicopter

Compactness

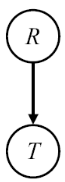
- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values
- Each row requires one number p for X_i = *true* (the number for X_i = *false* is just $1-p$)



- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, 1 + 1 + 4 + 2 + 2 = 10 numbers (vs. 2^5-1 = 31)
- We utilize the property of **locally structured system**

Casualty?

- **R**ain causes **T**raffic
- Let's build the joint:



$P(R)$

r	1/4
¬r	3/4

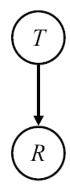
$P(T|R)$

r	t	3/4
	¬t	1/4
¬r	t	1/2
	¬t	1/2

$P(T,R)$

r	t	3/16
r	¬t	1/16
¬r	t	6/16
¬r	¬t	6/16

Reverse Casualty?



$P(T)$

t	9/16
¬t	7/16

$P(R|T)$

t	r	1/3
	¬r	2/3
¬t	r	1/7
	¬r	6/7

$P(T,R)$

r	t	3/16
r	¬t	1/16
¬r	t	6/16
¬r	¬t	6/16

Casualty?

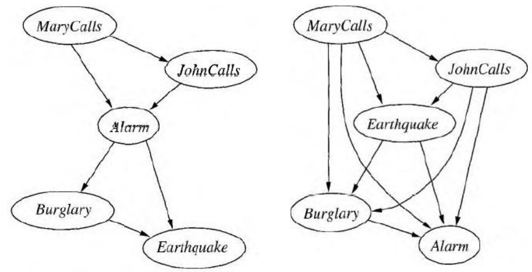
- What do the arrows really mean?
- Topology may happen to encode causal structure
- Topology really encodes conditional independencies

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain
  - E.g. consider the variables Traffic and RoofDrips
  - End up with arrows that reflect correlation, not causation

Example 2, Again

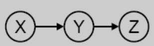
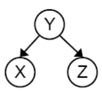
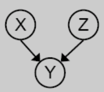
Consider the following 2 orders for insertion:

- (a) MaryCalls, JohnCalls, Alarm, Burglary, Earthquake
  - Since, P(Burglary|Alarm, JohnCalls, MaryCalls) = P(Burglary|Alarm)
- (b) Mary Calls, JohnCalls, Earthquake, Burglary, Alarm.



(a) (b)

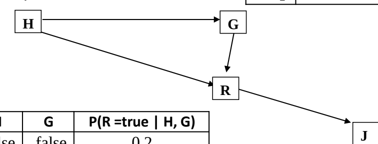
Connection Types

Name	Diagram	X ind. Z?	X ind. Z, given Y?
Casual chain		Not necessarily	Yes
Common Cause		No	Yes
Common Effect		Yes	No

Test Question

H	P(G=true | H)
T	.4
F	.8

P(H=true) = 0.1



H	G	P(R =true | H, G)
false	false	0.2
false	true	0.9
true	false	0.3
true	true	0.8

R	P(J=true | R)
false	0.2
true	0.7

H - Hardworking
G - Good Grader
R - Excellent Recommendation
J - Landed a good Job

What can be inferred?

i: $P(H,G)=P(H)\cdot P(G)$ ✗

ii $P(J|R,H)=P(J|R)$ ✓

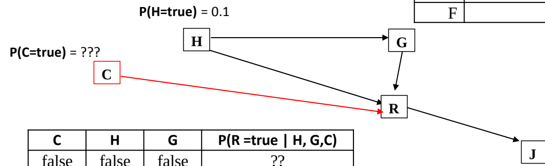
iii $P(J)\neq P(J|H)$ ✗

Q: What is the value of P(H,G,¬R,¬J)?
A: P(H,G, ¬R, ¬J) = P(H)*P(G|H)*P(¬R|H,G)*P(¬J|H,G, ¬R) = P(H)*P(G|H)*P(¬R|H,G)*P(¬J| ¬R) = 0.1 * 0.4 * 0.2 * 0.8 = 0.0064

Q: What if we want to add another parameter, C= Has The Right Connections?

Answer

H	P(G=true | H)
T	.4
F	.8

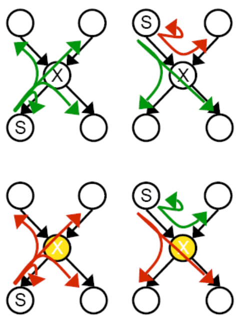


C	H	G	P(R =true | H, G,C)
false	false	false	??
false	false	true	??
false	true	false	??
false	true	true	??
true	false	false	??
true	false	true	??
true	true	false	??
true	true	true	??

R	P(J=true | R)
false	0.2
true	0.7

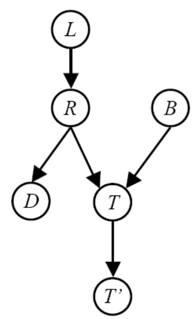
Reachability (the Bayes Ball)

- Shade evidence nodes
- Start at source node
- Try to reach target by search
- States: node, along with previous arc
- Successor function:
  - Unobserved nodes:
    - To any child
    - To any parent if coming from a child
  - Observed nodes:
    - From parent to parent
- If you can't reach a node, it's conditionally independent of the start node

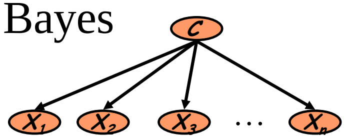


Example

- L ind. T', given T?
  Yes
- L ind. B?
  Yes
- L ind. B, given T?
  No
- L ind. B, given T'?
  No
- L ind. B, given T and R?
  Yes



Naïve Bayes



- Conditional Independence Assumption: features are independent of each other given the class:

$$P(X_1,\ldots,X_n \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_n \mid C)$$

- What can we model with naïve bayes?
- Any process where,
  - Each cause has lots of "independent" effects
  - Easy to estimate the CPT fro each effect
  - We want to reason about the probability of different causes given observed effects

Naive Bayes Classifiers

Task: Classify a new instance D based on a tuple of attribute values into one of the classes $c_j \in C$

$$D = \langle x_1, x_2, \ldots, x_n \rangle$$

$$c_{MAP} = \operatorname*{argmax}_{c \in C} P(c \mid x_1, x_2, \ldots, x_n)$$

$$= \operatorname*{argmax}_{c \in C} \frac{P(x_1, x_2, \ldots, x_n \mid c) P(c)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \operatorname*{argmax}_{c \in C} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct