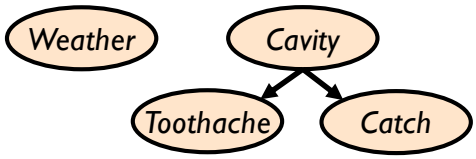
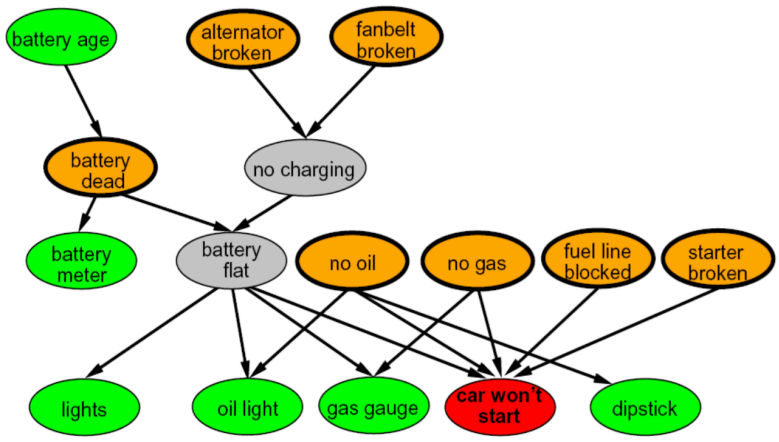
Inference in Bayes Nets
AI Class 10 (Ch. 14.1–14.4.2; skim 14.3)



1

Bookkeeping

- Last time
  - Probability review
  - Bayesian Belief Nets
- This class
  - Review of BBNs
  - The Chain Rule
  - A little more about conditional probability
  - But mostly… **inference in Bayes' nets**



2

How is the Bayesian network created?

1. Get an expert to design it
   - Expert must determine the structure of the Bayesian network
     - This is best done by modeling direct causes of a variable as its parents
   - Expert must determine the values of the CPT entries
     - These values could come from the expert's informed opinion
     - Or an external source like census information
     - Or they are estimated from data
     - Or a combination of the above
2. Learn it from data
   - This is a much better option but it usually requires a large amount of data
   - This is where Bayesian statistics comes in!

3

Probability, redux

- Worlds, random variables, events, sample space
- **Joint probabilities** of multiple connected variables
- **Conditional probabilities** of a variable, given another variable(s)
- **Marginalizing out** unwanted variables
- **Inference** from the joint probability

The big idea: figuring out the probability of variable(s) taking certain value(s)

4

Review: Independence

What does it mean for A and B to be **independent**?

- $P(A) \perp\!\!\!\perp P(B)$
- A and B do not affect each other's probability
- $P(A \wedge B) = P(A)\ P(B)$

6

Review: Conditioning

What does it mean for A and B to be **conditionally independent given C?**

- A and B don't affect each other **if C is known**
- $P(A \wedge B \mid C) = P(A \mid C)\ P(B \mid C)$

7

Review: Bayes' Rule

What is **Bayes' Rule**?

$$P(H_i \mid E_j) = \frac{P(E_j \mid H_i)P(H_i)}{P(E_j)}$$

What's it useful for?

- Diagnosis
- Effect is perceived, want to know (probability of) cause

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

R&N, 495–496

8

Review: Bayes' Rule

What is **Bayes' Rule**?

$$P(H_i \mid E_j) = \frac{P(E_j \mid H_i)P(H_i)}{P(E_j)}$$

What's it useful for?

- Diagnosis
- Effect is perceived, want to know (probability of) cause

$$P(hidden \mid observed) = \frac{P(observed \mid hidden)P(hidden)}{P(observed)}$$

R&N, 495–496

9

Review: Joint Probability

- What is the **joint probability** of A and B?
  - $P(A,B)$
- The probability of **any pair** of legal assignments.
  - Generalizing to > 2, of course
- Booleans: expressed as a matrix/table

	alarm	¬alarm
burglary	0.09	0.01
¬burglary	0.1	0.8

≡

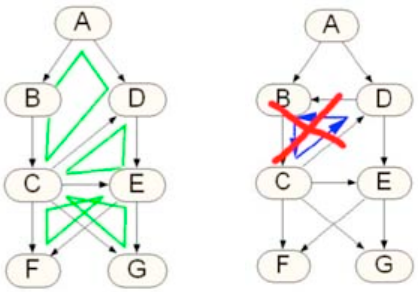
A	B	
T	T	0.09
T	F	0.1
F	T	0.01
F	F	0.8

- Continuous domains: probability functions

10

Next Up

- Bayesian networks
  - Network structure and independence
- Inference in Bayesian networks
  - Exact inference
  - Approximate inference



DAG: In a Bayes net, the links may form loops, but they may not form cycles.

www.norsys.com/tutorials/netica/secA/tut_A1.htm

11

Review: Bayes' Nets: Big Picture

- Problems with full joint distribution tables as our probabilistic models:
  - Joint gets **way** too big to represent explicitly
    - Unless there are only a few variables
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

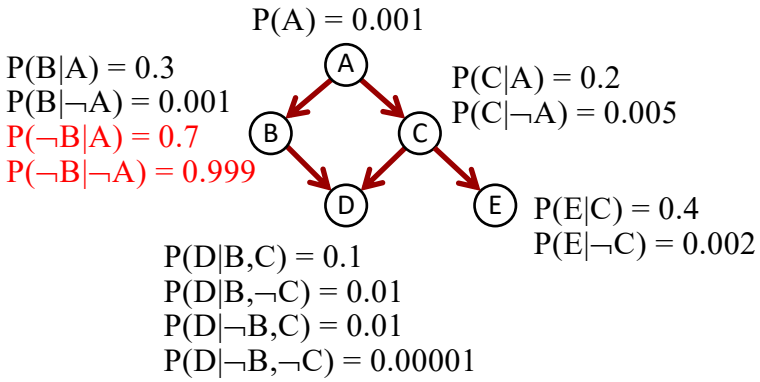
	A		¬A	
	E	**¬E**	**E**	**¬E**
B	0.01	0.08	0.001	0.009
¬B	0.01	0.09	0.01	0.79

Slides derived from Matt E. Taylor, U Alberta

12

Review: Bayes' Nets

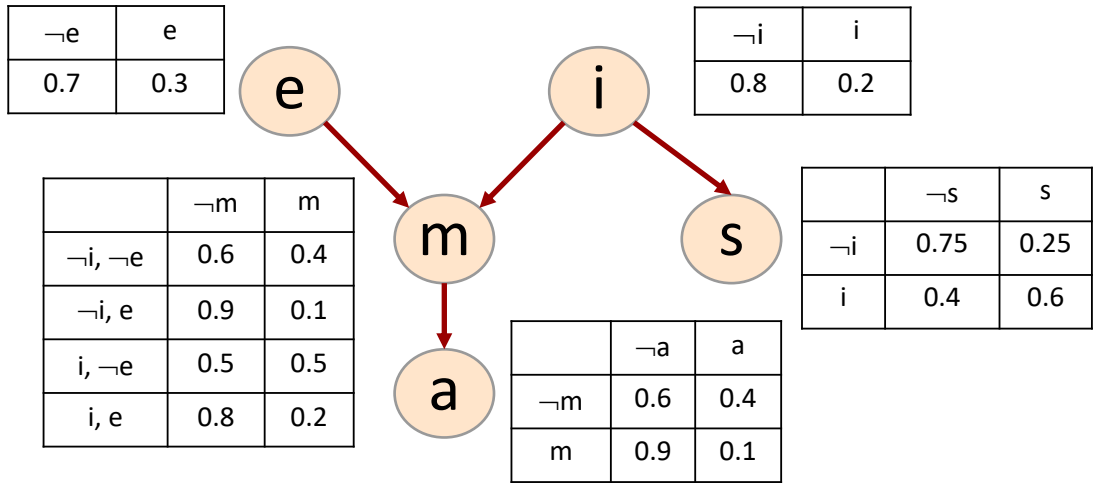
- Bayesian Network: **BN = (DAG, CPD)**
  - **DAG**: directed acyclic graph (BN's structure)
  - **CPD**: conditional probability distribution (BN's parameters)



13

Review: Bayes' Nets

- P(a, m, i, e, s) = P(a | m) * P(m | i, e) * P(i) * P(e) * P(s | i)



¬e	e
0.7	0.3

¬i	i
0.8	0.2

	¬m	m
¬i, ¬e	0.6	0.4
¬i, e	0.9	0.1
i, ¬e	0.5	0.5
i, e	0.8	0.2

	¬s	s
¬i	0.75	0.25
i	0.4	0.6

	¬a	a
¬m	0.6	0.4
m	0.9	0.1

www.upgrad.com/blog/bayesian-network-example/

14

The Chain Rule

- $P(\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n) = P(\alpha_1) \times P(\alpha_2 \mid \alpha_1) \times P(\alpha_3 \mid \alpha_1 \wedge \alpha_2) \times \ldots \times P(\alpha_n \mid \alpha_1 \wedge \cdots \wedge \alpha_{n-1})$

$= \Pi_{i=1..n} P(\alpha_i \mid \alpha_1 \wedge \cdots \wedge \alpha_{i-1})$

$= P(x_1, \ldots, x_n) = \Pi_{i=1}^{n} P(x_i \mid \pi_i)$

artint.info/html/ArtInt_143.html

15

The Chain Rule

$$P(x_1,\ldots,x_n) = \Pi_{i=1}^{n} P(x_i \mid \pi_i)$$

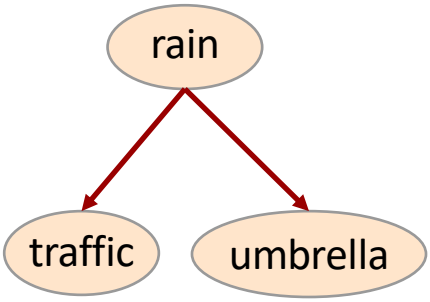
- Decomposition: $P(x_1,\ldots,x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x2)\ldots$

  $P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$
  $P(\text{Rain})\ P(\text{Traffic} \mid \text{Rain})\ P(\text{Umbrella} \mid \text{Rain}, \text{Traffic})$

- With assumption of conditional independence:

  $P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$
  $P(\text{Rain})\ P(\text{Traffic} \mid \text{Rain})\ P(\text{Umbrella} \mid \text{Rain})$

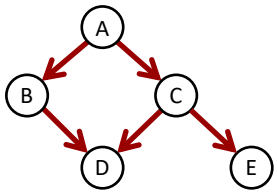
- Bayes' nets express conditional independences
  - (Assumptions)



Slides derived from Matt E. Taylor, U Alberta

16

Chaining: Example



Computing the joint probability for all variables is easy:

P(a, b, c, d, e) = P(e | a, b, ***c,*** d) P(a, b, c, d) ← ***By product rule***
= P(e | c) P(a, b, c, d) ← ***By conditional independence assumption***
= P(e | c) P(d | a, ***b, c***) P(a, b, c)
= P(e | c) P(d | b, c) P(c | ***a,*** b) P(a, b)
= P(e | c) P(d | b, c) P(c | a) P(b | a) P(a)

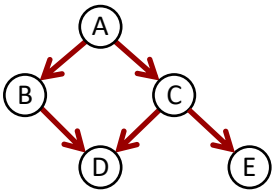
We're reducing distributions—P(x,y)—to single values.

17

Bayesian Networks

Two important properties:

1. Encodes the **conditional independence** relationships between the variables in the graph structure
2. Is a compact representation of the **joint probability distribution** over the variables



18

Topological Semantics

- A node is **conditionally independent** of its non-descendants given its parents
- A node is **conditionally independent** of all other nodes in the network given its parents, children, and children's parents (also known as its **Markov blanket**)

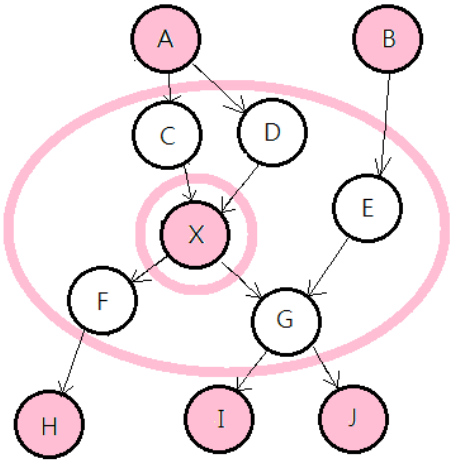


Image: mitsai1974.github.io/DevBlog/2018/07/11/bayesian-ml-net-profound

19

The Joint Probability Distribution

- Due to the Markov condition, we can compute the joint probability distribution over all the variables X_1, …, X_n in the Bayesian net using the formula:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid Parents(X_i))$$

Where Parents(X_i) means the values of the Parents of the node X_i with respect to the graph

21

Independence and Causal Chains

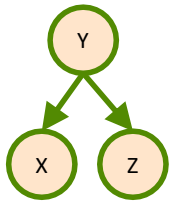
- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, we can it prove using algebra (tedious)
  - If no, can prove it with a counter-example
- Question: are X and Z necessarily independent?
  - No.
  - Ex: Clouds (X) cause rain (Y), which causes traffic (Z)
  - X can influence Z, Z can influence X (via Y)
- This configuration is a "causal chain"

Slides derived from Matt E. Taylor, WSU

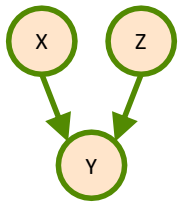
22

Two More Main Patterns

- Common Cause:
  - Y causes X and Y causes Z
  - Are X and Z independent? **No**
  - Are X and Z independent given Y? **Yes**



- Common Effect:
  - Two causes of one effect
  - Are X and Z independent? **Yes**
  - Are X and Z independent given Y?
  - No!
  - Observing an effect "activates" influence between possible causes.

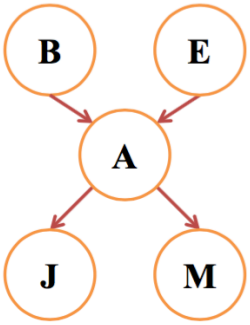


Slides derived from Matt E. Taylor, WSU

23

Conditionality Example

- Hidden: A, B, E. You don't know:
  - If there's a burglar.
  - If there was an earthquake.
  - If the alarm is going off.
- Observed: J and M.
  - John and/or Mary have some chance of calling if the alarm rings.
  - You know who called you.

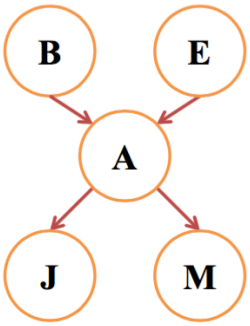


Slides derived from Matt E. Taylor, WSU

24

Conditionality Example 2

- At first (before observations):
  - Is the probability of John calling affected by whether there's an earthquake? **Yes**
  - Is the probability of Mary calling affected by John calling? **Yes**
- Then: Your alarm is going off! (observation)
  - Is the probability of Mary calling affected by John calling? **No**

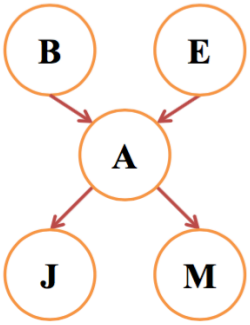


Slides derived from Matt E. Taylor, WSU

25

Conditionality Example 3

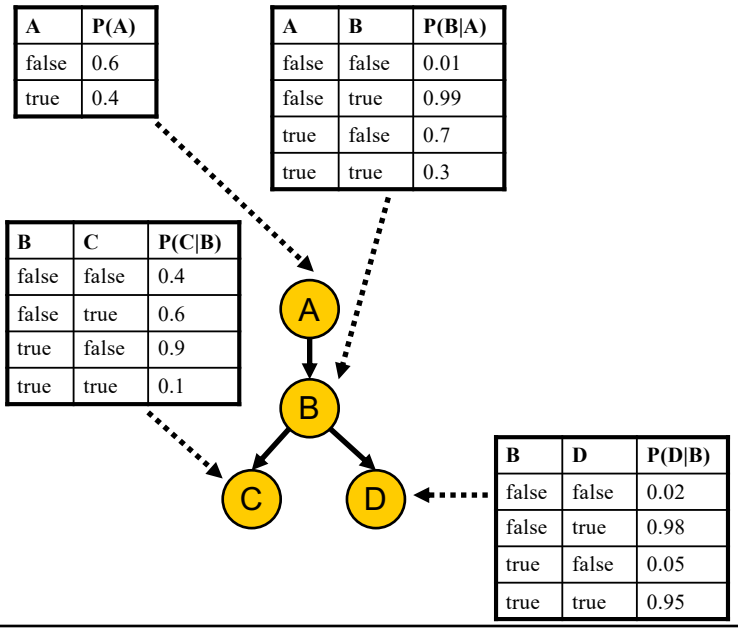
- At first (before observations):
  - Is whether there's an earthquake affected by whether there's a burglary in progress (and vice versa)? **No**
- Then: Your alarm is going off! (observation)
  - Does the probability a burglary is happening depend on whether there's an earthquake? **YES!**



Slides derived from Matt E. Taylor, WSU

26

A Set of Tables for Each Node



A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Each node X_i has a conditional probability distribution $P(X_i \mid \text{Parents}(X_i))$ that quantifies the effect of the parents on the node

The parameters are the probabilities in these conditional probability tables (CPTs)

28

Using a Bayesian Network Example

- Using the network in the example, suppose you want to calculate:

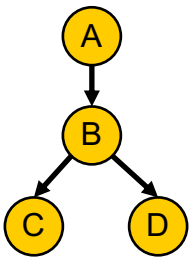
$$P(A = \text{true}, B = \text{true}, C = \text{true}, D = \text{true})$$

$$= P(A = \text{true}) * P(B = \text{true} \mid A = \text{true}) * P(C = \text{true} \mid B = \text{true}) P(D = \text{true} \mid B = \text{true})$$

This is from the graph structure

$$= (0.4)*(0.3)*(0.1)*(0.95)$$

These numbers are from the conditional probability tables



29

Representational Extensions

- Conditional probability tables (CPTs) for large networks can require a large number of parameters
  - $O(2^k)$ where k is the branching factor of the network
- There are ways of compactly representing CPTs
  - Deterministic relationships
  - Noisy-OR
  - Noisy-MAX
- What about continuous variables?
  - Discretization
  - Use density functions (usually mixtures of Gaussians) to build hybrid Bayesian networks (with discrete and continuous variables)

30

Bayes' Net Inference

Some material from Lise Getoor

31

What Are We Trying To Do?

- Now we know the **semantics** of Bayes' nets
- But how do we use it?
- Say we have some **evidence** (that is, some variables are instantiated)
- We usually want to know the probability of some **other** variables
- Why?
  - Reason about **hidden** (non-observed) information
  - What **caused** something?
  - What is the **probability** of something?

32

Inference

- Instead of computing the joint, suppose we just want the probability for one variable (or a subset)
- Using a Bayesian network to compute probabilities is called **inference**
- In general, inference involves queries of the form:

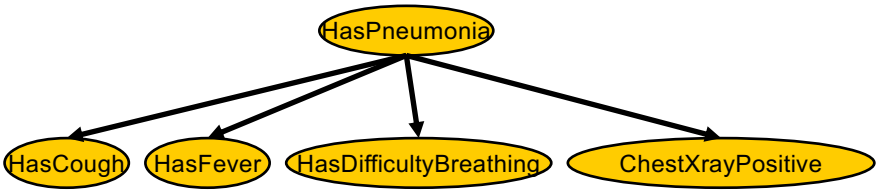
P(X | E)

E = The evidence variable(s)

X = The query variable(s)

33

Inference



- An example of a query would be:

  P(HasPneumonia = true | HasFever = true, HasCough = true)

- Note: Even though HasDifficultyBreathing and ChestXrayPositive are in the Bayesian network, they are not given values in the query (ie. they do not appear either as query variables or evidence variables)
- They are treated as unobserved variables

34

Inference Techniques

- **Exact inference:** Analytically compute the conditional probability distribution over the variables we care about
- **Approximate inference:** Sometimes exact inference is too hard
  - Come up with approximate solutions based on **statistical sampling**

Exact inference
- Enumeration
- Belief propagation in polytrees
- Variable elimination
- Clustering / join tree algorithms

Approximate inference
- Stochastic simulation / sampling methods
- Markov chain Monte Carlo methods
- Genetic algorithms
- Neural networks
- Simulated annealing
- Mean field theory

35

Query Types

Given a Bayesian network, what questions might we want to ask?

- Conditional probability query: P(x | e)
  - Given instantiations for some of the variables (e = values of all instantiated variables; it doesn't have to be just one), what is the probability that node X has a particular value *x*?

36

Query Types

Given a Bayesian network, what questions might we want to ask?

- Conditional probability query: P(x | e)
  - Given instantiations for some of the variables (e = values of all instantiated variables; it doesn't have to be just one), what is the probability that node X has a particular value *x*?
- Maximum a posteriori probability: What value of x maximizes P(x|e) ?
  - What is the most likely explanation for some evidence?
  - That is, what is the value of node(s) X that maximizes the probability that you would have seen the evidence you did?
  - This is called a **maximum a posteriori probability** or MAP query

37

Query Types

Given a Bayesian network, what questions might we want to ask?

- Conditional probability query: P(x | e)
- Maximum a posteriori probability: What value of x maximizes P(x|e) ?
- **General question: What's the whole probability distribution over variable X given evidence e — what is P(X | e)?**

38

Inference Tasks

- **Simple queries:** Compute posterior marginal $P(X_i \mid E=value)$
  - E.g., P(NoGas | Gauge=*empty*, Lights=*on*, Starts=*false*)
- **Conjunctive queries:**
  - $P(X_i, X_j \mid E=value) = P(X_i \mid E=value)\, P(X_j \mid X_i, E=value)$
- **Optimal decisions:**
  - *Decision networks* include utility information
  - Probabilistic inference gives P(outcome | action, evidence)
- **Value of information:** Which evidence should we seek next?
- **Sensitivity analysis:** Which probability values are most critical?
- **Explanation:** Why do I need a new starter motor?

39

Using the joint distribution

- To answer any query involving a conjunction of variables, sum over the variables not involved in the query

40

Using the joint distribution

- To answer any query involving a conjunction of variables, sum over the variables not involved in the query

$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$

$$= \sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} \sum_{c \in \mathrm{dom}(C)} \Pr(A = a \wedge B = b \wedge C = c)$$

41

Using the joint distribution

- To answer any query involving a conjunction of variables, sum over the variables not involved in the query

$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$

$$= \sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} \sum_{c \in \mathrm{dom}(C)} \Pr(A = a \wedge B = b \wedge C = c)$$

42

Using the joint distribution

- To answer any query involving a conjunction of variables, sum over the variables not involved in the query

$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$

$$= \sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} \sum_{c \in \mathrm{dom}(C)} \Pr(A = a \wedge B = b \wedge C = c)$$

$$\Pr(d \mid b) = \frac{\Pr(b,d)}{\Pr(b)} = \frac{\sum_{AC} \Pr(a,b,c,d)}{\sum_{ACD} \Pr(a,b,c,d)}$$

Summing over A and C, because b and d are instantiated in the query

43

Inference by Enumeration

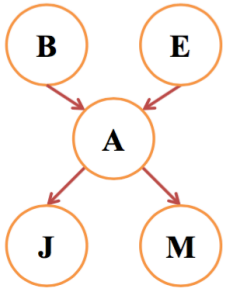
Reminder: P(E) is known (observed), so 1/P(E) is a constant that makes everything sum to 1: the *normalizing constant*

- Add all of the terms (atomic event probabilities) from the full joint distribution
- If **E** are the evidence (observed) variables and **Y** are the other (unobserved) variables, then:
  - $P(X \mid \mathbf{E}) = \alpha\, P(X, \mathbf{E}) = \alpha \sum P(X, \mathbf{E}, \mathbf{Y})$
- Each P(X, **E, Y**) term can be computed using the chain rule
- Computationally expensive!

44

Example 2: Enumeration

- Recipe:
  - State the marginal probabilities you need
  - Figure out ALL the atomic probabilities you need
  - Calculate and combine them
- Example:
  - $P(+b \mid +j, +m) = \dfrac{P(+b, +j, +m)}{P(+j, +m)}$



Slides derived from Matt E. Taylor, WSU; Russell&Norvig

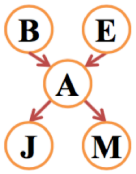
45

Example 2 cont'd

$$P(+b,+j,+m) =$$
$$P(+b)P(+e)P(+a|+b,+e)P(+j|+a)P(+m|+a)+$$
$$P(+b)P(+e)P(-a|+b,+e)P(+j|-a)P(+m|-a)+$$
$$P(+b)P(-e)P(+a|+b,-e)P(+j|+a)P(+m|+a)+$$
$$P(+b)P(-e)P(-a|+b,-e)P(+j|-a)P(+m|-a)$$

Because E and A are unobserved, we have to consider all cases

P(+m | +b, +e)?



46

Example 2 cont'd

$$P(+b,+j,+m) =$$
$$P(+b)P(+e)P(+a|+b,+e)P(+j|+a)P(+m|+a)+$$
$$P(+b)P(+e)P(-a|+b,+e)P(+j|-a)P(+m|-a)+$$
$$P(+b)P($$
$$P(+b)P($$

Because E and A are unobserved, we have to consider all cases

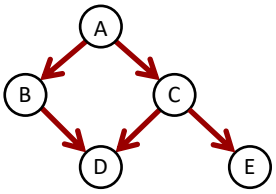
The number of possible assignments is exponential in the unobserved variables.

That is, unfortunately, the best we can do. General querying of Bayesian networks is NP-complete

47

Example 3: Enumeration

- $P(x_i) = \Sigma_{\pi_i} P(x_i \mid \pi_i) P(\pi_i)$
- Say we want to know $P(D=t)$
- Only E is *given* as true
- $P(d \mid e) = \alpha \Sigma_{ABC} P(a, b, c, d, e)$
  $= \alpha \Sigma_{ABC} P(a) P(b \mid a) P(c \mid a) P(d \mid b, c) P(e \mid c)$
- With simple iteration, that's a lot of repetition!
  - P(e | c) has to be recomputed every time we iterate over C=true



reminder:
π_i = parents of x_i

reminder:
$\alpha = 1/P(e)$

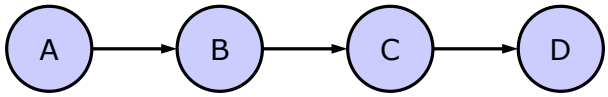
48

Better: Variable Elimination

- Basically just enumeration with caching of local calculations
- Linear for polytrees (singly connected BNs)
- Potentially exponential for multiply connected BNs
  - **Exact inference in Bayesian networks is still NP-hard!**
- Join tree algorithms are an extension of variable elimination methods that compute posterior probabilities for all nodes in a BN simultaneously

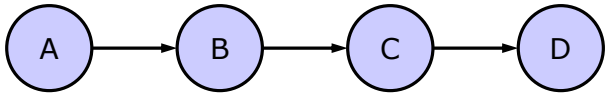
49

Simple Case



50

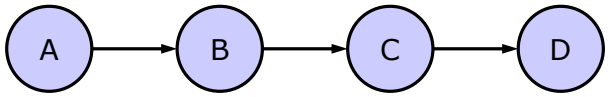
Simple Case



$$\Pr(d) = \sum \Pr(a,b,c,d)$$

51

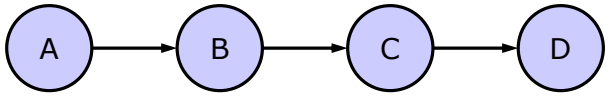
Simple Case



$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$
$$= \sum_{ABC} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$

52

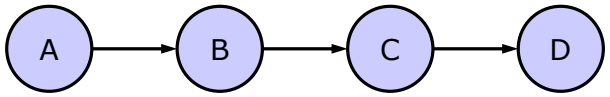
Simple Case



$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$
$$= \sum_{ABC} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$

53

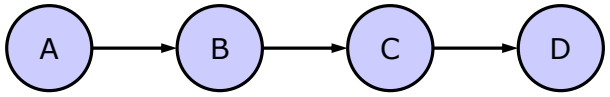
Simple Case



$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$
$$= \sum_{ABC} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$
$$= \sum_{C}\sum_{B}\sum_{A} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$

54

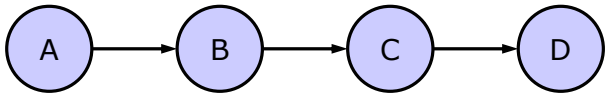
Simple Case



$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$
$$= \sum_{ABC} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$
$$= \sum_{C}\sum_{B}\sum_{A} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$
$$= \sum_{C} \Pr(d \mid c) \sum_{B} \Pr(c \mid b) \sum_{A} \Pr(b \mid a)\Pr(a)$$

55

Simple Case

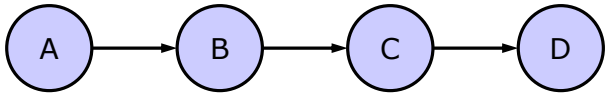


How can we do better?

$$\Pr(d) = \sum_{ABC} \Pr(a,b,c,d)$$
$$= \sum_{ABC} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$
$$= \sum_{C}\sum_{B}\sum_{A} \Pr(d \mid c)\Pr(c \mid b)\Pr(b \mid a)\Pr(a)$$
$$= \sum_{C} \Pr(d \mid c) \sum_{B} \Pr(c \mid b) \sum_{A} \Pr(b \mid a)\Pr(a)$$

56

Simple Case

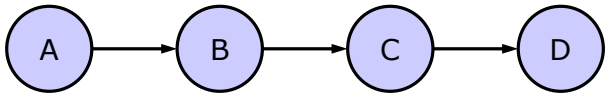


$$\Pr(d) = \sum_{C} \Pr(d \mid c) \sum_{B} \Pr(c \mid b) \sum_{A} \Pr(b \mid a)\Pr(a)$$

$$\begin{bmatrix} \Pr(b_1 \mid a_1)\Pr(a_1) & \Pr(b_1 \mid a_2)\Pr(a_2) \\ \Pr(b_2 \mid a_1)\Pr(a_1) & \Pr(b_2 \mid a_2)\Pr(a_2) \end{bmatrix}$$

57

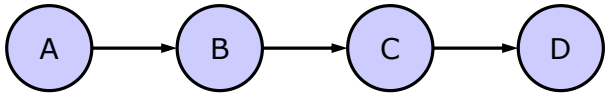
Simple Case



$$\Pr(d) = \sum_C \Pr(d \mid c) \sum_B \Pr(c \mid b) \underbrace{\sum_A \Pr(b \mid a) \Pr(a)}_{\begin{bmatrix} \sum_A \Pr(b_1 \mid a) \Pr(a) \\ \sum_A \Pr(b_2 \mid a) \Pr(a) \end{bmatrix}}$$

58

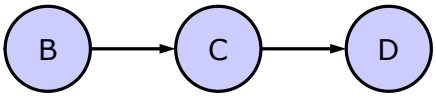
Simple Case



$$\Pr(d) = \sum_C \Pr(d \mid c) \sum_B \Pr(c \mid b) \underbrace{\sum_A \Pr(b \mid a) \Pr(a)}_{f_1(b)}$$

59

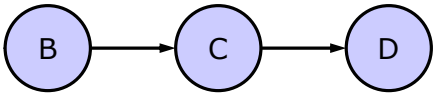
Simple Case



$$\Pr(d) = \sum_C \Pr(d \mid c) \sum_B \Pr(c \mid b) f_1(b)$$

60

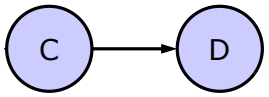
Simple Case



$$\Pr(d) = \sum_C \Pr(d \mid c) \underbrace{\sum_B \Pr(c \mid b) f_1(b)}_{f_2(c)}$$

61

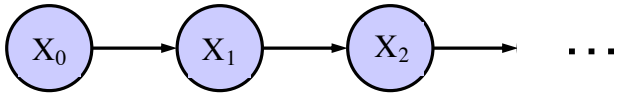
Simple Case



$$\Pr(d) = \sum_{C} \Pr(d \mid c)\, f_2(c)$$

62

Markov Chains

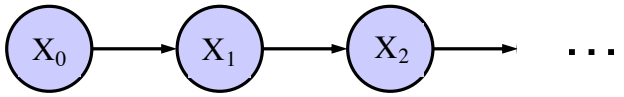


- A **Markov chain** is a special kind of BBN
- X_t represents **state at time t**
- Its dependence structure gives: $P(X_t | X_1, \ldots, X_{t-1}) = P(X_t | X_{t-1})$
  - This CPD is called the **state transition probability**
    - What is the probability of going from a particular state to the next?
  - Intuitively, X_t conveys all information about the past that can affect future states
- JPD of a Markov chain: $P(X_0,\ldots,X_T) = P(X_0) \times \prod_{t=1}^{T} P(X_t|X_{t-1})$

Poole & Mackworth

63

Markov Chains



- **Stationary Markov chain:** all state transition probability tables are the same
  - $P(X_3|X_2) = P(X_2|X_1) = P(X_1|X_0) = \ldots$
  - For all $t > 0$, $t' > 0$: $P(X_t|X_{t-1}) = P(X_{t'}|X_{t'-1})$
- We only need to specify $P(X_0)$ and $P(X_t|X_{t-1})$
  - Simple, easy model to specify
  - Often the natural model
  - The network can extend indefinitely through time
- Consider: weather; stock prices

64

Variable Elimination Algorithm

- Given a Bayesian network, and an *elimination order* for the non-query variables

66

Variable Elimination Algorithm

- Given a Bayesian network, and an *elimination order* for the non-query variables, compute

$$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_m}\prod_{j}\Pr(x_j \mid Pa(x_j))$$

67

Variable Elimination Algorithm

- Given a Bayesian network, and an *elimination order* for the non-query variables, compute

$$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_m}\prod_{j}\Pr(x_j \mid Pa(x_j))$$

- For i = m downto 1

68

Variable Elimination Algorithm

- Given a Bayesian network, and an *elimination order* for the non-query variables, compute

$$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_m}\prod_{j}\Pr(x_j \mid Pa(x_j))$$

- For i = m downto 1
  - remove all the factors that mention X_i
  - multiply those factors, getting a value for each combination of mentioned variables

69

Variable Elimination Algorithm

- Given a Bayesian network, and an *elimination order* for the non-query variables, compute

$$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_m}\prod_{j}\Pr(x_j \mid Pa(x_j))$$

- For i = m downto 1
  - remove all the factors that mention X_i
  - multiply those factors, getting a value for each combination of mentioned variables
  - sum over X_i

70

Variable Elimination Algorithm

- Given a Bayesian network, and an *elimination order* for the non-query variables, compute

$$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_m}\prod_{j}\Pr(x_j \mid Pa(x_j))$$

- For i = m downto 1
  - remove all the factors that mention X_i
  - multiply those factors, getting a value for each combination of mentioned variables
  - sum over X_i
  - put this new factor into the factor set

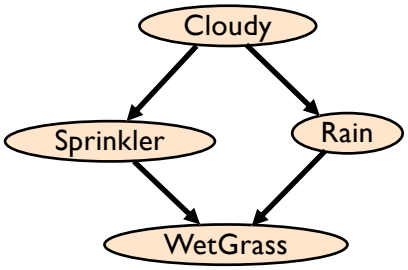
71

Variable Elimination: Example

$$P(w) = \sum_{r,s,c} P(w \mid r,s)P(r \mid c)P(s \mid c)P(c)$$
$$= \sum_{r,s} P(w \mid r,s)\sum_{c} P(r \mid c)P(s \mid c)P(c)$$
$$= \sum_{r,s} P(w \mid r,s)f_1(r,s)$$

"factors"

$f_1(r,s)$



72

Factors

- A **factor** is a function from a tuple of random variables to real numbers
- We write a factor as $f(X_1,...X_j)$
- Example:
- P(Z|X,Y) is a factor f (X,Y,Z)
  - Factors do not have to sum to one
  - P(Z|X,Y) is a set of probability distributions: one for each combination of values of X and Y
- P(Z=f|X,Y) is a factor f(X,Y)

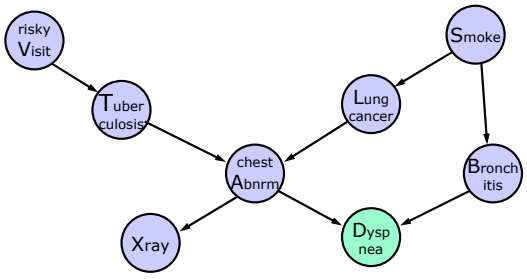
$f(X, Y)_{Z=f}$

X	Y	Z	val
t	t	t	0.1
t	t	f	0.9
t	f	t	0.2
t	f	f	0.8
f	t	t	0.4
f	t	f	0.6
f	f	t	0.3
f	f	f	0.7

Poole & Mackworth

73

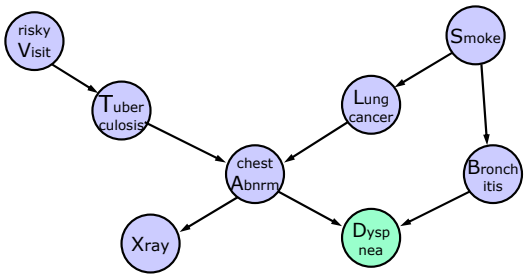
A More Complex Example



drawn from ocw.mit.edu/courses/6-825-techniques-in-artificial-intelligence-sma-5504-fall-2002

74

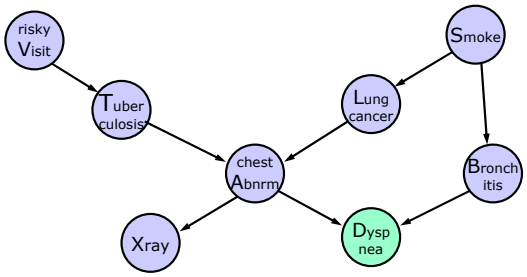
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T,S,X,V} \Pr(d \mid a,b)\Pr(a \mid t,l)\Pr(b \mid s)\Pr(l \mid s)\Pr(s)\Pr(x \mid a)\Pr(t \mid v)\Pr(v)$$

75

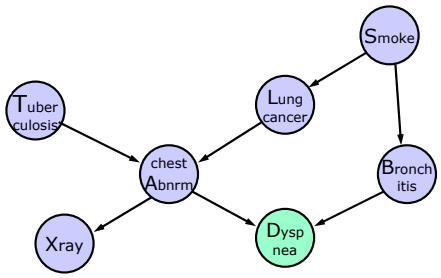
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T,S,X} \Pr(d \mid a,b)\Pr(a \mid t,l)\Pr(b \mid s)\Pr(l \mid s)\Pr(s)\Pr(x \mid a)\underbrace{\sum_{V}\Pr(t \mid v)\Pr(v)}_{f_1(t)}$$

76

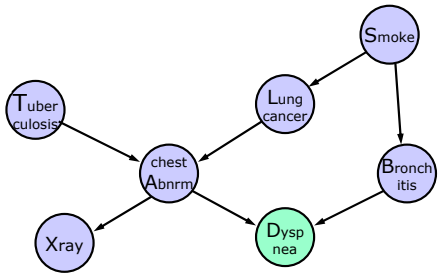
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T,S,X} \Pr(d \mid a,b)\Pr(a \mid t,l)\Pr(b \mid s)\Pr(l \mid s)\Pr(s)\Pr(x \mid a) f_1(t)$$

77

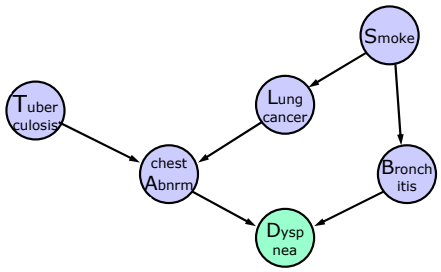
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T,S} \Pr(d \mid a,b)\Pr(a \mid t,l)\Pr(b \mid s)\Pr(l \mid s)\Pr(s) f_1(t) \underbrace{\sum_X \Pr(x \mid a)}_{1}$$

78

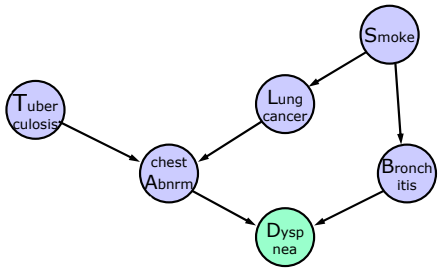
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T,S} \Pr(d \mid a,b) \Pr(a \mid t,l) \Pr(b \mid s) \Pr(l \mid s) \Pr(s) f_1(t)$$

79

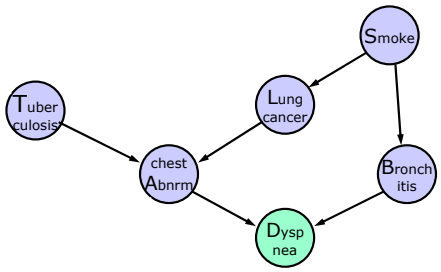
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T} \Pr(d \mid a,b) \Pr(a \mid t,l) f_1(t) \underbrace{\sum_{S} \Pr(b \mid s) \Pr(l \mid s) \Pr(s)}$$

80

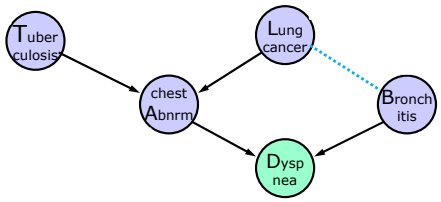
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T} \Pr(d \mid a,b) \Pr(a \mid t,l) f_1(t) \underbrace{\sum_{S} \Pr(b \mid s) \Pr(l \mid s) \Pr(s)}_{f_2(b,l)}$$

81

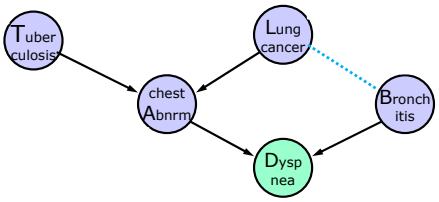
A More Complex Example



$$\Pr(d) = \sum_{A,B,L,T} \Pr(d \mid a,b) \Pr(a \mid t,l) f_1(t)\ f_2(b,l)$$

83

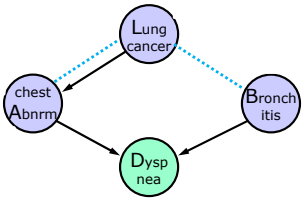
A More Complex Example



$$\Pr(d) = \sum_{A,B,L} \Pr(d \mid a,b) f_2(b,l) \underbrace{\sum_{T} \Pr(a \mid t,l) f_1(t)}_{f_3(a,l)}$$

84

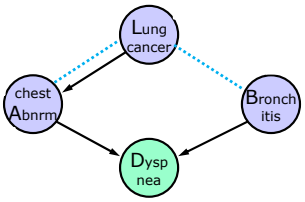
A More Complex Example



$$\Pr(d) = \sum_{A,B,L} \Pr(d \mid a,b) f_2(b,l) f_3(a,l)$$

85

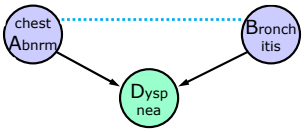
A More Complex Example



$$\Pr(d) = \sum_{A,B} \Pr(d \mid a,b) \underbrace{\sum_{L} f_2(b,l) f_3(a,l)}_{f_4(a,b)}$$

86

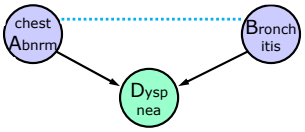
A More Complex Example



$$\Pr(d) = \sum_{A,B} \Pr(d \mid a,b) \underbrace{\sum_{L} f_2(b,l) f_3(a,l)}_{f_4(a,b)}$$

87

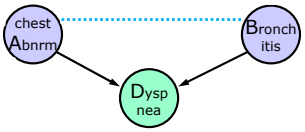
A More Complex Example



$$\Pr(d) = \sum_{A,B} \Pr(d \mid a,b) f_4(a,b)$$

88

A More Complex Example



$$\Pr(d) = \sum_{A} \underbrace{\sum_{B} \Pr(d \mid a,b) f_4(a,b)}_{f_5(a)}$$

89

A More Complex Example

$$\Pr(d) = \sum_{A} f_5(a)$$

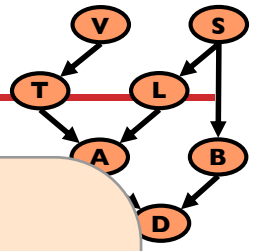
90

Properties of Variable Elimination

- Time is exponential in size of largest factor
- Bad elimination order can generate huge factors
- NP Hard to find the best elimination order
- Even the best elimination order may generate large factors
- There are reasonable heuristics for picking an elimination order (such as choosing the variable that results in the smallest next factor)
- Inference in polytrees (nets with no cycles) is linear in size of the network (the largest CPT)

91

Lungs Finale



- We want to compute *P(d)*
- Need

Initial f

Elimina

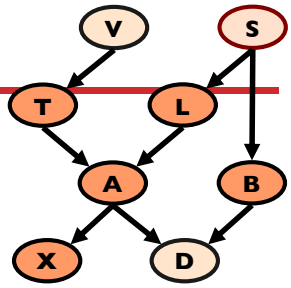
Compu

Variable elimination is just enumeration with caching of local calculations!

Order in which variables are eliminated can strongly affect efficiency.

100

Dealing with Evidence



- How do we deal with evidence?
  - And what is "evidence?"
  - Variables whose value has been observed

102

Evidence: The Core Idea

- We can make **new factors** out of an existing factor
- By **assigning** some or all of the variables

X	Y	Z	val
t	t	t	0.1
t	t	f	0.9
t	f	t	0.2
t	f	f	0.8
f	t	t	0.4
f	t	f	0.6
f	f	t	0.3
f	f	f	0.7

f(X,Y,Z):

X	Y	Z	val
t	t	t	0.1
t	t	f	0.9
t	f	t	0.2
t	f	f	0.8
~~f~~	~~t~~	~~t~~	~~0.4~~
~~f~~	~~t~~	~~f~~	~~0.6~~
~~f~~	~~f~~	~~t~~	~~0.3~~
~~f~~	~~f~~	~~f~~	~~0.7~~

What is the result of assigning X= t ?

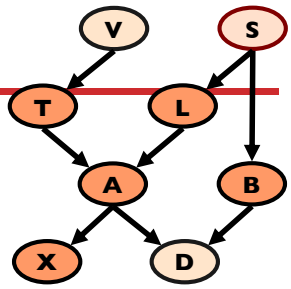
$f(X=t,Y,Z) = f(X, Y, Z)_{X=t}$

Y	Z	val
t	t	0.1
t	f	0.9
f	t	0.2
f	f	0.8

Factor of Y, Z

103

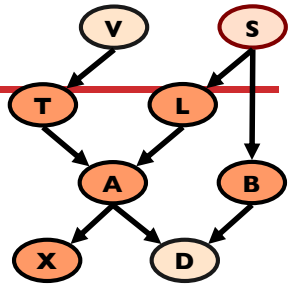
Dealing with Evidence



- How do we deal with evidence?
  - And what is "evidence?"
  - Variables whose value has been observed
- Suppose we are given evidence: $V = t, S = f, D = t$
- We want to compute $P(L, V = t, S = f, D = t)$
- (What is the probability of lung cancer, given a visit to someplace risky, no history of smoking, and dyspnea?)

104

Dealing with Evidence



- We start by writing the factors:

$$P(v)P(s)P(t\,|\,v)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

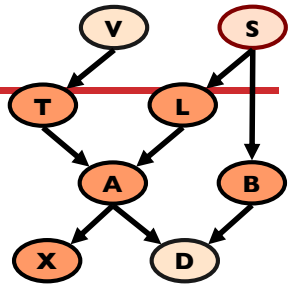
- Since we know that V = t, we don't need to eliminate V
- Instead, we can replace the factors P(V) and P(T |V) with

$$f_{P(V)} = P(V = t) \quad f_{p(T|V)}(T) = P(T\,|\,V = t)$$

- Note that $f_{P(V)}$ is a constant, so does not appear in elimination of other variables

105

Dealing with Evidence

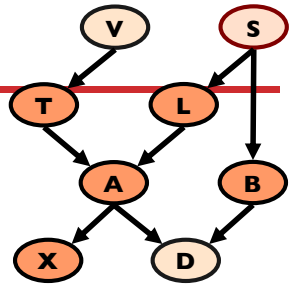


- So now…
  - Given evidence $V = t,\ S = f,\ D = t$
  - Compute $P(L,\ V = t,\ S = f,\ D = t\ )$
  - Initial factors, after setting evidence:

$$f_{P(v)}f_{P(s)}f_{P(t|v)}(t)f_{P(l|s)}(l)f_{P(b|s)}(b)P(a\,|\,t,l)P(x\,|\,a)f_{P(d|a,b)}(a,b)$$

106

Dealing with Evidence



- Given evidence $V = t$, $S = f$, $D = t$, we want to compute $P(L, V = t, S = f, D = t)$
- Initial factors, after setting evidence:

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a \mid t,l) P(x \mid a) f_{P(d|a,b)}(a,b)$$

- Eliminating x, we get

$$f_{P(v)} f_{P(s)} f_{P(t|v)}(t) f_{P(l|s)}(l) f_{P(b|s)}(b) P(a \mid t,l) f_x(a) f_{P(d|a,b)}(a,b)$$

- Eliminating t, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_{P(b|s)}(b) f_t(a,l) f_x(a) f_{P(d|a,b)}(a,b)$$

- Eliminating a, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_{P(b|s)}(b) f_a(b,l)$$

- Eliminating b, we get

$$f_{P(v)} f_{P(s)} f_{P(l|s)}(l) f_b(l)$$

$$P(L, V = t, S = f, D = t) = f_{P(v)} f_{P(s)} f_b(l) f_{P(l|s)}(l)$$

107

Learning CPTs from Data

- Given a data set, can you learn what a Bayesian network with variables A, B, C and D would look like?

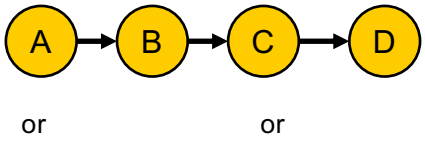
Say we collect a lot of **samples:**

A	B	C	D
true	false	false	true
true	false	true	false
true	false	false	true
false	true	false	false
false	true	false	true
false	true	false	false
false	true	false	false
:	:	:	:

109

Learning Bayesian Networks from Data

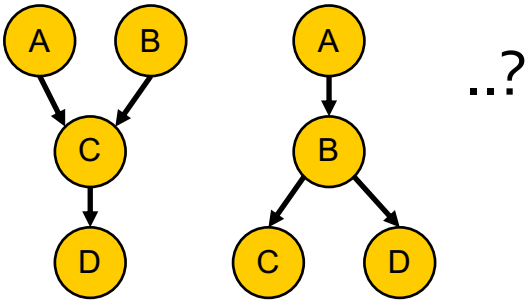
- Given a data set, can you learn what a Bayesian network with variables A, B, C and D would look like?



or or or

Say we collect a lot of **samples:**

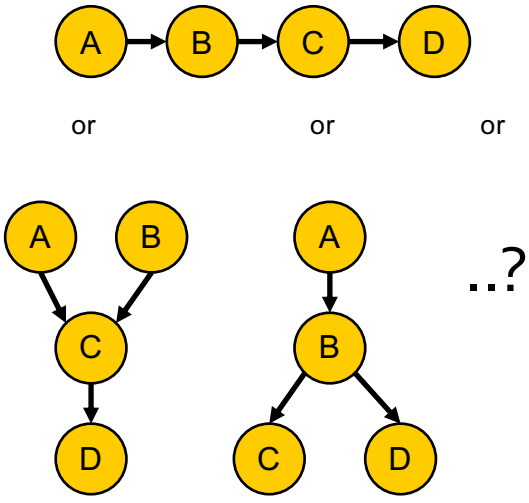
A	B	C	D
true	false	false	true
true	false	true	false
true	false	false	true
false	true	false	false
false	true	false	true
false	true	false	false
false	true	false	false
:	:	:	:



..?

110

Learning Bayesian Networks from Data



or or or

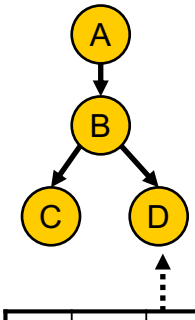
..?

- Each possible structure contains information about the conditional independence relationships between A, B, C and D
- We would like a structure that contains conditional independence relationships that are supported by the data
- Note that we still need to learn the values in the CPTs from data

111

Learning Bayesian Networks from Data

- How does Bayesian statistics help?



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

1. I might have a prior belief about what the structure should look like.
2. I might have a prior belief about what the values in the CPTs should be.

These beliefs get updated as I see more data

112

Learning Bayesian Networks from Data

- We won't have enough time to describe how we actually learn Bayesian networks from data ☹
- Some references:
  - Gregory F. Cooper and Edward Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 9:309-347, 1992.
  - David Heckerman. A Tutorial on Learning Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research. 1995. (Available online)

113

Other Inference Methods

- Convert network to a polytree
  - In a polytree no two nodes have more than one path between them
  - For such a graph there is a linear time algorithm
  - However, converting into a polytree requires a large increase in the size of the graph (number of nodes)
- Why is inference in polytrees easy?
- Given a variable X we can always divide the other variables into two sets:
  - E+: Variables 'above' X
  - E-: Variables 'below' X
- These sets are mutually exclusive
- Using these sets we can efficiently compute conditional and joint distributions

114

Summary

- Bayes nets
  - Structure
  - Parameters
  - Conditional independence
  - Chaining
- BN inference
  - Enumeration
  - Variable elimination
  - Sampling methods..?

115

Constructing a BBN: Paper Delivery

- Bill has noticed that his morning newspaper delivery has been sporadic
- Delivery is dependent on the paper having been successfully printed the previous night
- Possible explanations for a paper not having been printed are a malfunction at the printing press, or the end of civilization as we know it
- The prior probability of a printer malfunction is 0.05
- Bill expects the end of civilization with a probability of 0.001 (!)
- If the end of civilization is here, then the paper not be printed
- If there is a printing malfunction and no end of civilization, there is a probability of 0.05 that the paper will be printed
- If there is no malfunction and no end of civilization, there is a probability of 0.99 that the paper will be printed
- If the paper is not printed it will not be delivered
- If it is printed, there is a probability of 0.9 that it will be delivered

What variables should we include?

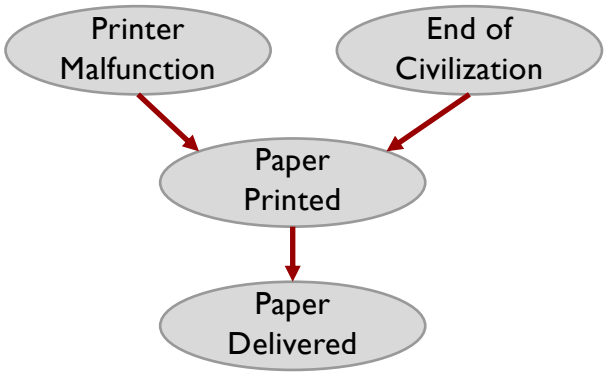
What does our BN look like?

What do the (complete) CPTs look like?

www.cs.ubc.ca/~mack/CS322/exercises/11_ex-varelim-sol.pdf

116

Constructing a BBN: Paper Delivery

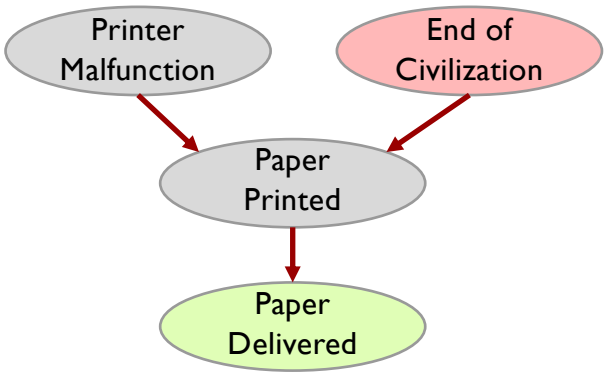


- Bill's paper fails to arrive one morning. He'd like to know the probability that civilization has ended. What are the observed variable(s), and what is the query variable(s)?

www.cs.ubc.ca/~mack/CS322/exercises/11_ex-varelim-sol.pdf

117

Constructing a BBN: Paper Delivery



The observed variable is Paper Delivered and the query is End of Civilization

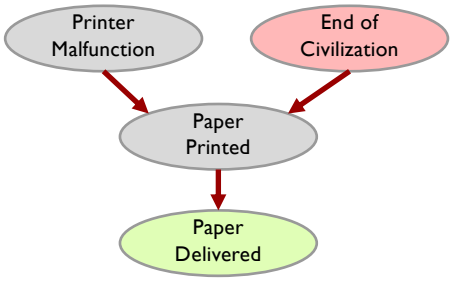
- Bill's paper fails to arrive one morning. He'd like to know the probability that civilization has ended. What are the observed variable(s), and what is the query variable(s)?

www.cs.ubc.ca/~mack/CS322/exercises/11_ex-varelim-sol.pdf

118

Inference in a BBN: Paper Delivery

- What are the initial factors (prior to variable elimination)?
  - f_0(Malfunction), f_1(EndOfCiv), f_2(Malfunction,endOfCiv,paperPrinted), f_3(paperPrinted,paperDelivered)
- Carry out variable elimination for this inference problem. Specify the elimination ordering you are using. Show each step of your work.

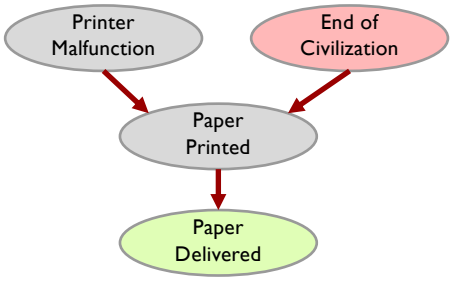


www.cs.ubc.ca/~mack/CS322/exercises/11_ex-varelim-sol.pdf

119

Inference in a BBN: Paper Delivery

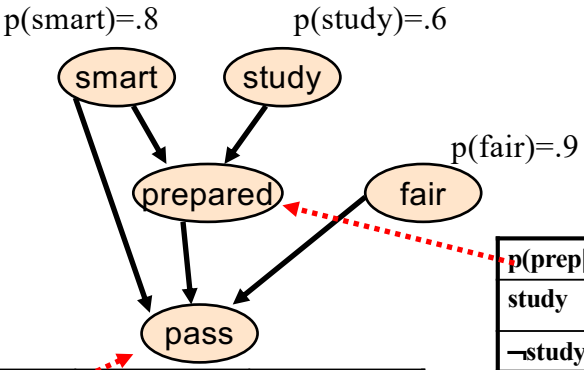
- f_0(Malfunction), f_1(EndOfCiv), f_2(Malfunction,endOfCiv,paperPrinted), f_3(paperPrinted,paperDelivered)
- First assign the value *false* to paperDelivered. f_3 is removed and a new factor f_4(paperPrinted) is added.
- We can next sum out Malfunction, which removes factors f_0 and f_2. A new factor f_5(EndOfCiv,paperPrinted) is added.
- If we next sum out paperPrinted, then f_4 and f_5 are removed and f_6(EndOfCiv) is added.
- At this point we are left with just two factors, f_1(EndOfCiv) and f_6(EndOfCiv). Note that f_1 corresponds to the prior for that variable.
- We then just need to multiply those final factors f_1 and f_6 to get f_7 and normalize to get f_8.



www.cs.ubc.ca/~mack/CS322/exercises/11_ex-varelim-sol.pdf

120

Exercise: Variable Elimination



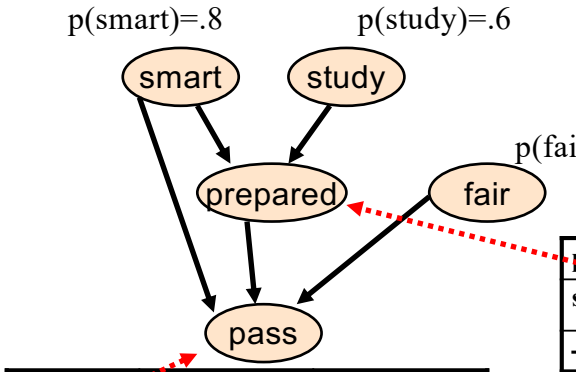
p(prep|…)	smart	¬smart
study	.9	.7
¬study	.5	.1

p(pass|…)	smart		¬smart	
	prep	¬prep	prep	¬prep
fair	.9	.7	.7	.2
¬fair	.1	.1	.1	.1

Query: What is the probability that a student **studied**, given that they pass the exam?

121

Exercise: Variable Elimination



p(prep|…)	smart	¬smart
study	.9	.7
¬study	.5	.1

p(pass|…)	smart		¬smart	
	prep	¬prep	prep	¬prep
fair	.9	.7	.7	.2
¬fair	.1	.1	.1	.1

Query: What is the probability that a student **is smart**, given that they pass the exam?

122