Artificial Intelligence

13. Bayesian Networks

Shashi Prabh

School of Engineering and Applied Science
Ahmedabad University

Reminder: Elementary Probability

- Basic laws: $0 \leq P(\omega) \leq 1$ $\quad \sum_{\omega \in \Omega} P(\omega) = 1$
- Events: subsets of Ω: $P(A) = \sum_{\omega \in A} P(\omega)$
- Random variable $X(\omega)$ has a value in each ω
  - Distribution $P(X)$ gives probability for each possible value x
  - Joint distribution $P(X,Y)$ gives total probability for each combination x,y
- Summing out/marginalization: $P(X=x) = \sum_y P(X=x,Y=y)$
- Conditional probability: $P(X|Y) = P(X,Y)/P(Y)$
- Product rule: $P(X|Y)P(Y) = P(X,Y) = P(Y|X)P(X)$
  - Generalize to chain rule: $P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,..,X_{i-1})$

Bayes' Rule

- The product rule both ways: $P(a \mid b)\, P(b) = P(a, b) = P(b \mid a)\, P(a)$
- Dividing left and right expressions, we get the Bayes' Rule

$$P(a \mid b) = \frac{P(b \mid a)\, P(a)}{P(b)}$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Describes an "update" step from prior $P(a)$ to posterior $P(a \mid b)$
  - Foundation of many AI systems
- In the running for the most important AI equation!

Inference with Bayes' Rule

- Diagnostic probability from causal probability or likelihood

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

- Example:
  - M: meningitis
  - S: stiff neck

$$\left.\begin{aligned} P(s \mid m) &= 0.8 \\ P(s \mid \neg m) &= 0.01 \\ P(m) &= 0.0001 \end{aligned}\right\} \text{Example givens}$$

$$P(m \mid s) = \frac{P(s \mid m)\, P(m)}{P(s)} \simeq \frac{0.8 \text{ x } 0.0001}{0.01}$$

- Posterior probability of meningitis still very small: 0.008
  - You should still get stiff necks checked out! Why?

Independence

- Two variables X and Y are (absolutely) **independent** if

$$\forall x,y \qquad P(x, y) = P(x)\, P(y)$$

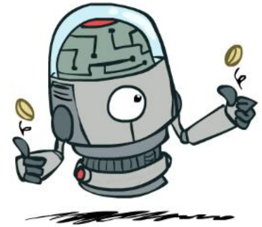
  - The joint distribution **factors** into a product of two simpler distributions

- Equivalently, via the product rule $P(x,y) = P(x|y)\, P(y)$,

$$P(x \mid y) = P(x) \quad \text{or} \quad P(y \mid x) = P(y)$$

Independence

- Example: two dice rolls $Roll_1$ and $Roll_2$
  - $P(Roll_1=5, Roll_2=3) = P(Roll_1=5)\ P(Roll_2=3) = 1/6 \times 1/6 = 1/36$
  - $P(Roll_2=3 \mid Roll_1=5) = P(Roll_2=3)$



Conditional Independence

- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

- X is conditionally independent of Y given Z if and only if:

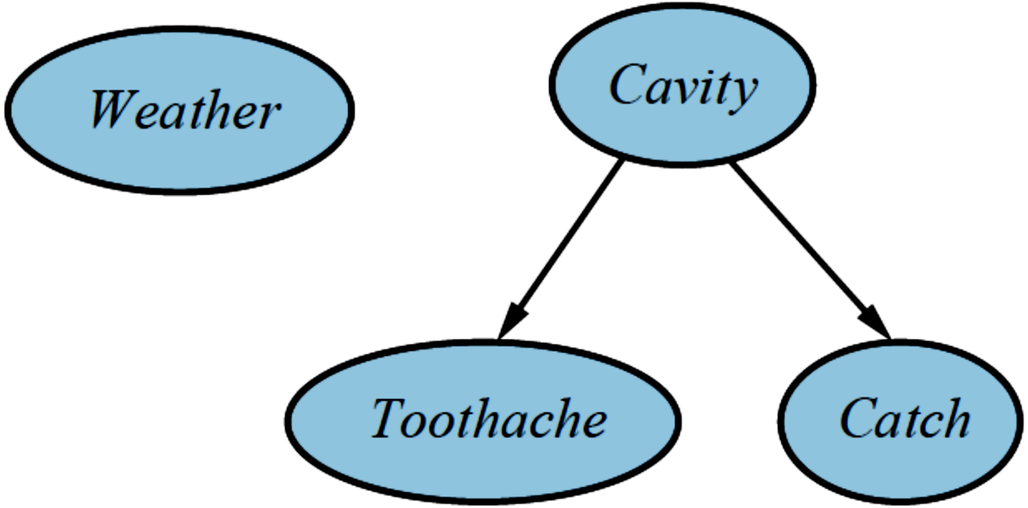
$$\forall x,y,z \qquad P(x \mid y, z) = P(x \mid z)$$
$$P(y \mid x, z) = P(y \mid z)$$

or, equivalently, if and only if

$$\forall x,y,z \qquad P(x, y \mid z) = P(x \mid z)\, P(y \mid z)$$

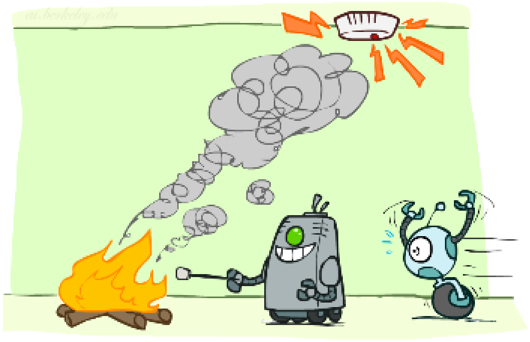
Conditional Independence

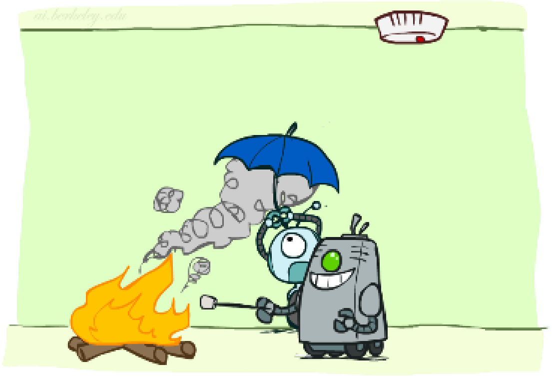
- Example
  - Cavity
  - Toothache
  - Catch
  - Weather



Conditional Independence

- What about this domain?
  - Fire
  - Smoke
  - Alarm



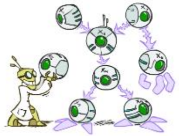


Bayesian Networks

Bayes Net Syntax and Semantics

Bayesian Networks (Bayes Nets)

- Full joint probability distribution can answer any query but at the cost of exponentially large joint probability tables
- Absolute and conditional independence among variables can greatly reduce the number of probabilities that need to be specified for defining the full joint distribution



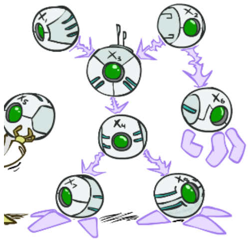
- Bayes nets, also called belief networks, is a data structure used to represent dependencies among variables
- A Bayes Net is a directed graph where each node is annotated with conditional probability distributions
  - A subset of the general class of probabilistic graphical models

Bayes Net Syntax

- A set of nodes, one per random variable X_i
  - Can be discrete or continuous
  - Can be assigned (observed) or unassigned (unobserved)
- Directed arrows connect node pairs in a Parent-Child relationship
  - Indicates "direct influence" between variables
  - Absence of arc encodes conditional independence (more later)
  - The resulting graph is a DAG

Bayes Net Syntax

- Each node has associated conditional probability distribution that quantifies the effects of its parents
- Local causality and conditional independence leads to compact representation of the joint distribution
  - Each variable interacting locally with a few others

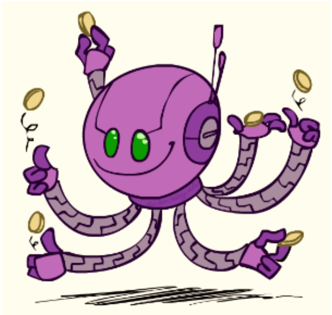


Bayes net = Topology (graph) + Local Conditional Probabilities

Example: Coin Flips

- N independent coin flips

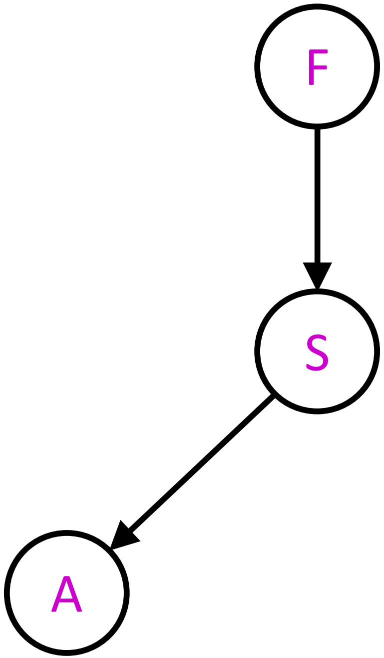
X_1 X_2 . . . X_n

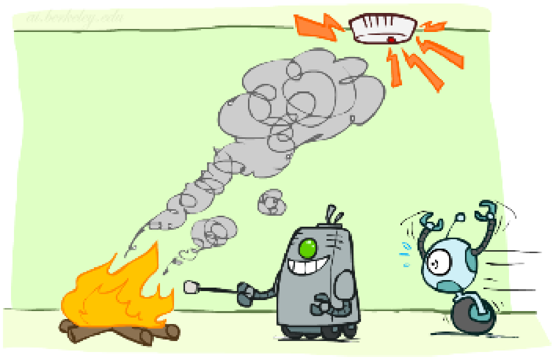


- No interactions between variables: absolute independence

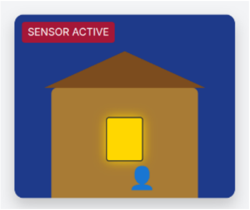
Example: Smoke alarm

- Variables:
  - F: There is fire
  - S: There is smoke
  - A: Alarm sounds





Example: IoT Network

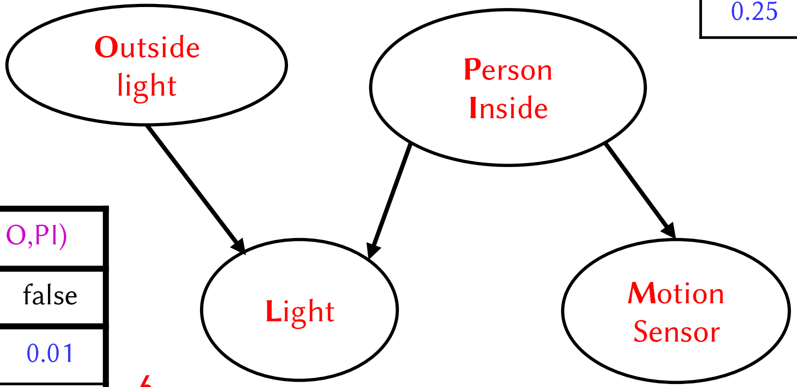


	P(O)	
dark	dim	bright
0.4	0.3	0.3

2

P(PI)	
yes	no
0.25	0.75

1



O	PI	P(L | O,PI)	
		true	false
dark	yes	0.99	0.01
dim	yes	0.80	0.20
bright	yes	0.10	0.90
*	no	0.01	0.99

6

PI	P(MS | PI)	
	true	false
true	0.95	0.05
false	0.01	0.99

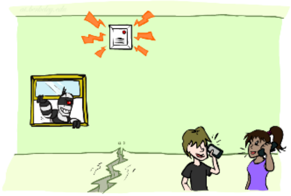
2

Show: $P(L) = \langle 0.1815, 0.8185 \rangle$

$P(PI \mid MS = y) = \langle 0.8333, 0.1667 \rangle$

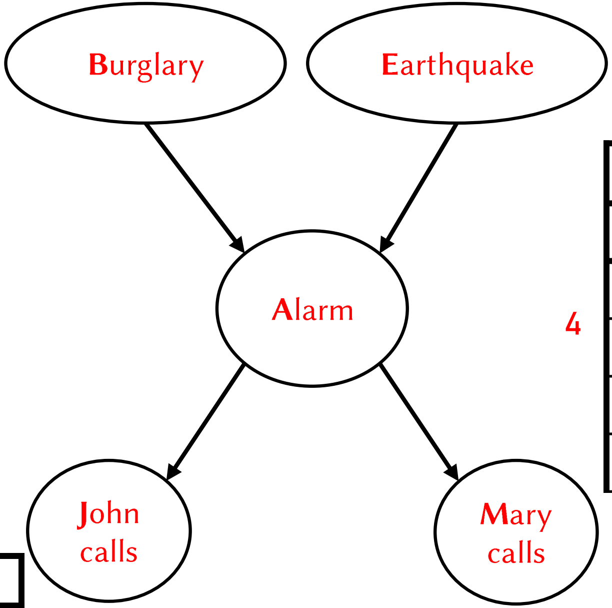
$P(L \mid MS = y, O = dark) = \langle 0.8287, 0.1733 \rangle$

Example: Alarm Network



P(B)	
true	false
0.001	0.999

1



1

P(E)	
true	false
0.002	0.998

4

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

2

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

Number of **free parameters** in each CPT:

1. Parent range sizes $d_1,\ldots,d_k$
2. Child range size d
3. Each row must sum to 1

$$(d-1)\ \Pi_i\ d_i$$

Sparse Bayes Nets

- Suppose
  - n variables
  - Maximum domain size is d
  - Maximum number of parents is k
- Then, full joint distribution has size $O(d^n)$
- But Bayes Net has size $O(n \cdot d^k)$
  - Linear scaling with n as long as causal structure is local
    - Called sparse networks

Bayes Net global semantics

- Bayes nets encode joint distributions as product of conditional distributions on each variable

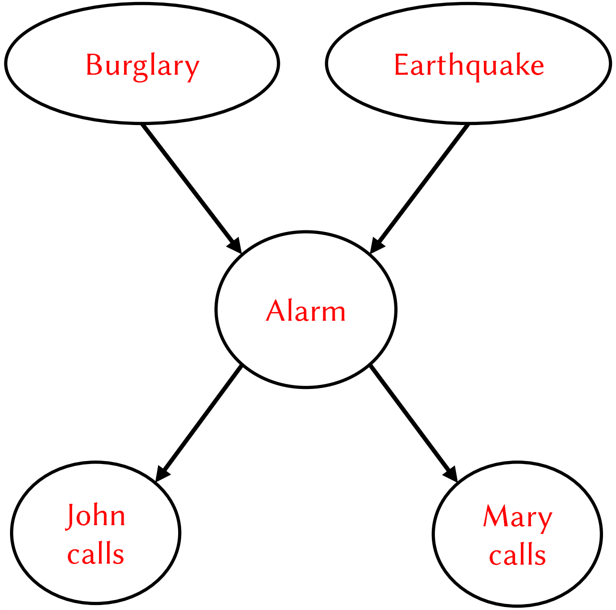
$$P(X_1,..,X_n) \;=\; \prod_i P(X_i \mid Parents(X_i))$$

Example

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

$$=.001x.998x.94x.1x.3=.000028$$

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

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

Conditional independence in BNs

- Let $X_1,..,X_n$ be sorted in topological order according to the graph, i.e., parents before children, so

  $$\text{Parents}(X_i) \subseteq X_1,\ldots,X_{i-1}$$

- So the Bayes net asserts conditional independences

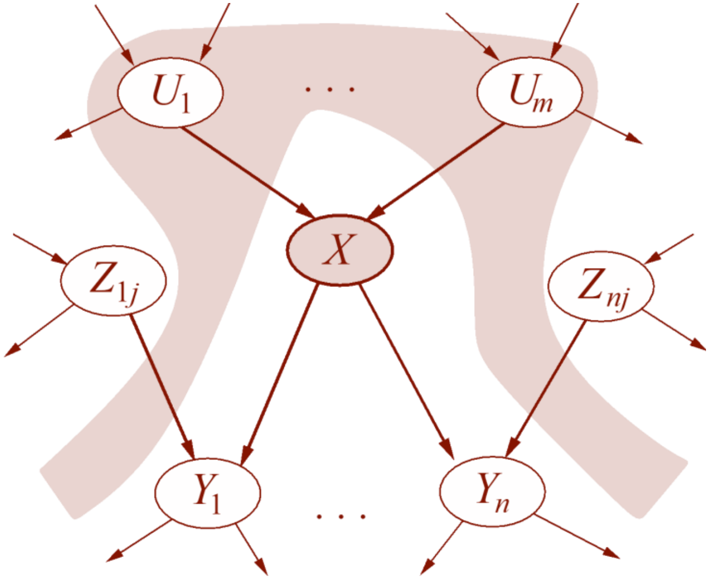
  $$P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid \text{Parents}(X_i))$$

  - To ensure these are valid, choose parents for node X_i that "shield" it from other predecessors

- $P(M \mid J, A, E, B) = P(M \mid A)$

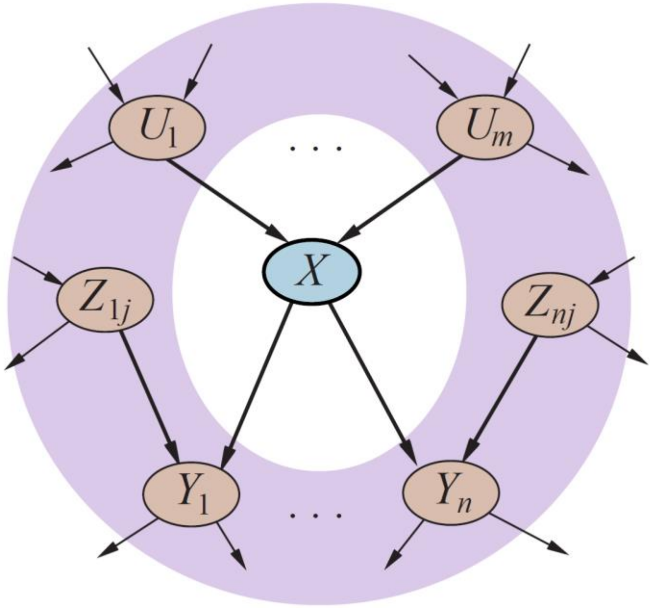
Conditional independence semantics

- Every variable is conditionally independent of its
  - Other predecessors given its parents
  - Non-descendants given its parents



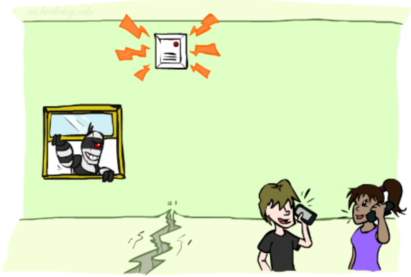
Conditional independence semantics

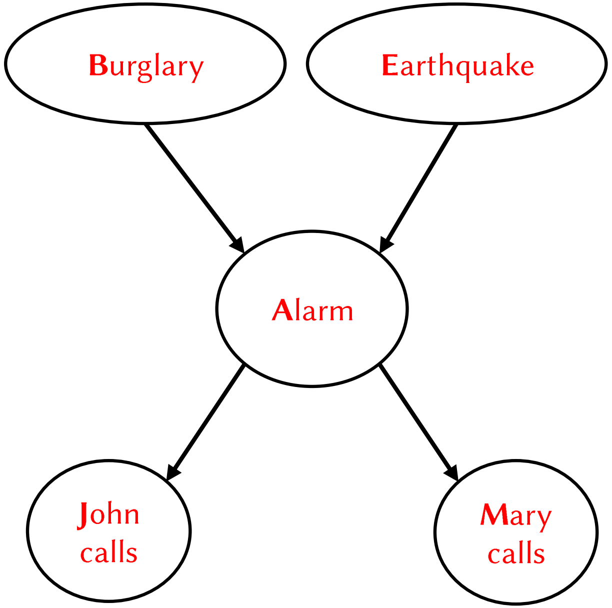
- Markov blanket of a node: parents, children and children's parents
- Every variable is conditionally independent of all other nodes given its Markov blanket
- d-Separation is yet another test
  - Moralize the graph
  - Test whether Z blocks all paths from X to Y. If yes, $X \perp Y \mid Z$



Example: Burglary

- $J \perp M \mid A$
  - Markov blanket of J includes A only
- B is not independent of E given A
- $B \perp J, M \mid A, E$





Quiz

1. What does a Bayesian network use to represent dependencies between variables?
   A. Undirected graph
   B. Directed acyclic graph (DAG)
   C. Tree structure
   D. Bipartite graph
2. Which property is explicitly encoded by the structure of a Bayesian network?
   A. Conditional independence
   B. Causal strength
   C. Temporal ordering
   D. Clustering coefficient

Quiz

1. What does a Bayesian network use to represent dependencies between variables?
   A. Undirected graph
   B. Directed acyclic graph (DAG)
   C. Tree structure
   D. Bipartite graph

2. Which property is explicitly encoded by the structure of a Bayesian network?
   A. Conditional independence
   B. Causal strength
   C. Temporal ordering
   D. Clustering coefficient

Quiz

1. Suppose variable X is independent of Y given Z in a Bayesian network. Which expression represents this?
   A. $P(X|Y,Z)=P(X|Z)$
   B. $P(X,Y|Z)=P(X|Z)$
   C. $P(X,Y,Z)=P(X|Y,Z)$
   D. $P(X|Y)=P(X)$
2. What is the role of *evidence* in a Bayesian network?
   A. Defines the prior probabilities.
   B. Enables updating probability distributions.
   C. It rearranges the network structure.

Quiz

1. Suppose variable X is independent of Y given Z in a Bayesian network. Which expression represents this?
   A. $P(X|Y,Z)=P(X|Z)$
   B. $P(X,Y|Z)=P(X|Z)$
   C. $P(X,Y,Z)=P(X|Y,Z)$
   D. $P(X|Y)=P(X)$

2. What is the role of *evidence* in a Bayesian network?
   A. Defines the prior probabilities.
   B. Enables updating probability distributions.
   C. It rearranges the network structure.

Quiz

1. How is the joint probability distribution of all variables in a Bayesian network computed?
   A. By summing the probabilities of each node
   B. By multiplying probabilities along the longest path
   C. By multiplying the conditional probabilities of each node given its parents
   D. By dividing the total probability equally among all nodes

Inference by Enumeration in Bayes Net

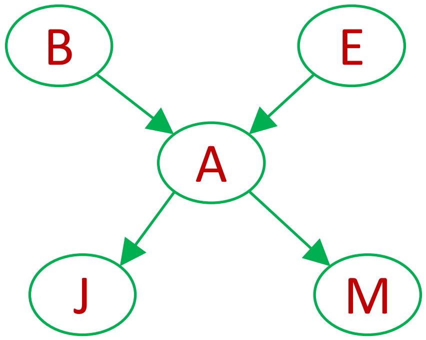
- Reminder of inference by enumeration
  - Any probability of interest can be computed by summing entries from the joint distribution:
    - $P(\mathbf{Q} \mid \mathbf{e}) = \alpha \sum_{\mathbf{h}} P(\mathbf{Q}, \mathbf{h}, \mathbf{e})$
  - Entries from the joint distribution can be obtained from a BN by multiplying the corresponding conditional probabilities

Inference by Enumeration in Bayes Net

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

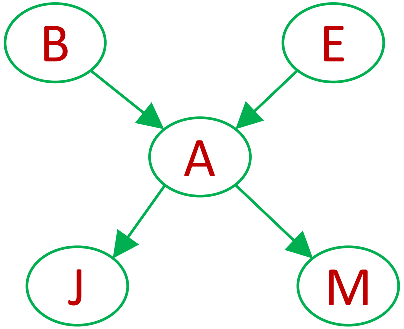
  $= \alpha \sum_e \sum_a P(B)\, P(e)\, P(a \mid B, e)\, P(j \mid a)\, P(m \mid a)$

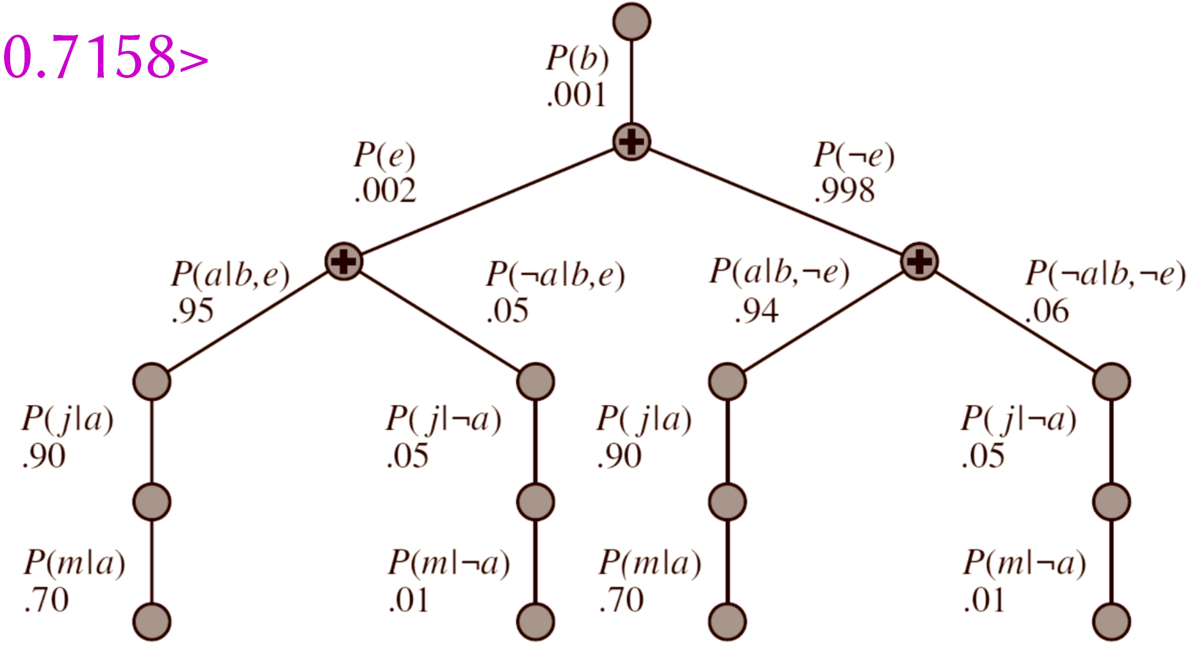
- So inference in Bayes nets means computing sums of products of numbers: sounds easy!!
- Problem: sums of **exponentially many** products!



Inference by Enumeration in Bayes Net

- $P(B \mid j, m) = \alpha\, P(B) \sum_{e,a} P(e)\, P(a \mid B, e)\, P(j \mid a)\, P(m \mid a)$
- $P(b \mid j, m) = \alpha * 0.00059224$, $P(\neg b \mid j, m) = \alpha * 0.0014919$
- $P(B \mid j, m) = < 0.2842, 0.7158>$





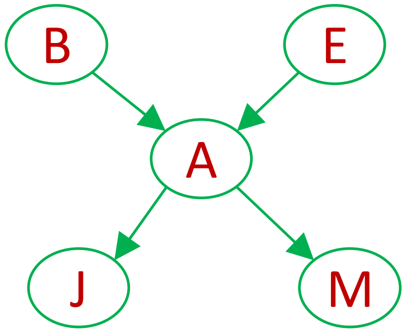
Inference by Enumeration in Bayes Net

- Note: $P(B \mid j) = <0.0163, 0.9873>$

  BUT

  $P(B \mid j, \neg m) = <0.0051, 0.9949>$!!

- Homework.
  - Show that
    - $P(B \mid \neg j, m) = <0.0069, 0.9931>$
    - $P(A) = <0.0025, 0.9975>$
    - $P(A \mid j, m) = <0.7607, 0.2393>$



Can we do better?

$$\sum_{e,a} P(B)\ P(e)\ P(a|B,e)\ P(j|a)\ P(m|a)$$

$$= \quad P(B)\ P(e) \quad P(a|B,e) \quad P(j|a) \quad P(m|a)$$

$$+ \ P(B)\ P(\neg e)\ P(a|B,\neg e) \quad P(j|a) \quad P(m|a)$$

$$+ \ P(B)\ P(e) \quad P(\neg a|B,e) \quad P(j|\neg a)\ P(m|\neg a)$$

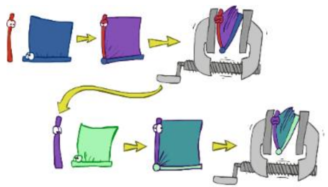
$$+ \ P(B)\ P(\neg e)\ P(\neg a|B,\neg e)\ P(j|\neg a)\ P(m|\neg a)$$

Lots of repeated subexpressions!

Can we do better?

- Consider $uwy + uwz + uxy + uxz + vwy + vwz + vxy + vxz$
  - 16 multiplies, 7 adds
  - Lots of repeated subexpressions!
- Rewrite as $(u+v)(w+x)(y+z)$
  - 2 multiplies, 3 adds

Variable elimination



- Store calculations to eliminate repeated evaluations
- Move summations inwards as far as possible

$$P(B \mid j, m) = \alpha \sum_{e,a} P(B)\, P(e)\, P(a \mid B, e)\, P(j \mid a)\, P(m \mid a)$$

$$= \alpha\, P(B) \sum_{\mathbf{e}} P(e) \sum_{\mathbf{a}} P(a \mid B, e)\, P(j \mid a)\, P(m \mid a)$$

- Do the calculation from right to left (inside out)
  - Sum over a first, then sum over e

Operation 1: Pointwise product

- In **pointwise product** of factors (similar to a database join, not matrix multiply!)
  - New factor has union of variables of the two original factors
  - Each entry is the product of the corresponding entries from the original factors

P(J|A) x P(A) = P(A,J)

- Example:

P(A)

true	0.1
false	0.9

x

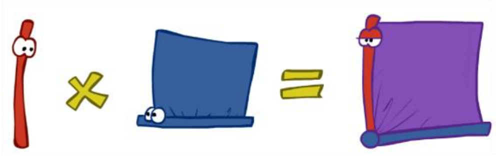
P(J|A)

A \ J	true	false
true	0.9	0.1
false	0.05	0.95

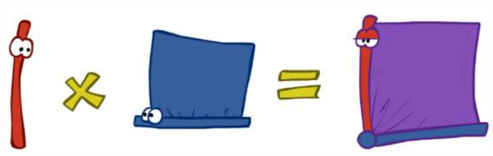
=

P(A,J)

A \ J	true	false
true	0.09	0.01
false	0.045	0.855



Operation 1: Pointwise product



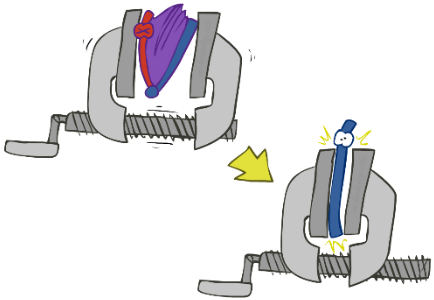
- Another example:

X	Y	$\mathbf{f}(X,Y)$	Y	Z	$\mathbf{g}(Y,Z)$	X	Y	Z	$\mathbf{h}(X,Y,Z)$
t	t	.3	t	t	.2	t	t	t	$.3 \times .2 = .06$
t	f	.7	t	f	.8	t	t	f	$.3 \times .8 = .24$
f	t	.9	f	t	.6	t	f	t	$.7 \times .6 = .42$
f	f	.1	f	f	.4	t	f	f	$.7 \times .4 = .28$
						f	t	t	$.9 \times .2 = .18$
						f	t	f	$.9 \times .8 = .72$
						f	f	t	$.1 \times .6 = .06$
						f	f	f	$.1 \times .4 = .04$

Figure 13.12 Illustrating pointwise multiplication: $\mathbf{f}(X,Y) \times \mathbf{g}(Y,Z) = \mathbf{h}(X,Y,Z)$.

Operation 2: Summing out a variable

- Summing out or Marginalizing a variable from a factor shrinks a factor to a smaller one
- Example: $\sum_j P(A,J) = P(A,j) + P(A,\neg j) = P(A)$



P(A,J)

A , J	true	false
true	0.09	0.01
false	0.045	0.855

Marginalize J

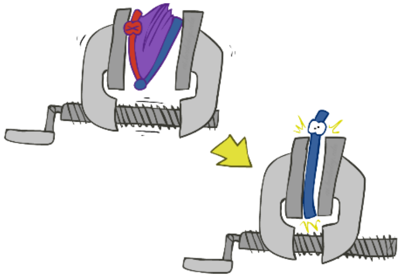
P(A)

true	0.1
false	0.9

Summing out from a product of factors

- Project the factors each way first, then sum the products

Example:



$\sum_a P(a|B,e) \times P(j|a) \times P(m|a)$

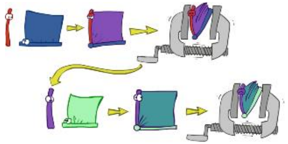
$= \quad P(a|B,e) \quad \times \quad P(j|a) \quad \times \quad P(m|a)$

$+ \quad P(\neg a|B,e) \times P(j|\neg a) \times P(m|\neg a)$

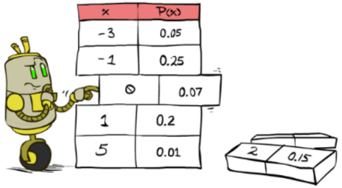
$$\mathbf{h}_2(Y,Z) = \sum_x \mathbf{h}(X,Y,Z) = \mathbf{h}(x,Y,Z) + \mathbf{h}(\neg x,Y,Z)$$

$$= \begin{pmatrix} .06 & .24 \\ .42 & .28 \end{pmatrix} + \begin{pmatrix} .18 & .72 \\ .06 & .04 \end{pmatrix} = \begin{pmatrix} .24 & .96 \\ .48 & .32 \end{pmatrix}$$

Variable Elimination



- Query: $P(Q \mid E_1=e_1, \ldots, E_k=e_k)$
- Start with initial factors:
  - Local CPTs (but instantiated by evidence)



- While there are still hidden variables (not Q or evidence):
  - Pick a hidden variable H_j
  - Eliminate (marginalize) H_j from the product of all factors mentioning H_j
- Join all remaining factors and normalize

 x α

Example

Query P(B | j,m)

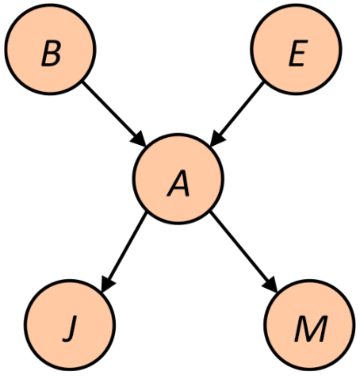
$$\mathbf{P}(B \mid j,m) = \alpha \mathbf{f}_1(B) \times \sum_e \mathbf{f}_2(E) \times \sum_a \mathbf{f}_3(A,B,E) \times \mathbf{f}_4(A) \times \mathbf{f}_5(A)$$

P(B) P(E) P(A|B,E) P(j|A) P(m|A)

Choose A

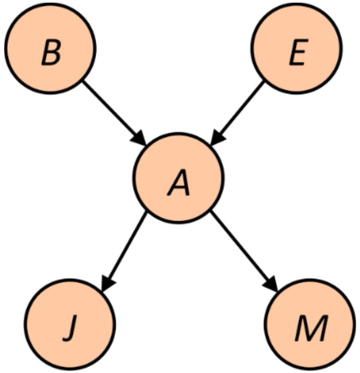
P(A|B,E)
P(j|A)
P(m|A)

$\xrightarrow{\times} \xrightarrow{\sum}$ P(j,m|B,E)



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

Example



Query P(B | j,m)

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

Choose E

P(E)
P(j,m|B,E) $\xrightarrow{\times}$ $\xrightarrow{\Sigma}$ P(j,m|B)

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

Finish with B

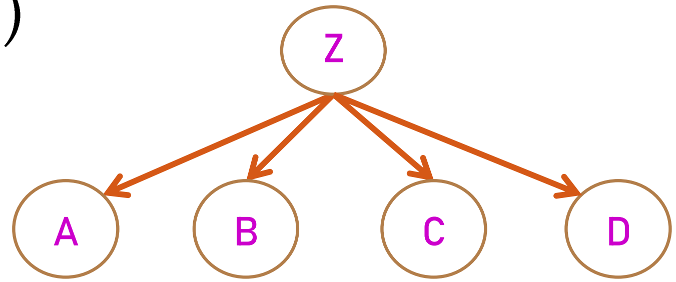
P(B)
P(j,m|B) $\xrightarrow{\times}$ P(j,m,B) $\xrightarrow{\text{Normalize}}$ P(B | j,m)

Order matters

- Order the terms Z, A, B C, D

$$P(D) = \alpha \sum_{z,a,b,c} P(z)\, P(a|z)\, P(b|z)\, P(c|z)\, P(D|z)$$
$$= \alpha \sum_{z} P(z) \sum_{a} P(a|z) \sum_{b} P(b|z) \sum_{c} P(c|z)\, P(D|z)$$

- Largest factor has 2 variables (D, Z)

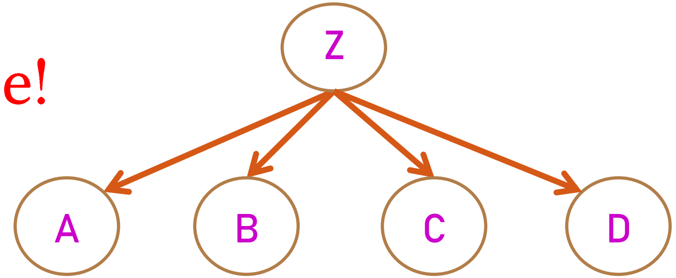


Order matters

- Order the terms A, B C, D, Z

$$P(D) = \alpha \textstyle\sum_{a,b,c,z} P(a|z)\, P(b|z)\, P(c|z)\, P(D|z)\, P(z)$$
$$= \alpha \textstyle\sum_{a} \sum_{b} \sum_{c} \sum_{z} P(a|z)\, P(b|z)\, P(c|z)\, P(D|z)\, P(z)$$

  - Largest factor has 4 variables (A,B,C,D)
  - In general, with n leaves, factor of size 2^n
- Finding optimal ordering is intractable!



Quiz

- Order the terms for $P(J \mid b)$

Quiz

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

- Every variable that is not an ancestor of query or evidence does not matter - M in this example

Quiz

1. What does a factor f(X,Y) represent in the context of Bayes net?
   A. A function giving scores to each value pair of X and Y
   B. A way to normalize probabilities
   C. A method for determining variable elimination order
2. What is the reason for carefully choosing the variable elimination order?
   A. Reduces the number of required normalizations
   B. Minimizes the size of intermediate factors
   C. Eliminates the need for conditioning on evidence

Quiz

1. What does a factor f(X,Y) represent in the context of Bayes net?
   A. A function giving scores to each value pair of X and Y
   B. A way to normalize probabilities
   C. A method for determining variable elimination order
2. What is the reason for carefully choosing the variable elimination order?
   A. Reduces the number of required normalizations
   B. Minimizes the size of intermediate factors
   C. Eliminates the need for conditioning on evidence

Quiz

- When do you normalize the final factor in variable elimination?
    A. After each marginalization step
    B. Before any evidence is conditioned
    C. Once all hidden variables are eliminated
    D. Never. Normalization is not needed

Quiz

- When do you normalize the final factor in variable elimination?
    A. After each marginalization step
    B. Before any evidence is conditioned
    C. Once all hidden variables are eliminated
    D. Never. Normalization is not needed

VE: Computational and Space Complexity

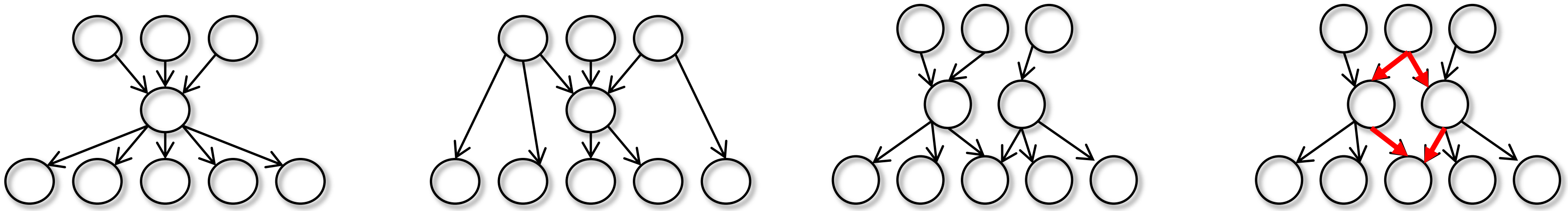
- The computational and space complexity of variable elimination is determined by the largest factor (and it's space that kills you)

- The elimination ordering can greatly affect the size of the largest factor.
  - E.g., previous slide's example 2^n vs. 2

- Does there always exist an ordering that only results in small factors?
  - No!

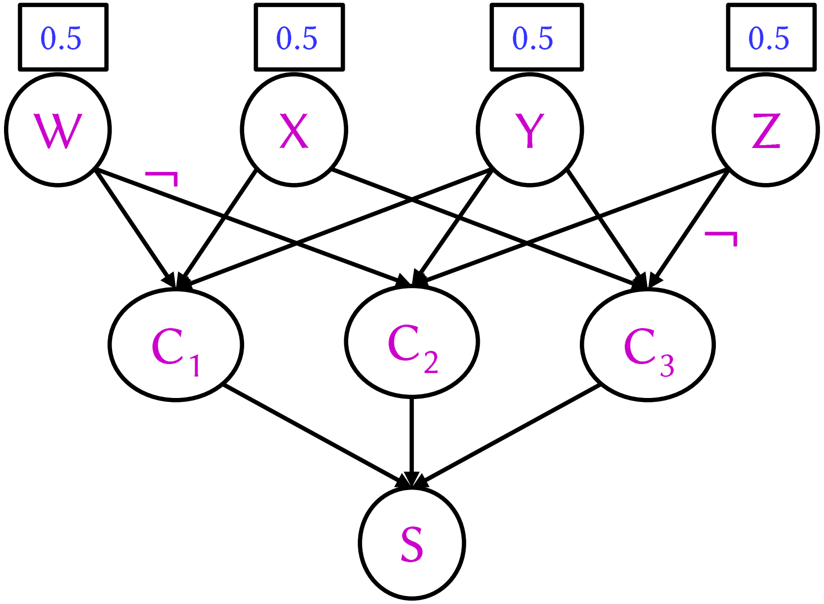
Polytrees

- A polytree is a directed graph with no undirected cycles
- For polytrees the complexity of variable elimination is linear in the network size (number of CPT entries) if you eliminate from the leave towards the roots



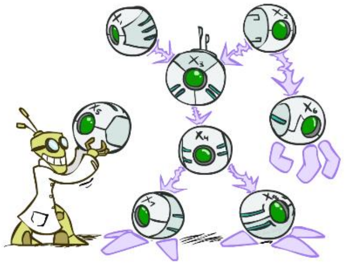
Worst Case Complexity - Reduction from SAT

- Variables: W, X, Y, Z
- CNF clauses:
  1. $C_1 = W \vee X \vee Y$
  2. $C_2 = Y \vee Z \vee \neg W$
  3. $C_3 = X \vee Y \vee \neg Z$
- Sentence $S = C_1 \wedge C_2 \wedge C_3$
- $P(S) > 0$ iff S is satisfiable
  $\Longrightarrow$ **NP-hard**
- $P(S) = K \times 0.5^n$ where K is the number of satisfying assignments for clauses
  $\Longrightarrow$ **#P-hard**



Summary

- Independence and conditional independence are important forms of probabilistic knowledge
- Bayes nets encode joint distributions efficiently by taking advantage of conditional independence
  - Global joint probability = product of local conditionals
- Exact inference = sums of products of conditional probabilities from the network



- Reading: Chapter 13
- Assignments: PS 7
- Next:
  - Chapter 14 - Markov Models