

## Uncertain environments. Probabilistic reasoning.

Ch. 13 & 14.1 4

Oct 20, 2008

Lecture 11, COMP 131

1

## Last time

- Planning

Oct 20, 2008

Lecture 11, COMP 131

2

## Today

- Agents in uncertain environments
- Quick probability review
- Probabilistic reasoning
- Bayesian networks

Oct 20, 2008

Lecture 11, COMP 131

3

## Uncertain environments

- Partial observability
- Sensor error and noise
- Stochastic transitions between states
  - with some probability lightning will strike and burn down your house
- Complexity too much to handle logically

### Example:

agent is planning a trip route that involves taking a flight at time T

action A45 is available which starts driving to the airport at time T-45 minutes

will this action make the plan succeed?

Oct 20, 2008

Lecture 11, COMP 131

4

## Methods for handling uncertainty

- Default or nonmonotonic logic:
  - Assume my car does not have a flat tire
  - Assume A45 works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?
- Rules with fudge factors:
  - A45  $\mapsto$  0.3 get there on time (0.3 is the degree of belief in this rule)
  - Sprinkler  $\mapsto$  0.99 WetGrass (causal rules)
  - WetGrass  $\mapsto$  0.7 Rain (diagnostic rules)
- Issues: Problems with evidence combination, e.g., Sprinkler causes Rain?
- Probability
  - Model agent's degree of belief
  - Given the available evidence, A45 will get me there on time with probability 0.3

Oct 20, 2008

Lecture 11, COMP 131

5

## Rational decisions under uncertainty

- Suppose an action *AI* is successful with probability 0.9; is it a rational choice?
- That depends on the agent's preferences between possible outcomes (their **utility** to the agent)
  - outcome of *AI* succeeding has utility of +2
  - outcome of *AI* failing has utility of -2,000

maybe there's a better action...
- Remember expectiminimax from Ch. 6
  - utility of chance nodes =  $\sum_{outcome} P(outcome) \times Utility(outcome)$
- A **decision-theoretic** agent combines *probability theory* (for knowledge representation and reasoning) with *utility theory* (for stating preferences)

Oct 20, 2008

Lecture 11, COMP 131

6

## Probability: quick review

- Random variables
  - discrete (incl. Boolean): *Weather*, domain: < sunny, cloudy, rain, snow >
  - $P(\text{Weather}) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$  (sums to 1)
- Atomic events
  - complete specification of state with all variables' values assigned
  - atomic events are mutually exclusive
- Axioms of probability
  - $0 \leq P(a) \leq 1$
  - $P(\text{true}) = 1, P(\text{false}) = 0$
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
- Frequentist vs. Bayesian approach
  - **frequentist**:  $P(a) = 0.3$  means  $a$  is observed in 3 out 10 cases
  - **bayesian/subjectivist**:  $P(a) = 0.3$  means I am 30% likely to believe that  $a$  (degree of belief)

Oct 20, 2008

Lecture 11, COMP 131

7

## Conditional probability

- Prior (unconditional) probability distribution
  - $P(\text{Traffic}) = \langle 0.7, 0.3 \rangle$  with domain: < high, low >
- Joint probability distribution
  - $P(\text{Weather}, \text{Traffic})$  for all value combinations for *Weather* and *Traffic* (table)
- Conditional probability
  - $P(\text{toothache} | \text{cavity}) = 0.8$
  - $P(a|b) = P(a,b)/P(b)$  when  $P(b) > 0$  or  $P(a,b) = P(a|b)P(b)$
- Independence
  - $P(\text{Weather} | \text{Toothache}) = P(\text{Weather})$  or
  - $P(\text{Weather}, \text{Toothache}) = P(\text{Weather})P(\text{Toothache})$

upper case: random variables,  
lower case: values

Oct 20, 2008

Lecture 11, COMP 131

8

## Inference by enumeration

- Using full joint distribution  $P(\text{Toothache}, \text{Catch}, \text{Cavity})$ 
  - *Catch*: Boolean variable for the dentist's tool catching on the tooth

	toothache		$\neg$ toothache	
	catch	$\neg$ catch	catch	$\neg$ catch
cavity	.108	.012	.072	.008
$\neg$ cavity	.016	.064	.144	.576

- Can infer that:  $P(\text{cavity}) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$
- **Marginalization** or **summing out**: for any variable  $Y, Z$ 

$$P(Y) = \sum_z P(Y, z)$$

Oct 20, 2008

Lecture 11, COMP 131

9

## Bayes rule and inference

- Bayes rule
 
$$P(b | a) = \frac{P(a | b)P(b)}{P(a)}$$

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$



Rev. Thomas Bayes  
1702-1761

- Evidence-based Bayesian inference (simple case)

$$P(s | m) = 0.5$$

$s$  = stiff neck

$$P(m) = 0.00002$$

$m$  = meningitis

$$P(s) = 0.05$$

$$P(m | s) = \frac{P(s | m)P(m)}{P(s)} = 0.0002$$

Oct 20, 2008

Lecture 11, COMP 131

10

## Naïve Bayes

- Combining evidence
  - $P(\text{Cavity} | \text{toothache}, \text{catch})?$
  - for  $n$  variables there are  $2^n$  joint table entries
- Conditional independence
  - $P(\text{Toothache}, \text{Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity})P(\text{Catch} | \text{Cavity})$
  - given the presence or absence of a cause, variables can be independent conditioned on that cause
- Naïve Bayes model
 
$$P(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause})$$

$$P(\text{Cause} | \text{Effect}_1, \dots, \text{Effect}_n) = \frac{P(\text{Effect}_1, \dots, \text{Effect}_n | \text{Cause})P(\text{Cause})}{P(\text{Effect}_1, \dots, \text{Effect}_n)}$$

$$= \frac{P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause})}{P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause}) + P(\neg \text{Cause}) \prod_i P(\text{Effect}_i | \neg \text{Cause})}$$

Oct 20, 2008

Lecture 11, COMP 131

11

## Bayesian networks (Bayes nets)

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions ( $O(n)$ )
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link  $\approx$  "directly influences")
  - a conditional distribution for each node given its parents:
 
$$P(X_i | \text{Parents}(X_i))$$
- In the simplest case, conditional distribution represented as a **conditional probability table** (CPT) giving the distribution over  $X_i$  for each combination of parent values

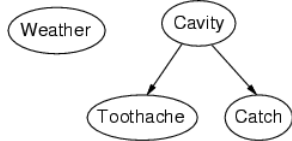
Oct 20, 2008

Lecture 11, COMP 131

12

## Example

- Topology of network encodes conditional independence assertions:



- Weather* is independent of the other variables
- Toothache* and *Catch* are conditionally independent given *Cavity*

Oct 20, 2008

Lecture 11, COMP 131

13

## Example

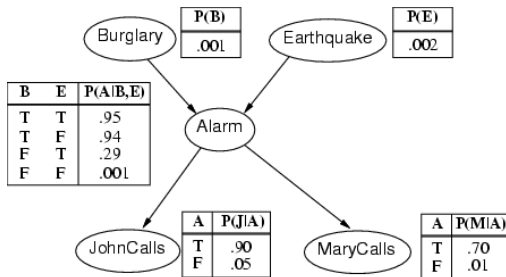
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Network topology reflects "causal" knowledge:
  - A burglar can set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

Oct 20, 2008

Lecture 11, COMP 131

14

## Example contd.



Oct 20, 2008

Lecture 11, COMP 131

15

## 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  $X_i = \text{true}$  (the number for  $X_i = \text{false}$  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.  $2^5 - 1 = 31$  for full joint distribution)



Oct 20, 2008

Lecture 11, COMP 131

16

## Global semantics

The full joint distribution is defined as the product of the local conditional distributions:

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

e.g.,  $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= P(j | a) P(m | a) P(a | \neg b, \neg e) P(\neg b) P(\neg e)$$



Oct 20, 2008

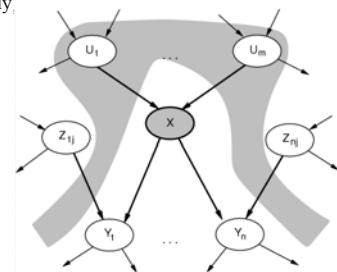
Lecture 11, COMP 131

17

## Local semantics

Each node is conditionally independent of its nondescendants given its parents

**Theorem:**  
local semantics  $\leftrightarrow$   
global semantics



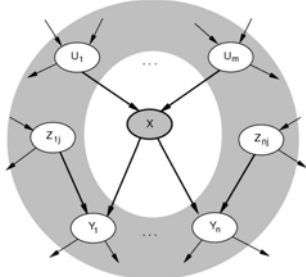
Oct 20, 2008

Lecture 11, COMP 131

18

## Markov blanket

Each node is conditionally independent of all others given its  
**Markov blanket**: *parents + children + children's parents*



Oct 20, 2008

Lecture 11, COMP 131

19

## Constructing Bayesian networks

1. Choose an ordering of variables  $X_1, \dots, X_n$ 
  - add “root causes” first, then their effects, etc. until “leaves”
2. For  $i = 1$  to  $n$ 
  - add  $X_i$  to the network
  - select parents from  $X_1, \dots, X_{i-1}$  such that  

$$P(X_i | Parents(X_i)) = P(X_i | X_1, \dots, X_{i-1})$$
**independence given parents**

This choice of parents guarantees:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) \text{ (chain rule)}$$

$$= \prod_{i=1}^n P(X_i | Parents(X_i)) \text{ (by construction)}$$

Oct 20, 2008

Lecture 11, COMP 131

20

## Variable order important! example

- Suppose we choose the ordering  $M, J, A, B, E$



$$P(J | M) = P(J)?$$

Oct 20, 2008

Lecture 11, COMP 131

21

## Example

- Suppose we choose the ordering  $M, J, A, B, E$



$$P(J | M) = P(J)? \text{ No}$$

$$P(A | J, M) = P(A | J)? P(A | J, M) = P(A)?$$

Oct 20, 2008

Lecture 11, COMP 131

22

## Example

- Suppose we choose the ordering  $M, J, A, B, E$



$$P(J | M) = P(J)? \text{ No}$$

$$P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? \text{ No}$$

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

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

Oct 20, 2008

Lecture 11, COMP 131

23

## Example

- Suppose we choose the ordering  $M, J, A, B, E$



$$P(J | M) = P(J)? \text{ No}$$

$$P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? \text{ No}$$

$$P(B | A, J, M) = P(B | A)? \text{ Yes}$$

$$P(B | A, J, M) = P(B)? \text{ No}$$

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

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

Oct 20, 2008

Lecture 11, COMP 131

24

## Example

- Suppose we choose the ordering M, J, A, B, E



$P(J | M) = P(J)$ ? No  
 $P(A | J, M) = P(A | J)$ ?  $P(A | J, M) = P(A)$ ? No  
 $P(B | A, J, M) = P(B | A)$ ? Yes  
 $P(B | A, J, M) = P(B)$ ? No  
 $P(E | B, A, J, M) = P(E | A)$ ? No  
 $P(E | B, A, J, M) = P(E | A, B)$ ? Yes

Oct 20, 2008

Lecture 11, COMP 131

25

## Example contd.



- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact:  $1 + 2 + 4 + 2 + 4 = 13$  numbers needed

Oct 20, 2008

Lecture 11, COMP 131

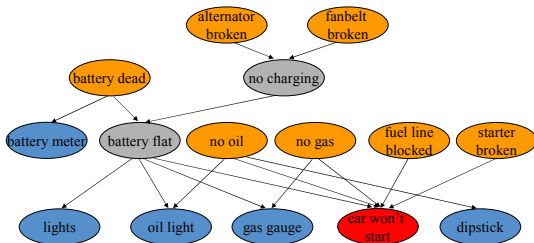
26

## Example: car diagnosis

Initial evidence: car won't start

Testable variables (blue), fixable cause variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters



Oct 20, 2008

Lecture 11, COMP 131

27

## Compact conditional distributions

- CPT grows exponentially with number of parents
- CPT becomes infinite with continuous-valued parent or child
- Solution: canonical distributions that are defined compactly
  - common pattern, specify a few parameters
- Deterministic nodes are the simplest case:
  - $X = f(\text{Parents}(X))$  for some function  $f$
  - e.g., Boolean functions
    - $\text{NorthAmerican} \leftrightarrow \text{Canadian} \vee \text{US} \vee \text{Mexican}$
  - e.g., numerical relationships among continuous variables
    - $\partial \text{Level} / \partial t = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation}$

Oct 20, 2008

Lecture 11, COMP 131

28

## Compact conditional distributions

Noisy-Or distributions model multiple non-interacting causes

- parents  $U_1, \dots, U_k$  include all causes (can add leak nodes)
- independent "causation failure" probability  $q_i$  for each cause alone

$$P(X | u_1, \dots, u_j, \neg u_{j+1}, \dots, \neg u_k) = 1 - \prod_{i=1}^k q_i$$

Cold	Flu	Malaria	$P(\text{Fever})$	$P(\neg \text{Fever})$	num parameters linear in num parents
F	F	F	0.0	1.0	
F	F	T	0.9	0.1	
F	T	F	0.8	0.2	
F	T	T	0.98	$0.02 = 0.2 \times 0.1$	
T	F	F	0.4	0.6	
T	F	T	0.94	$0.06 = 0.6 \times 0.1$	
T	T	F	0.88	$0.12 = 0.6 \times 0.2$	
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$	

Oct 20, 2008

Lecture 11, COMP 131

29

## Inference in Bayes nets

- Simple queries: compute **posterior marginal**  $P(X | E=e)$ 
  - e.g.,  $P(\text{NoGas} | \text{Gauge} = \text{empty}, \text{Lights} = \text{on}, \text{Starts} = \text{false})$
- Conjunctive queries:
  - $P(X_p, X_c | E=e) = P(X_p | E=e) P(X_c | X_p, E=e)$
- Optimal decisions: networks include utility information
  - $P(\text{outcome} | \text{action}, \text{evidence})$
- Value of information: which evidence to seek out?
- Sensitivity analysis: which probability values are most critical?
- Explanation: why do I need a new starter motor?

Oct 20, 2008

Lecture 11, COMP 131

30

## Exact inference by enumeration

- Compute conditional probability by summing terms from joint distribution

$$P(X | z) = \alpha P(X, z) = \alpha \sum_y P(X, z, y)$$

$\alpha$  is normalization in Bayes rule

- $P(B | j, m)$ ?

$$P(B | j, m) = \alpha \sum_e \sum_a P(B, e, a, j, m)$$

for  $B=true$ :

$$P(b | j, m) = \alpha \sum_e \sum_a P(b)P(e)P(a|b,e)P(j|a)P(m|a)$$

$$= \alpha P(b) \sum_e P(e) \sum_a P(a|b,e)P(j|a)P(m|a)$$

$O(2^n)$

- Loop through variables in order, multiplying CPT entries



Oct 20, 2008

Lecture 11, COMP 131

31

## Variable elimination

- A better form of exact inference
- **Variable elimination** algorithm: carry out summations right-to-left, storing immediate results (**factors**) to avoid repeating computation



$$P(B | j, m)$$

$$= \alpha P(B) \sum_e P(e) \sum_a P(a|B,e)P(j|a)P(m|a)$$

$$= \alpha P(B) \sum_e P(e) \sum_a P(a|B,e)P(j|a)f_M(a)$$

$$= \alpha P(B) \sum_e P(e) \sum_a P(a|B,e) f_J(a) f_M(a)$$

$$= \alpha P(B) \sum_e P(e) \sum_a f_A(a,b,e) f_J(a) f_M(a)$$

$$= \alpha P(B) \sum_e P(e) f_{E,AM}(b,e) \quad (\text{sum out } A)$$

$$= \alpha P(B) f_{E,AM}(b) \quad (\text{sum out } E)$$

$$= \alpha f_B(b) \times f_{E,AM}(b)$$

$f_M(A) = [P(m|a), P(m|\neg a)]$

$f_A(A,B,E)$  is a  $2 \times 2 \times 2$  matrix:  $P(a|B,E)$

Oct 20, 2008

Lecture 11, COMP 131

32

## Variable elimination: basic ops

**Summing out** a variable from a product of factors:

- move any constant factors outside the summation and add up submatrices in **pointwise product** of remaining factors

$$\sum_x f_1 \times \dots \times f_k = f_1 \times \dots \times f_i \sum_x f_{i+1} \times \dots \times f_k = f_1 \times \dots \times f_i f_X$$

assuming  $f_1, \dots, f_i$  do not depend on  $X$

**Pointwise product** of factors  $f_1$  and  $f_2$ :

$$f_1(x_1, \dots, x_p, y_1, \dots, y_k) \times f_2(y_1, \dots, y_k, z_1, \dots, z_l)$$

$$= f(x_1, \dots, x_p, y_1, \dots, y_k, z_1, \dots, z_l)$$

Oct 20, 2008

Lecture 11, COMP 131

33

## Pointwise product example

$$\text{E.g., } f_1(A,B) \times f_2(B,C) = f(A,B,C)$$

A	B	$f_1(A,B)$	B	C	$f_2(B,C)$	A	B	C	$f(A,B,C)$
T	T	.3	T	T	.2	T	T	T	.3×.2
T	F	.7	T	F	.8	T	T	F	.3×.8
F	T	.9	F	T	.6	T	F	T	.7×.6
F	F	.1	F	F	.4	T	F	F	.7×.4
						F	T	T	.9×.2
						F	T	F	.9×.8
						F	F	T	.1×.6
						F	F	F	.1×.4

Oct 20, 2008

Lecture 11, COMP 131

34

## Variable elimination algorithm

```
function ELIMINATION_ASK(X, e, bn)
  returns prob. distribution over X
  inputs: X, the query variable
         e, evidence specified as an event
         bn, a belief network specifying joint
            distribution  $P(X_1, \dots, X_n)$ 
```

```
factors := []; vars := REVERSE(VARS[bn]) right to left
```

```
for each var in vars do
```

```
  factors := [MAKE_FACTOR(var,e)|factors]
```

```
  if var is a hidden variable then
```

```
    factors := SUM_OUT(var,factors) if not X and not e
```

```
return NORMALIZE(POINTWISE_PRODUCT(factors))
```

Oct 20, 2008

Lecture 11, COMP 131

35

## Irrelevant variables

- Query:  $P(\text{JohnCalls} | \text{Burglary} = \text{true})$

$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(J,a) \sum_m P(m|a)$$

Sum over  $m$  is identically 1;  $M$  is **irrelevant** to the query

- Theorem 1:  $Y$  is irrelevant unless  $Y \in \text{Ancestors}(X \cup E)$

Here,  $X = \text{JohnCalls}$ ,  $E = \{\text{Burglary}\}$

$\text{Ancestors}(X \cup E) = \{\text{Alarm}, \text{Earthquake}\}$

so  $\text{MaryCalls}$  is irrelevant

- Compare to backward chaining in Horn clause KBs



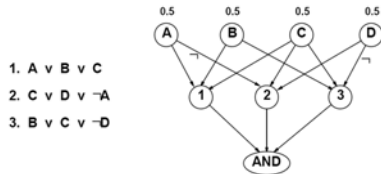
Oct 20, 2008

Lecture 11, COMP 131

36

## Complexity of exact inference

- Singly connected networks (**polytrees**):
  - any two nodes are connected by at most one (undirected) path
  - time and space cost of variable elimination are  $O(d^n)$
- Multiply connected networks:
  - can reduce 3SAT to exact inference, so exact inference is NP-hard



Oct 20, 2008

Lecture 11, COMP 131

37

## Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology (DAG with nodes = variables) + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct
- Exact inference by variable elimination: evaluate sums of products of conditional probabilities efficiently
- In polytrees, exact inference takes time linear in the size of the network; in the general case, intractable.

Oct 20, 2008

Lecture 11, COMP 131

38

## Next time

- Approximate inference in Bayes Nets
- Dynamic Bayes Nets (probabilistic reasoning over time)

Oct 20, 2008

Lecture 11, COMP 131

39