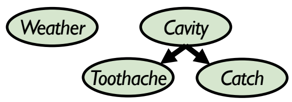
Bayes Nets

AI Class 10 (Ch. 14.1–14.4.2; skim 14.3)



Based on slides by Dr. Marie desJardin. Some material also adapted from slides by Matt E. Taylor @ WSU, Lise Getoor @ UCSC, Dr. P. Matuszek @ Villanova University, and Weng-Keen Wong at OSU. Based in part on www.csc.calpoly.edu/~fkurfess/Courses/CSC-481/W02/Slides/Uncertainty.ppt .

1

Bookkeeping

- HW3 out at HW4 time
  - We have to sort out problems with 1 and 2.
  - Can you see your Hwk1 annotations?
  - There will only be 5 homeworks
- This lecture: Bayes, Bayes, Bayes
- Next lecture: Games 2, Uncertain Reasoning
  - Presented by Pat

 2

2

Probability

- 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)

3

Bayes' Rule

- Derive the probability of some **event**, *given* **another event**
  - Assumption of attribute independency (AKA the Naïve assumption)
  - Naïve Bayes assumes that all *attributes* are independent.
- Also the basis of modern machine learning
- Bayes' rule is derived from the product rule

R&N 495

 4

4

Bayes' Rule

- $P(Y \mid X) = P(X \mid Y)\, P(Y) \,/\, P(X)$
- Often useful for **diagnosis**.
- If we have:
  - X = (observable) effects, e.g., symptoms
  - Y = (hidden) causes, e.g., illnesses
  - A model for how causes lead to effects: $P(X \mid Y)$
  - Prior beliefs about frequency of occurrence of effects: $P(Y)$
- We can reason from effects to causes: $P(Y \mid X)$

 5

5

Naïve Bayes Algorithm

- Estimate the probability of each class:
  - Compute the posterior probability (Bayes rule)

$$P(c_i \mid D) = \frac{P(c_i)P(D \mid c_i)}{P(D)}$$

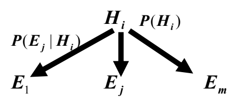
  - Choose the class with the highest probability
- Assumption of attribute independency (Naïve assumption): Naïve Bayes assumes that all of the attributes are independent.

 6

6

Bayesian Inference

- In the setting of diagnostic/evidential reasoning



hypotheses

evidence/manifestations

  - Know: prior probability of hypothesis $P(H_i)$
    conditional probability $P(E_j \mid H_i)$
  - Want to compute the *posterior probability* $P(H_i \mid E_j)$
- Bayes' theorem (formula 1):

$$P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)$$

Simple Bayesian Diagnostic Reasoning

- We know:
  - Evidence / manifestations: $E_1, \ldots E_m$
  - Hypotheses / disorders: $H_1, \ldots H_n$
    - E_j and H_i are **binary**; hypotheses are **mutually exclusive** (non-overlapping) and **exhaustive** (cover all possible cases)
  - Conditional probabilities: $P(E_j \mid H_i)$, i = 1, … n; j = 1, … m
- Cases (evidence for a particular instance): $E_1, \ldots, E_m$
- Goal: Find the hypothesis H_i with the highest posterior
  - $\text{Max}_i\ P(H_i \mid E_1, \ldots, E_m)$

Priors

- Four values total here:
  - P(H|E) = (P(E|H) * P(H)) / P(E)
- P(H|E) — what we want to compute
- Three we already know, called the *priors*
  - P(E|H)
  - P(H)
  - P(E)

(In ML we use the training set to estimate the priors)

Bayesian Diagnostic Reasoning II

- Bayes' rule says that
  - $P(H_i \mid E_1, \ldots, E_m) = P(E_1, \ldots, E_m \mid H_i)\, P(H_i) / P(E_1, \ldots, E_m)$
- Assume each piece of evidence E_i is conditionally independent of the others, **given** a hypothesis H_i, then:
  - $P(E_1, \ldots, E_m \mid H_i) = \prod^{l}_{j=1} P(E_j \mid H_i)$
- If we only care about relative probabilities for the H_i, then we have:
  - $P(H_i \mid E_1, \ldots, E_m) = \alpha\, P(H_i) \prod^{l}_{j=1} P(E_j \mid H_i)$

Bayes Example: Diagnosing Meningitis

$$P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)$$

- Your patient comes in with a stiff neck.
- Is it meningitis?
- Suppose we know that
  - Stiff neck is a symptom in 50% of meningitis cases
  - Meningitis (m) occurs in 1/50,000 patients
  - Stiff neck (s) occurs in 1/20 patients
- So probably not. But specifically?

Bayes Exercise: Diagnosing Meningitis

$$P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)$$

- Stiff neck is a symptom in 50% of meningitis cases
- Meningitis (m) occurs in 1/50,000 patients
- Stiff neck (s) occurs in 1/20 patients

Analysis of Naïve Bayes Algorithm

- Advantages:
  - Sound theoretical basis
  - Works well on numeric and textual data
  - Easy implementation and computation
  - Has been effective in practice (e.g., typical spam filter)

13

Limitations of Naïve Bayes

- Cannot easily handle:
  - Multi-fault situations
  - Cases where intermediate (hidden) causes exist:
    - Disease D causes syndrome S, which causes correlated manifestations M_1 and M_2

14

Limitations of Naïve Bayes

- Consider a composite hypothesis $H_1 \wedge H_2$, where H_1 and H_2 are independent. What is the relative posterior?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_m) = \alpha P(E_1, \ldots, E_m \mid H_1 \wedge H_2) P(H_1 \wedge H_2)$

    $= \alpha P(E_1, \ldots, E_m \mid H_1 \wedge H_2) P(H_1) P(H_2)$

    $= \alpha \prod^{l}_{m=1} P(E_m \mid H_1 \wedge H_2) P(H_1) P(H_2)$
- How do we compute $P(E_j \mid H_1 \wedge H_2)$??

15

Limitations of Simple Bayesian Inference II

- Assume H_1 and H_2 are independent, given $E_1, \ldots, E_j$?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_j) = P(H_1 \mid E_1, \ldots, E_j) P(H_2 \mid E_1, \ldots, E_j)$
- This is a very unreasonable assumption
  - Earthquake and Burglar are independent, but *not* given Alarm:
    - P(burglar | alarm, earthquake) << P(burglar | alarm)
- Simple application of Bayes' rule doesn't handle causal chaining:
  - A: this year's weather; B: cotton production; C: next year's cotton price
  - A influences C indirectly: $A \rightarrow B \rightarrow C$
  - $P(C \mid B, A) = P(C \mid B)$

16

Beyond Simple Bayes

- Need a richer representation to model:
  - Interacting hypotheses
  - Conditional independence
  - Causal chaining
- So: **conditional independence** and **Bayesian networks**!

17

Next Up

- Bayesian networks
  - Network structure
  - Conditional probability tables
  - Conditional independence
- Inference in Bayesian networks
  - Exact inference
  - Approximate inference

18

Review: Independence

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

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

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**
- $P(A \wedge B \mid C) = P(A \mid C)\, P(B \mid C)$

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

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

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

Bayes' Nets: Big Picture

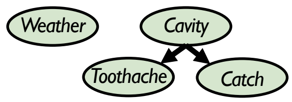
- Problems with full joint distribution tables as our probabilistic models:
  - Joint gets way too big to represent explicitly
    - Unless there are only a few variables
  - Hard to learn (estimate) anything empirically about more than a few variables at a time
    - Why?

	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, WSU

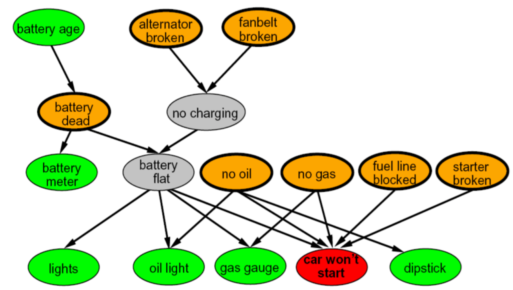
Bayes' Nets: Big Picture

- **Bayes' nets:** a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - A type of graphical models
- We describe how variables interact **locally**
  - Local interactions chain together to give global, indirect interactions



28

Example: Car Won't Start



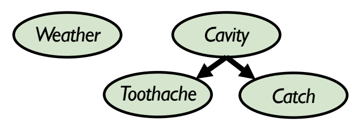
29

Example: Insurance

30

Example: Toothache

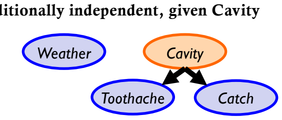
- Random variables:
  - How's the weather?
  - Do you have a toothache?
  - Does the dentist's probe catch when she pokes your tooth?
  - Do you have a cavity?



31

Graphical Model Notation

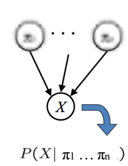
- Nodes: variables (with domains)
- Can be assigned (observed) or unassigned (hidden)
- Arcs: interactions
  - Indicate "direct influence" between
  - Formally: **encode conditional independence**
    - Toothache and Catch are **conditionally independent, given Cavity**
- For now: imagine that arrows mean *causation*
  - (in general, they don't!)



32

Bayesian Belief Networks (BNs)

- Let's formalize the semantics of a BN
  - A set of nodes, one per variable X
- A directed arc between each co-influential node
  - $X \rightarrow Y$ means X **has an influence on** Y
- A directed, acyclic graph



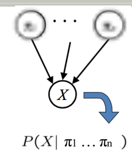
33

Bayesian Belief Networks (BNs)

- Each node X has a **conditional probability distribution**:

  $P(X_i \mid Parents(X_i))$

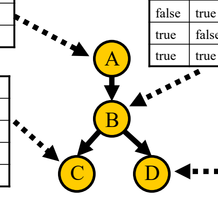
  - A collection of distributions over X
  - One for each combination of parents' values
  - Quantifies the effects of the parents on a node
- CPT: conditional probability table
  - Description of a noisy "causal" process



34

Conditional Probability Tables

- For X_i, CPD $P(X_i \mid Parents(X_i))$ quantifies effect of parents on X_i
- **Parameters** are probabilities in conditional probability tables (CPTs):



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

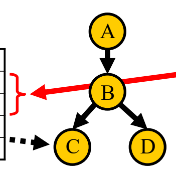
B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

35

CPTs cont'd

- Conditional Probability Distribution for C given B
- If you have a Boolean variable with k Boolean parents, this table has 2^{k+1} probabilities

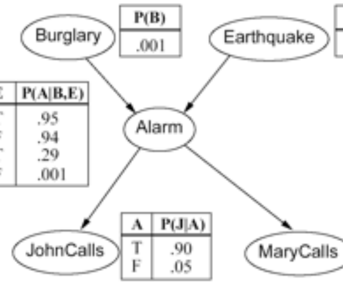
B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



For a given combination of values of the parents (B in this example), the entries for P(C=true | B) and P(C=false | B) must sum to 1
Example:
P(C=true | B=false) +
P(C=false | B=false) = 1

36

Bayesian Belief Networks (BNs)



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

37

Bayesian Belief Networks (BNs)

- Making a BN: **BN = (DAG, CPD)**
  - **DAG**: directed acyclic graph (BN's **structure**)
    - **Nodes**: random variables
      - Typically binary or discrete
      - Methods exist for continuous variables
    - **Arcs**: indicate probabilistic dependencies between nodes
      - *Lack* of link signifies conditional independence
  - **CPD**: conditional probability distribution (BN's **parameters**)
    - Conditional probabilities at each node, usually stored as a table (conditional probability table, or **CPT**)

38

Bayesian Belief Networks (BNs)

- Making a BN: **BN = (DAG, CPD)**
  - **DAG**: directed acyclic graph (BN's **structure**)
  - **CPD**: conditional probability distribution (BN's **parameters**)
    - Conditional probabilities at each node, usually stored as a table (conditional probability table, or **CPT**)

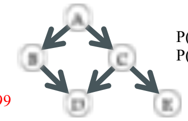
      $P(x_i \mid \pi_i)$ where π_i is the set of all parent nodes of x_i
  - Root nodes are a special case
    - No parents, so use priors in CPD:

      $\pi_i = \varnothing$, so $P(x_i \mid \pi_i) = P(x_i)$

39

Example BN

P(A) = 0.001



P(B|A) = 0.3
P(B|¬A) = 0.001
P(¬B|A) = 0.7
P(¬B|¬A) = 0.999

P(C|A) = 0.2
P(C|¬A) = 0.005

P(D|B,C) = 0.1
P(D|B,¬C) = 0.01
P(D|¬B,C) = 0.01
P(D|¬B,¬C) = 0.00001

P(E|C) = 0.4
P(E|¬C) = 0.002

We only specify P(A) etc., not P(¬A), since they have to sum to one

40

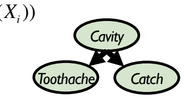
Probabilities in BNs

- Bayes' nets implicitly **encode joint distributions** as a **product of local conditional distributions**.
- To see probability of a **full assignment**, multiply all the relevant conditionals together:

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

  - Example:

$P(+\text{cavity}, +\text{catch}, \neg\text{toothache}) = ?$

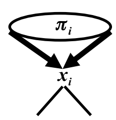


- This lets us reconstruct any entry of the full joint

41

Conditional Independence and Chaining

- Conditional independence assumption: $P(x_i \mid \pi_i, q) = P(x_i \mid \pi_i)$
  - ***q*** is any set of variables (nodes) other than x_i and its successors
  - π_i **blocks influence** of other nodes on x_i and its successors
  - That is, ***q*** influences x_i only through variables in π_i)



  - Then, complete joint probability distribution of all variables can be represented by local CPDs by chaining:

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

42

The Chain Rule

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

e.g, $P(x_1, \ldots, x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x2)\ldots$

- Decomposition:

  P(Traffic, Rain, Umbrella) =
  P(Rain) *P*(Traffic | Rain) *P*(Umbrella | Rain, Traffic)

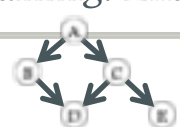
- With assumption of conditional independence:

  P(Traffic, Rain, Umbrella) =
  P(Rain) *P*(Traffic | Rain) *P*(Umbrella | Rain)

- Bayes' nets express conditional independences
  - (Assumptions)

44

Chaining: Example



By product rule

By conditional independence assumption

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)
= P(e | c) P(a, b, c, d)
= 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.

45

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**)
  - (For much later: a method called **d-separation** can be applied to decide whether a set of nodes X is independent of a set Y, given a third set Z)

46

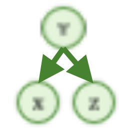
Independence and Causal Chains

- 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"

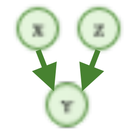
47

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**



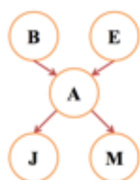
- Common Effect:
  - Two causes of one effect
  - Are X and Z independent? **Yes**
  - Are X and Z independent given Y?
  →**No**!
  - Observing an effect "*activates*" influence between possible causes.



48

Conditionality Example

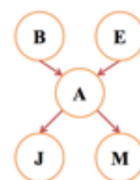
- Hidden: A, B, E. You don't know:
  - If there's a burglar.
  - If there was an earthquake.
  - If the alarm is going off.
- Observed: J and M.
  - John and/or Mary have some chance of calling if the alarm rings.
  - You know who called you.



49

Conditionality Example 2

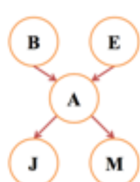
- At first:
  - Is the probability of John calling affected by whether there's an earthquake?
  - Is the probability of Mary calling affected by John calling?
- Your alarm *is* going off!
  - Is the probability of Mary calling affected by John calling?



50

Conditionality Example 3

- At first:
  - Is whether there's an earthquake affected by whether there's a burglary in progress (and vice versa)?
- Your alarm *is* going off!
  - Does the probability a burglary is happening depend on whether there's an earthquake?



51

Representational Extensions

- 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)

52

Bayes' Net Inference
Multi-Agent Systems

 53 *Some material borrowed from Lise Getoor*

53

Today's Class

- Bayes' nets inference
  - Inference by enumeration; by variable elimination
- Multi-agent systems
- Midterm study guide posted
- HW4 moved to after midterm
- Remember, Erfan is lecturing Thursday

 54

54

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?

 55

55

Direct Inference with BNs

- Instead of computing the joint, suppose we just want the probability for *one* variable.
- Exact methods of computation:
  - **Enumeration**
  - **Variable elimination**
  - **Join trees: get the probabilities associated with every query variable**

 57

57

Inference by Enumeration

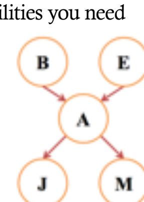
- Add all of the terms (atomic event probabilities) from the full joint distribution
- If **E** are the evidence (observed) variables and **Y** are the other (unobserved) variables, then:

  $P(X \mid \mathbf{E}) = \alpha\, P(X, \mathbf{E}) = \alpha \sum P(X, \mathbf{E}, \mathbf{Y})$
- Each P(X, **E**, **Y**) term can be computed using the chain rule
- Computationally expensive!

59

Example 1: Enumeration

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



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

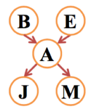
60

Example 1 cont'd

$P(+b,+j,+m) =$

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

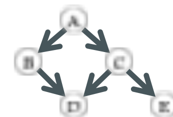
P(+m | +b, +e)?



61

Example 2: 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
- $P(d \mid e) = \alpha\, \Sigma_{ABC} P(a, b, c, d, e)$ *(where α = 1/P(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



62

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 NP-hard!**
- Join tree algorithms are an extension of variable elimination methods that compute posterior probabilities for **all** nodes in a BN simultaneously

63

Variable Elimination Approach

General idea:

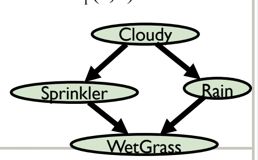
- Write query in the form
  $$P(X_n, e) = \sum_{x_k} \cdots \sum_{x_3} \sum_{x_2} \prod_i P(x_i \mid pa_i)$$
  - Note that there is no α term here
  - It's a conjunctive probability, not a conditional probability…
- Iteratively
  - Move all irrelevant terms outside of innermost sum
  - Perform innermost sum, getting a new term
  - Insert the new term into the product

64

Variable Elimination: Example

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

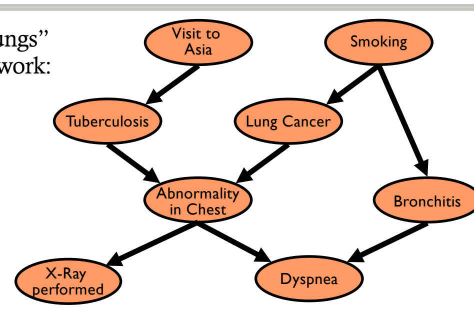
"factors"



65

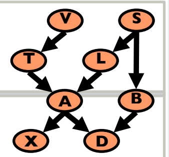
A More Complex Example

- "Lungs" network:



66

Lungs 1



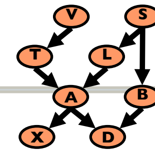
- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*

Initial factors:

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

67

Lungs 2



- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*

Initial factors:

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

Eliminate: *v*

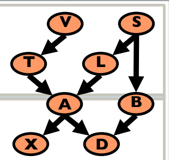
Compute: $f_v(t) = \sum_v P(v)P(t\mid v)$

$$\Rightarrow f_v(t)P(s)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

- Note: $f_v(t) = P(t)$
- Result of elimination is not **necessarily** a probability term

68

Lungs 3



- We want to compute *P(d)*
- Need to eliminate: *s,x,t,l,a,b*

Initial factors:

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

$$\Rightarrow f_v(t)P(s)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

Eliminate: *s*

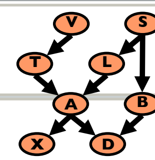
Compute: $f_s(b,l) = \sum_s P(s)P(b\mid s)P(l\mid s)$

$$\Rightarrow f_v(t)f_s(b,l)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

- Summing on *s* results in a factor with two arguments $f_s(b,l)$
- In general, result of elimination may be a function of several variables

69

Lungs 4



- We want to compute *P(d)*
- Need to eliminate: *x,t,l,a,b*

Initial factors

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

$$\Rightarrow f_v(t)P(s)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

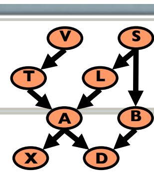
Eliminate: *x*

Compute: $f_x(a) = \sum_x P(x\mid a)$

$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a\mid t,l)P(d\mid a,b)$$

70

Lungs 5



- We want to compute *P(d)*
- Need to eliminate: *t,l,a,b*

Initial factors $P(v)P(s)P(t\mid v)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$

$$\Rightarrow f_v(t)P(s)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a\mid t,l)P(d\mid a,b)$$

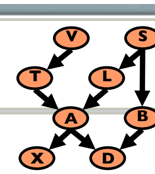
Eliminate: *t*

Compute: $f_t(a,l) = \sum_t f_v(t)P(a\mid t,l)$

$$\Rightarrow f_s(b,l)f_x(a)f_t(a,l)P(d\mid a,b)$$

71

Lungs 6



- We want to compute *P(d)*
- Need to eliminate: *l,a,b*

Initial factors $P(v)P(s)P(t\mid v)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$

$$\Rightarrow f_v(t)P(s)P(l\mid s)P(b\mid s)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a\mid t,l)P(x\mid a)P(d\mid a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a\mid t,l)P(d\mid a,b)$$

$$\Rightarrow f_s(b,l)f_x(a)f_t(a,l)P(d\mid a,b)$$

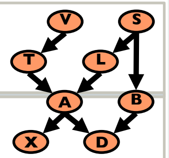
Eliminate: *l*

Compute: $f_l(a,b) = \sum_l f_s(b,l)f_t(a,l)$

$$\Rightarrow f_l(a,b)f_x(a)P(d\mid a,b)$$

72

Lungs Finale



- We want to compute *P(d)*
- Need to eliminate: *b*

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

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

$\Rightarrow f_v(t)f_s(b,l)P(a|t,l)P(x|a)P(d|a,b)$

$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a|t,l)P(d|a,b)$

$\Rightarrow f_s(b,l)f_x(a)f_t(a,l)P(d|a,b)$

$\Rightarrow \underline{f_l(a,b)}\,\underline{f_x(a)}\,\underline{P(d|a,b)} \Rightarrow \underline{f_a(b,d)} \Rightarrow \underline{f_b(d)}$

Eliminate: a,b

Compute: $f_a(b,d) = \sum_a f_l(a,b)f_x(a)p(d|a,b) \quad f_b(d) = \sum_b f_a(b,d)$

73

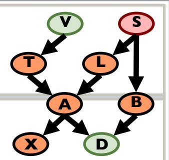
Computing Factors

R	S	C	P(R|C)	P(S|C)	P(C)	P(R|C) P(S|C) P(C)
T	T	T				
T	T	F				
T	F	T				
T	F	F				
F	T	T				
F	T	F				
F	F	T				
F	F	F				

R	S	$f_1(R,S) = \sum_c P(R\|S) P(S\|C) P(C)$
T	T	
T	F	
F	T	
F	F	

74

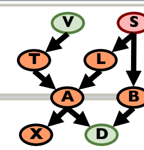
Dealing with Evidence



- 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 $P(L, V = t, S = f, D = t)$

75

Dealing with Evidence



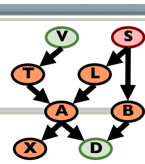
- We start by writing the factors:

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

  $f_{P(V)} = P(V = t) \quad f_{p(T|V)}(T) = P(T | V = t)$
- These "select" appropriate parts of original factors given evidence
- Note that $f_{P(V)}$ is a constant, so **does not appear** in elimination of other variables

76

Dealing with Evidence

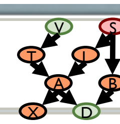


- So now…
  - Given evidence $V = t$, $S = f$, $D = t$
  - Compute $P(L, V = t, S = f, D = t)$
  - Initial factors, after setting evidence:

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

77

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:

  $f_{P(v)}f_{P(s)}f_{P(t|v)}(t)f_{P(l|s)}(l)f_{P(b|s)}(b)P(a|t,l)\underline{P(x|a)}f_{P(d|a,b)}(a,b)$
- Eliminating x, we get

  $f_{P(v)}f_{P(s)}\underline{f_{P(t|v)}}(t)f_{P(l|s)}(l)f_{P(b|s)}(b)\underline{P(a|t,l)}f_x(a)f_{P(d|a,b)}(a,b)$
- Eliminating t, we get

  $f_{P(v)}f_{P(s)}f_{P(l|s)}(l)f_{P(b|s)}(b)\underline{f_t(a,l)f_x(a)f_{P(d|a,b)}(a,b)}$
- Eliminating a, we get

  $f_{P(v)}f_{P(s)}f_{P(l|s)}(l)\underline{f_{P(b|s)}(b)f_a(b,l)}$
- Eliminating b, we get

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

78

Variable Elimination Algorithm

- Let $X_1, \ldots, X_m$ be an ordering on the non-query variables
- For i = m, …, 1 $\sum_{X_1}\sum_{X_2}\ldots\sum_{X_m}\prod_j P(X_j \mid Parents(X_j))$
  - In the summation for X_i, leave only factors mentioning X_i
  - Multiply the factors, getting a factor that contains a number for each value of the variables mentioned, including X_i
  - Sum out X_i, getting a factor f that contains a number for each value of the variables mentioned, not including X_i
  - Replace the multiplied factor in the summation

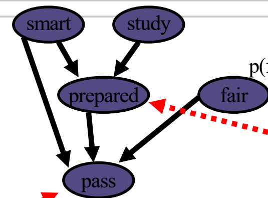
 79

79

Exercise: Enumeration

p(smart)=.8 p(study)=.6

p(fair)=.9



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

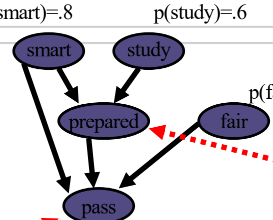
Query: What is the probability that a student **studied**, given that they pass the exam?

81

Exercise: Variable Elimination

p(smart)=.8 p(study)=.6

p(fair)=.9



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?

82