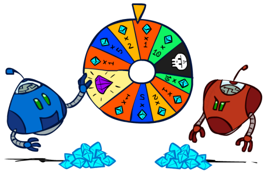
Artificial Intelligence

Module 6: Reasoning under Uncertainty

PART 6.1 : Quantifying Uncertainty
PART 6.2 : Probablistic Reasoning



Dr. Chandra Prakash
Assistant Professor
Department of Computer Science and Engineering

(Slides adapted from StuartJ. Russell, B Ravindran, Mausam, Dan Klein and Pieter Abbeel, Partha P Chakrabarti, Saikishor Jangiti)

Module 6: Reasoning under Uncertainty

- PART 6.1 : Quantifying Uncertainty
  - Basic of Probability
- PART 6.2 : Probablistic Reasoning
  - Bayes Rule
  - Bayesian Network
- PART 6.3 : Fuzzy Logic
- PART 6.4 : Probabilistic Reasoning over time
  - Hidden Markov Model
  - Kalman filter
  - Markov Chain Monte Carlo
- PART 6.5 : Decisions Theory
  - Utility Function
  - Decision Network
  - Markov Decision Proces

Probability Recap

- Conditional probability $P(x|y) = \frac{P(x,y)}{P(y)}$
- Product rule $P(x,y) = P(x|y)P(y)$
- Chain rule
$$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1}) \end{aligned}$$
- X, Y independent if and only if: $\forall x,y : P(x,y) = P(x)P(y)$
- X and Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y | Z$
$$\forall x,y,z : P(x,y|z) = P(x|z)P(y|z)$$

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- 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
  - Foundation of many systems we'll see later (e.g. ASR etc.)
- In the running for most important AI equation!

That's my rule!

Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$\boxed{P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}}$$

- This simple equation underlies all modern AI systems for probabilistic inference
- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \textit{Example givens}$$

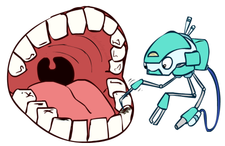
$P(+m|+s)$

  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

Inference by Enumeration

- What is the probability of a cavity given a toothache?
- What is the probability of a cavity given the probe catches?
- Example : Joint Distribution
  - Three Random Variables
    - cavity, toothache and catch

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576



- P(toothache) ?

=0.108+0.012+0.016+0.064 = 0.20 or 20%

- P(toothache ∨ cavity) ?

= 0.20 + 0.072 + 0.008 = 0.28

- conditional probabilities : P(¬cavity | toothache) ?

= 0.4

Quiz: Bayes' Rule

- Given:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

- What is P(W | dry) ?

W	P(W | dry)
sun	0.9 * 0.8 / (0.9*0.8 +0.3*0.2)
rain	1- (0.9 * 0.8 / (0.9*0.8 +0.3*0.2))

Complexity of Enumeration

- Worst case time: $O(d^n)$
  - d = maximum arity of the random variables
  - n = number of random variables
- Space complexity also $O(d^n)$
  - Size of joint distribution
- Problem
  - Hard/impossible to estimate all $O(d^n)$ entries of joint distribution for large problems
  - Prohibitive

Ghostbusters, Revisited

- Let’s say we have two distributions:
  - Prior distribution over ghost location: P(G)
    - Let’s say this is uniform
  - Sensor reading model: P(R | G)
    - Given: we know what our sensors do
    - R = reading color measured at (1,1)
    - E.g. P(R = yellow | G=(1,1)) = 0.1

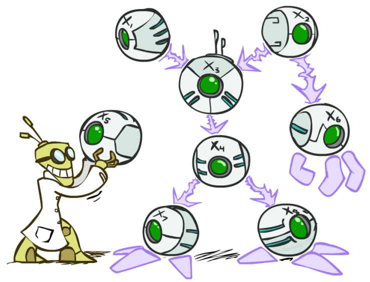
0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

- We can calculate the posterior distribution P(G|r) over ghost locations given a reading using Bayes’ rule:

$$P(g|r) \propto P(r|g)P(g)$$

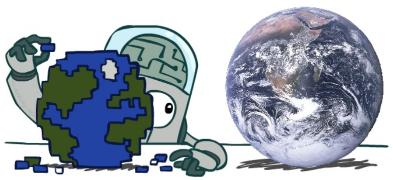
0.17	0.10	0.10
0.09	0.17	0.10
<0.01	0.09	0.17

Bayes’ Nets



Probabilistic Models

- Models describe how (a portion of) the world works
- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - “All models are wrong; but some are useful.”
    – George E. P. Box



- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information

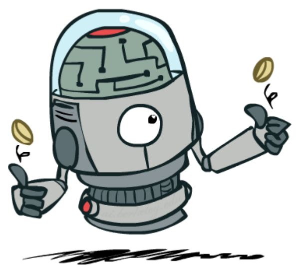
Independence derviation

- A and B are independent iff:

  P(A| B) = P(A)
  P(B | A) = P(B)

  These two constraints are logically equivalent

- Therefore, if A and B are independent:



$$P(A\,|\,B) = \frac{P(A \wedge B)}{P(B)} = P(A)$$

$$P(A \wedge B) = P(A)P(B)$$

Independence

- Two variables are independent if:

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

  - This says that their joint distribution factors into a product two simpler distributions
  - Another form: $$\forall x, y : P(x|y) = P(x)$$
  - We write: $X \perp\!\!\!\perp Y$
- Independence is a simplifying modeling assumption
  - Empirical joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?

Example: Independence?

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Not Same

Not Independent

$P_2(T, W)$

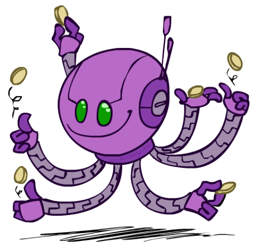
T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

Example: Independence

- N fair, independent coin flips:



$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

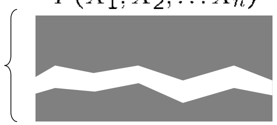
$\ldots$

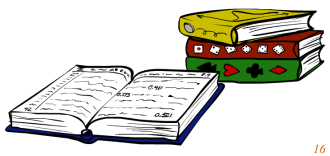
$P(X_n)$

H	0.5
T	0.5

$P(X_1, X_2, \ldots X_n)$

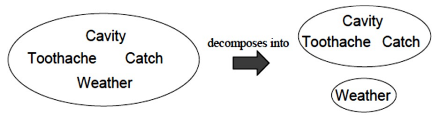
2^n





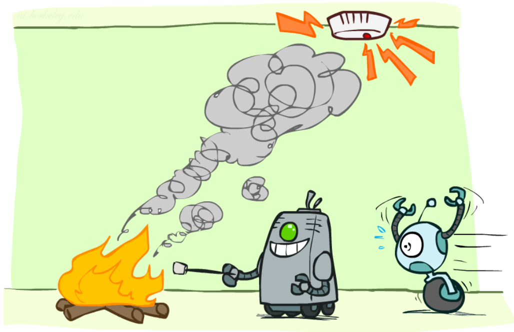
Independence

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



  – P(Tootache, cath, Cavity, Weather)= P(Tootache, cath, Cavity) P (Weather)
- Size of the joint probability distribution table?
  – 32 entries reduced to 12; for n indepedent biased coins, $2^n \rightarrow n$
  – Factorization
- Absolute independences are rare.
  – Complete independence is powerful but rare What to do if it doesn't hold?

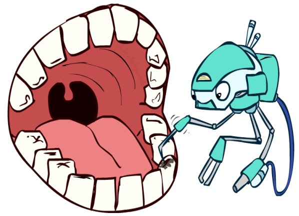
Conditional Independence



Weaker Assumption

Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)
- The same independence holds if I don't have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)
- Catch is conditionally independent of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)



- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily

Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z

$$X \perp\!\!\!\perp Y | Z$$

if and only if:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

$$P(x|z,y) = \frac{P(x,z,y)}{P(z,y)} = \frac{P(x,y|z)P(z)}{P(y|z)P(z)} = \frac{P(x|z)P(y|z)P(z)}{P(y|z)P(z)}$$

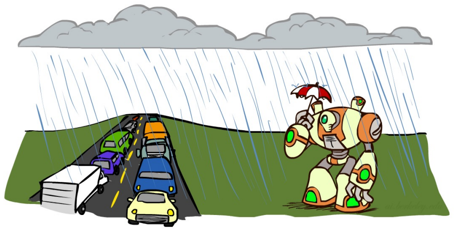
Power of Cond. Independence

- Often, using conditional independence reduces the storage complexity of the joint distribution from exponential to linear!!
- Conditional independence is the most basic & robust form of knowledge about uncertain environments.

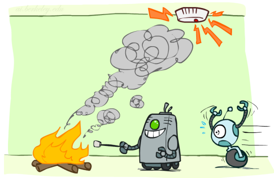
Conditional Independence

- What about this domain:
  - Traffic
  - Umbrella
  - Raining

$T \perp\!\!\!\perp U / R$

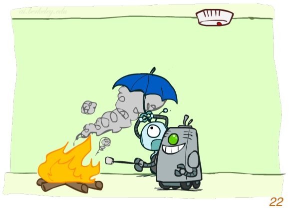


- Fire
- Smoke
- Alarm



$A \perp\!\!\!\perp S / F$

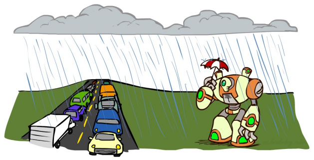
$A \perp\!\!\!\perp F / S$



Conditional Independence and the Chain Rule

- Chain rule: $P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$
- Trivial decomposition:

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



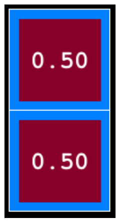
- With assumption of conditional independence:

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

- Bayes'nets / graphical models help us express conditional independence assumptions

Ghostbusters Chain Rule

- Each sensor depends only on where the ghost is
- That means, the two sensors are conditionally independent, given the ghost position
- T: Top square is red
  B: Bottom square is red
  G: Ghost is in the top
- Givens:
  P(+g) = 0.5
  P(-g) = 0.5
  P(+t | +g) = 0.8
  P(+t | -g) = 0.4
  P(+b | +g) = 0.4
  P(+b | -g) = 0.8



P(T,B,G) = P(G) P(T|G) P(B|G)

T	B	G	P(T,B,G)
+t	+b	+g	0.16
+t	+b	-g	0.16
+t	-b	+g	0.24
+t	-b	-g	0.04
-t	+b	+g	0.04
-t	+b	-g	0.24
-t	-b	+g	0.06
-t	-b	-g	0.06

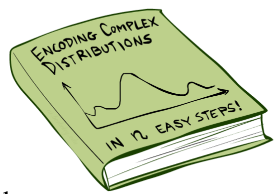
Bayes'Nets (Bayesian Network) : Big Picture

In Probabilities, we have seen that random variables can be related with joint distribution, conditionality, independence, etc.

• However, such relations are defined in a formal semantics using a network topology and probabilities

• Bayesian Networks is a field that represents such relations systematically

• We explore the use of probability theory to objects and relations (like in first-order logic)



Bayesian Network : Big Picture

- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

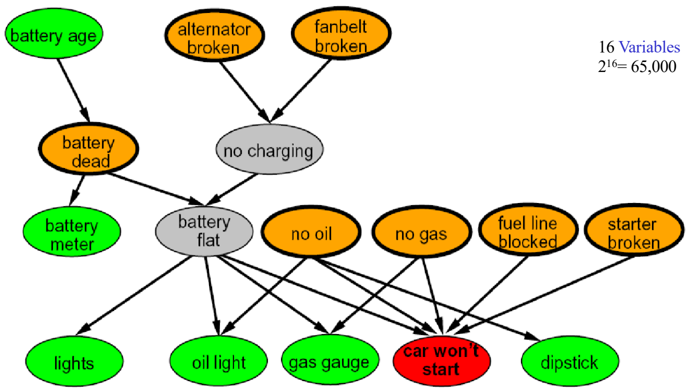
- Bayesian Network: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - Also known as Bayes Nets \ Belief Network
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions
  - For about 10 min, we'll be vague about how these interactions are specified

Example Bayesian Network : Insurance

27 Variables
2^{27}= 134 Millions

Example Bayes' Net: Car diagnosis

16 Variables
2^{16}= 65,000



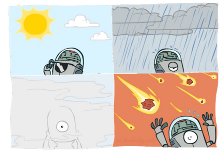
Bayesian Network

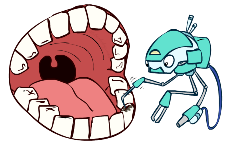
Bayes Nets

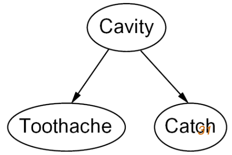
- In general, joint distribution P over set of variables (X1 x ... x Xn) requires exponential space for representation & inference
- BNs provide a graphical representation of conditional independence relations in P
  - –usually quite compact
  - –requires assessment of fewer parameters, those being quite natural (e.g., causal)
  - –efficient (usually) inference: query answering and belief update

Graphical Model Notation

- Nodes: variables (with domains)
  - Set of random variables
  - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
  - Indicate "direct influence" between variables
  - Formally: encode conditional independence (more later)
- For now: imagine that arrows mean direct causation (in general, they don't!)
- Each node has a conditional probability table that quantifies the effects that the parent have on the node.
  - The graph has no directed cycles. It is a directed acyclic graph (DAG).

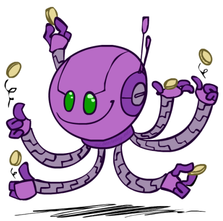






Example: Coin Flips

- N independent coin flips



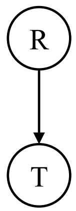
- No interactions between variables: absolute independence

Example: Traffic

- Variables:
  - R: It rains
  - T: There is traffic

- Model 1: independence

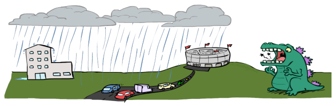
- Model 2: rain causes traffic

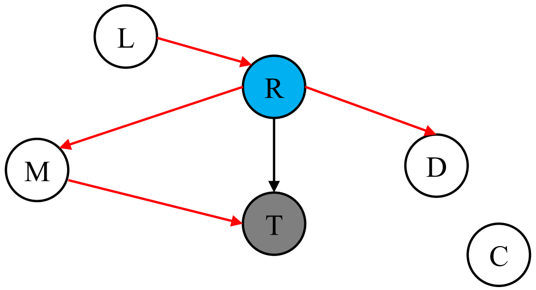


- Why is an agent using model 2 better?

Example: Traffic II

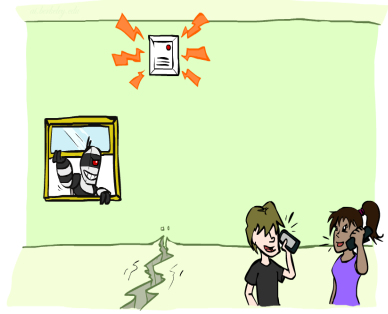
- **Let's build a causal graphical model**
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - M: Cricket Match
  - C: Cavity





Classical Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake
- Burglar alarm at home
  - Fairly reliable at detecting a burglary
  - Responds at times to minor earthquakes
- Two neighbors, on hearing alarm, calls police
  - John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too.
  - Mary likes loud music and sometimes misses the alarm altogether

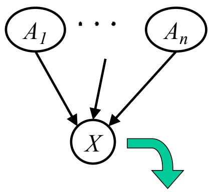


Bayesian Network Semantics

- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
  - A collection of distributions over X, one for each combination of parents' values

$$P(X|A_1 \ldots A_n)$$

  - CPT: Conditional Probability Table
  - Description of a noisy "causal" process



$$P(X|A_1 \ldots A_n)$$

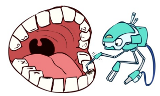
A Bayes net = Topology (graph) + Local Conditional Probabilities

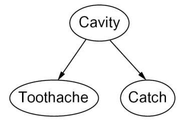
Probabilities in Bayesian Network (BNs)

- Bayesian Network implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

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

  - Example:





$$P(+cavity, +catch, -toothache)$$

= P(+cavity) P(+catch|+cavity) P(-toothache|+cavity)

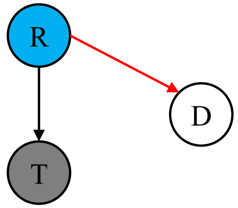
Why Is the Product Legitimate

- Chain rule (valid for all distributions): $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | x_1 \ldots x_{i-1})$
- Assume conditional independences: $P(x_i | x_1, \ldots x_{i-1}) = P(x_i | \mathit{parents}(X_i))$

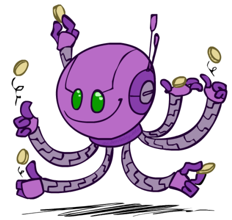
Consequence:

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

- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies
    - T⊥ D/R



Example: Coin Flips



$P(X_1)$	
h	0.5
t	0.5

$P(X_2)$	
h	0.5
t	0.5

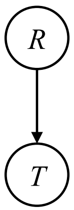
. . .

$P(X_n)$	
h	0.5
t	0.5

$P(h, h, t, h) =$ P(h)P(h)P(t)P(h)

Only distributions whose variables are absolutely independent can be represented by a Bayes' net with no arcs.

Example: Traffic



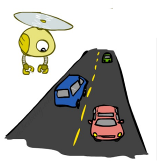
$P(R)$

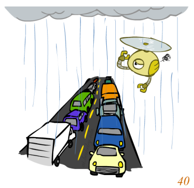
+r	1/4
-r	3/4

$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

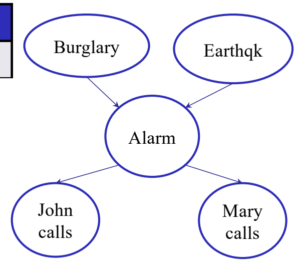
$P(+r, -t) =$ P(+r)P(-t|+r) = ¼*1/4



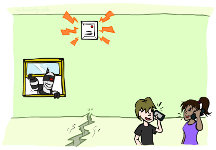


Example: Alarm Network

B	P(B)
+b	0.001



E	P(E)
+e	0.002



B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	-e	+a	0.94
-b	+e	+a	0.29
-b	-e	+a	0.001

A	J	P(J|A)
+a	+j	0.9
-a	+j	0.05

A	M	P(M|A)
+a	+m	0.7
-a	+m	0.01

P(M|A)P(J|A)P
(A|B,E)

Example: Traffic

- Causal direction

$$R \rightarrow T$$

$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Example: Reverse Traffic

- Reverse causality?

$$T \rightarrow R$$

$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

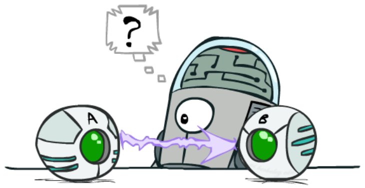
$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Causality?

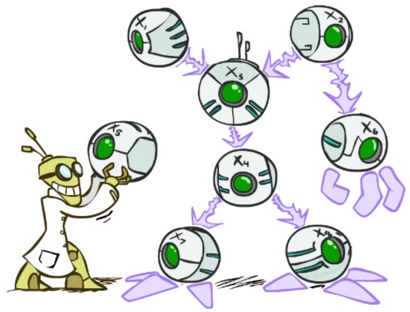
- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables Traffic and Drips
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

$$P(x_i|x_1, \ldots x_{i-1}) = P(x_i|\mathit{parents}(X_i))$$



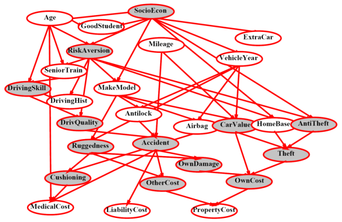
Bayes' Nets

- So far: how a Bayes' net encodes a joint distribution
- Next: how to answer queries about that distribution
  - Today:
    - First assembled BNs using an intuitive notion of conditional independence as causality
    - Then saw that key property is conditional independence
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)



Bayes Nets

- A Bayes net is an efficient encoding of a probabilistic model of a domain



- Questions we can ask:
  - Inference: given a fixed BN, what is P(X | e)?
  - Representation: given a BN graph, what kinds of distributions can it encode?
  - Modeling: what BN is most appropriate for a given domain?

Bayes Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over X, one for each combination of parents' values

$$P(X|a_1 \ldots a_n)$$

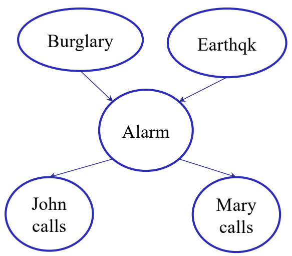
- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

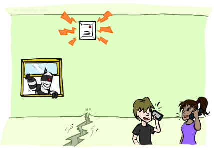
$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

Example: Alarm Network

P(B)
0.001

P(E)
0.002





P(M|A)
P(J|A)
P(A|B,E)
P(~A|~B∧E)

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

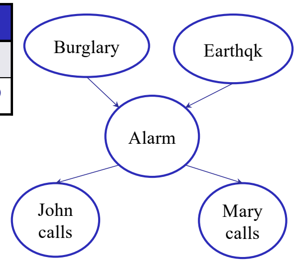
A	P(J|A)
T	0.9
F	0.05

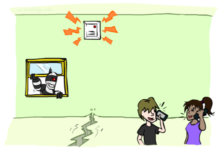
A	P(M|A)
T	0.7
F	0.01

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998





P(M|A)
P(J|A)
P(A|B,E)
P(A|~B∧E)
P(A) ??

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

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

P(A|~B∧E)
= P(~A|~B∧E). P(~B). P(E)

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

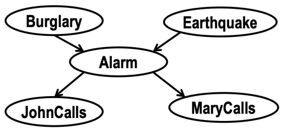
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

Computation of the probabilities of several different event combinations of the Burglary-Alarm belief network example

The joint probability distribution

- Computation of the probabilities of several different event combinations of the BurglaryAlarm belief network example:

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



- Probalility that alaram go off.

P(A) = P(A¬B¬E) + P(A¬BE) + P(AB¬E) + P(ABE)

let ¬ = '

so P(A)=P(AB'E') + P(AB'E) + P(ABE') + P(ABE)

= P(A | B'E').P(B'E') + P(A | B'E).P(B'E) + P(A | BE').P(BE') + P(A | BE).P(BE)

= 0.001 x 0.999 x 0.998 + 0.29 x 0.999 x 0.002 + 0.95 x 0.001 x 0.998 + 0.95 x 0.001 x 0.002

= 0.001 + 0.0006 + 0.0009 = 0.0025

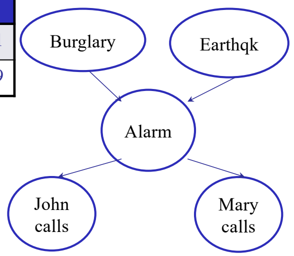
Find

- P(A'B),
- P(AE),
- P(J) ,
- P(AB) ??

Example

Probability of the event that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both Mary and John call:

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

P(J ∧ M ∧ A ∧ ¬B ∧ ¬E) = P(JMAB'E')

Apply joint probability distribution

= P(J | MAB'E') P(MAB'E')

= P(J | A) P(MAB'E')

= P(J | A) P(M | A) P(AB'E')

= P(J | A) P(M | A) P(A | B'E') P(B'E')

= P(J | A) P(M | A) P(A | B'E') P(B') P(E')

= P(J | A) P(M | A) P(A | ¬B ∧ ¬E) P(¬B) P(¬E)

= 0.9 X 0.7 X 0.001 X 0.999 X 0.998

= 0.00062

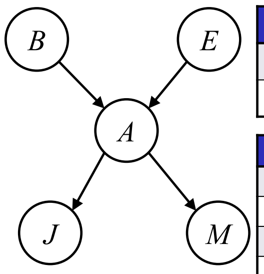
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

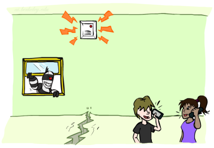
B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

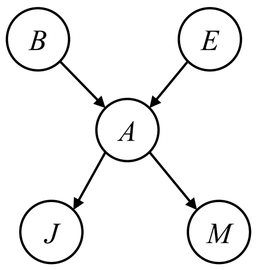
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

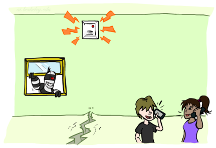
$$P(+b, -e, +a, -j, +m) =$$
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

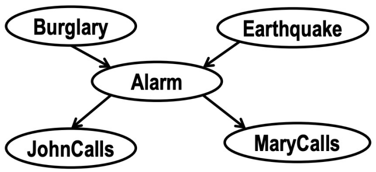
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b, -e, +a, -j, +m) =$$
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

More Exercise

- P(MB) = P(MBA) + P(MBA')
  - = P(M | AB).P(AB) + P(M | A'B).P(A'B)
  - = P(M | A).P(AB) + P(M | A').P(A'B)
  - = 0.7 x 0.00095 + 0.01 x 0.00005
  - = 0.00067
- P(M | B) = P(MB) / P(B)
  - = 0.00067 / 0.001 = 0.67
- P(B | J) = P(JB) / P(J)
  - = 0.00086 / 0.052125 = 0.016
- P(B | A) = P(AB) / P(A)
  - = 0.00095 / 0.0025 = 0.38
- P(B | AE) = P(ABE) / P(AE) = [P(A | BE).P(BE)] / P(AE)
  - = [0.95 x 0.001 x 0.002] / 0.00058
  - = 0.003



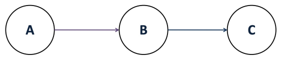
Other Applications

- Medical Diagnosis
- Computational Biology and Bioinformatics
- Natural Language Processing
- Document classification
- Image processing
- Decision support systems
- Ecology & natural resource management
- Robotics
- Forensic science…

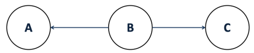
d-Separation

- Answer independence queries regarding variables in a BN.
  - Study independence properties for triples
    - Causal chain, common cause and common effect
- Analyze complex cases in terms of member triples
- D-Separation: **a condition/algorithm for answering these queries.**
- An undirected path between two nodes is "cut off"
  - if information cannot flow across one of the nodes in the path
- Two nodes are d-separated
  - if every undirected path between them is cut off
- Two sets of nodes are d-separated
  - if every pair of nodes, one from each set, is d-separated
- Two sets of nodes are independent
  - if they are d-separated

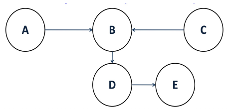
d-Separation



- Linear connection:
  - Information can flow between A and C if and only if we do have evidence at B



- Diverging connection:
  - Information can flow between A and C if and only if we do not have evidence at B
- Converging connection:
  - Information can flow between A and C if and only if we do have evidence at B or and descendent of B (such as D or E)



d-Separation : Causal Chains

- This configuration is a “causal chain”

X: Low pressure **Y: Rain** **Z: Traffic**

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Is it guaranteed that X independent of Z ?
  - No!
  - One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.
  - Example:
    - Low pressure causes rain causes traffic, high pressure causes no rain causes no traffic
    - In numbers:

      P(+y | +x) = 1, P(-y | - x) = 1,
      P(+z | +y) = 1, P(-z | -y) = 1

d-Separation : Causal Chains

- This configuration is a “causal chain”

X: Low pressure Y: Rain Z: Traffic

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Guaranteed X independent of Z given Y?

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

$$= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$

$$= P(z|y)$$

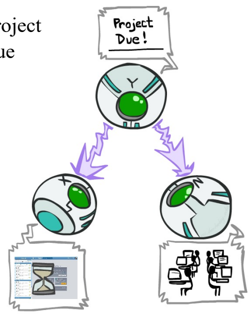
Yes!

- Evidence along the chain “blocks” the influence

d-Separation : Causal Chains

- This configuration is a “common cause”

Y: Project due



X: Forums busy

Z: Lab full

$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

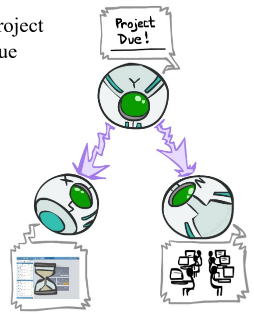
- Guaranteed X independent of Z ?
- No!
  - One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.
  - Example:
    - Project due causes both forums busy and lab full
    - In numbers:

      P(+x | +y) = 1, P(-x | -y) = 1,
      P(+z | +y) = 1, P(-z | -y) = 1

d-Separation : Causal Chains

- This configuration is a “common cause”

Y: Project due



X: Forums busy

Z: Lab full

$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Guaranteed X and Z independent given Y?

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

$$= \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)}$$

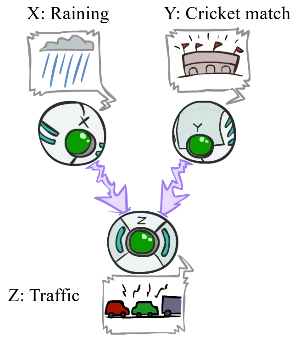
$$= P(z|y)$$

Yes!

- Observing the cause blocks influence between effects.

Common Effect

- Last configuration: two causes of one effect (v-structures)

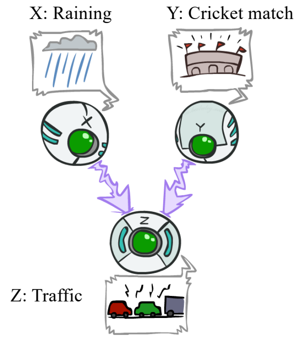


- Are X and Y independent?
  - Yes: the Y: Cricket match and the rain cause traffic, but they are not correlated
- Proof:

$$P(x,y) = \sum_z P(x,y,z)$$
$$= \sum_z P(x)P(y)P(z|x,y)$$
$$= P(x)P(y)\sum_z P(z|x,y)$$
$$= P(x)P(y)$$

Common Effect

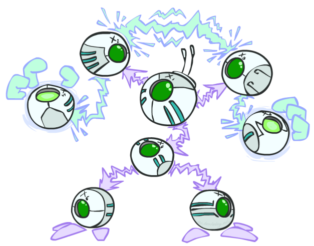
- Last configuration: two causes of one effect (v-structures)



- Are X and Y independent?
  - Yes: the Cricket match and the rain cause traffic, but they are not correlated
  - (Proved previously)
- Are X and Y independent given Z?
  - No: seeing traffic puts the rain and the Cricket match in competition as explanation.
- This is different from the other cases
  - Observing an effect activates influence between possible causes.

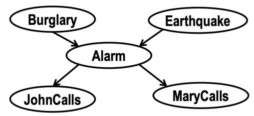
The General Case

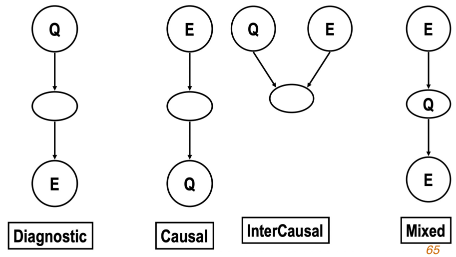
- General question: in a given BN, are two variables independent (given evidence)?
- Solution: analyze the graph
- Any complex example can be broken into repetitions of the three canonical cases



Inferences using belief networks

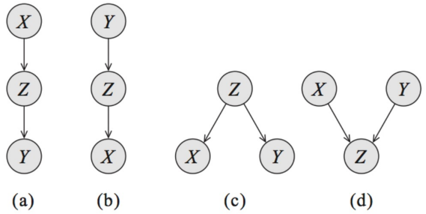
- Diagnostic inferences (from effects to causes)
  - Given that JohnCalls, infer that
    - P(Burglary | JohnCalls) = 0.016
- Causal inferences (from causes to effects)
  - Given Burglary, infer that
    - P(JohnCalls | Burglary) = 0.86
    - P(MaryCalls | Burglary) = 0.67
- Inter-causal inferences (between causes of a comm effect)
  - Given Alarm, we have P(Burglary | Alarm) = 0.376
  - If we add evidence that Earthquake is true, then P(Burglary | Alarm ∧ Earthquake) = 0.003
- Mixed inferences
  - Setting the effect JohnCalls to true and the cause Earthquake to false gives
    - P(Alarm | JohnCalls ∧ ¬ Earthquake) = 0.003





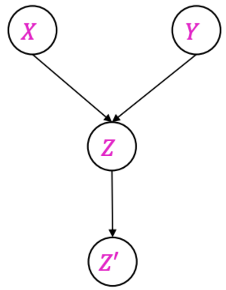
Conditional Independence Relations

- A path is blocked given a set of nodes E if there is a node Z on the path for which one of three conditions holds:
  1. Z is in E and Z has one arrow on the pat leading in and one arrow out (Case a and b)
  2. Z is in E and Z has both path arrows leading out (Case c)
  3. Neither Z nor any descendant of Z is in E, and both path arrows lead in to Z (Case d)

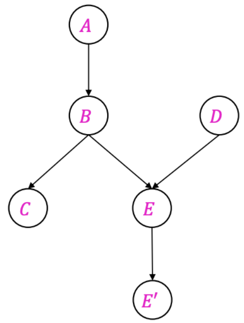


- If every undirected path from a node in X to a node in Y is d-separated (blocked) by a given set of evidence nodes E, then X and Y are conditionally independent given E.
- A set of nodes E d-separates two sets of nodes X an Y if every undirected path from a node in X to a node in Y is blocked given E.

Example

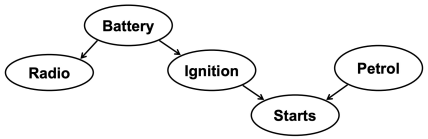


X independent of Y
X independent of Y given Z
X independent of Y given Z'



C independent of D
A independent of E' given E
A independent of D
A independent of D given E
A independent of D given B,E

Conditional Independence in Belief Networks

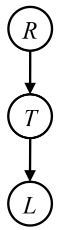


- Whether there is petrol and whether the radio plays are independent given evidence about whether the ignition takes place
- Petrol and Radio are independent if it is known whether the battery works

- Petrol and Radio are independent given no evidence at all.
- But they are dependent given evidence about whether the car starts.
- If the car does not start, then the radio playing is increased evidence that we are out of petrol.

Example: Traffic Domain

- Random Variables
  - R: Raining
  - T: Traffic
  - L: Late!



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$$P(L) = ?$$

$$= \sum_{r,t} P(r,t,L)$$

$$= \sum_{r,t} P(r)P(t|r)P(L|t)$$

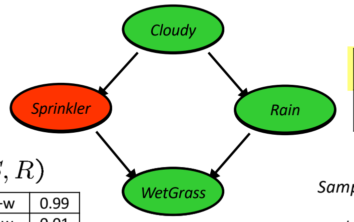
Inference in multiply connected Belief Networks

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

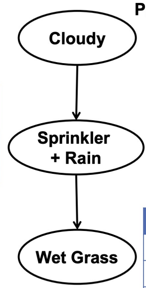
+c, -s, +r, +w

-c, +s, -r, +w

...

Clustering methods

- Transform the net into a probabilistically equivalent (but topologically different) poly-tree by merging offending nodes

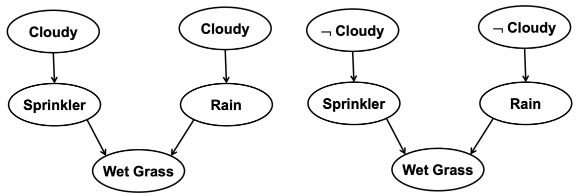


C	P(S+R = x) TT	TF	FT	FF
T	0.08	0.02	0.72	0.18
F	0.40	0.10	0.40	0.10

S+R		P(W)
T	T	0.99
T	F	0.90
F	T	0.90
F	F	0.00

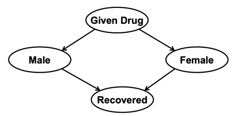
Cutset conditioning Methods

- A set of variables that can be instantiated to yield a poly-tree is called a cutset
- Instantiate the cutset variables to definite values
  - Then evaluate a poly-tree for each possible instantiation



Inference in multiply connected belief networks

- Stochastic simulation methods
  - Use the network to generate a large number of concrete models of the domain that are consistent with the network distribution.
  - They give an approximation of the exact evaluation.
  - Statistical bias can lead to misleading results – Simpson's paradox



- Should the drug be administered, or not?

Males	Recovered	Not recovered	Rec. Rate
Given drug	18	12	60%
Not given drug	7	3	70%

Females	Recovered	Not recovered	Rec. Rate
Given drug	2	8	20%
Not given drug	9	21	30%

Combined	Recovered	Not recovered	Rec. Rate
Given drug	20	20	50%
Not given drug	16	24	40%

Simpson's paradox

- P(recovery | male ∧ given_drug) = 0.6
- P(recovery | female ∧ given_drug) = 0.2
- P(recovery | given_drug)

  = P(recovery | male ∧ given_drug)P(given_drug | male) + P(recovery | female ∧ given_drug)P(given_drug | female)

  = (0.6 x 30/40) + (0.2 x 10/40) = 0.5

Default reasoning

- Some conclusions are made by default unless a counter-evidence is obtained
  - Non-monotonic reasoning
- Points to ponder
  - What is the semantic status of default rules?
  - What happens when the evidence matches the premises of two default rules with conflicting conclusions?
  - If a belief is retracted later, how can a system keep track of which conclusions need to be retracted as a consequence?

Issues in Rule-based methods for Uncertain Reasoning

- Locality
  - In logical reasoning systems, if we have A ⇒ B, then we can conclude B given evidence A, without worrying about any other rules. In probabilistic systems, we need to consider all available evidence.
- Detachment
  - Once a logical proof is found for proposition B, we can use it regardless of how it was derived (it can be detached from its justification). In probabilistic reasoning, the source of the evidence is important for subsequent reasoning.
- Truth functionality
  - In logic, the truth of complex sentences can be computed from the truth of the components. Probability combination does not work this way, except under strong independence assumptions.

- A famous example of a truth functional system for uncertain reasoning is the certainty factors model, developed for the Mycin medical diagnostic program

Drawbacks of using Bayesian theory

- The probabilities are described as a single numeric point value.
  - Distortion to precision that is actually available for supporting evidence.
  - When we assert with probability 0.7 that the dollar will fall against the Japensese Yen over the next six months, what we really mean is we have a fairly strong conviction there is a chance of about 0.6 to 0.8 say, that it will fall
- No way to differentiate between ignorance and uncertainty.
  - Example
    - One of the three A, B, C terrorist group has planted a bomb. Let C found guilty and P(C) = 0.8. According to traditional theory rest of the probability will distributes amongst other without having any knowledege about them.
- Forced to regard belief and disbelief as functional opposite.
- Ex. If P(A) = 0.3 then P(~A)= 0.7 so that P(A) + P(~A) =1
- AS A REMEDY FOR THE ABOVE PROBLEMS, GENERALISES THEORY HAS BEEN PROPOSED BY ARTHUR DEMPSTER (1968) AND EXTENDED BY STUDENT GLENN SHAFER (1976).

Dempster-Shafer Theory

- Separate probability masses may be assigned to all subsets of a universe of discourse rather than just to individual single members.
  - It Permit the inequality P(A) + P(~A) <=1
    - It assume a universe of discourse U
    - A set corresponding to n proposition, exactly one of which is true.
    - The Propositions are assumed to be exhaustive and mutually exclusive.
- Designed to deal with the distinction between uncertainty and ignorance.
- We use a belief function Bel(X) – probability that the evidence supports the proposition
- When we do not have any evidence about X, we assign Bel(X) = 0 as well as Bel(¬X) = 0
- For example, if we do not know whether a coin is fair, then:
    - Bel(Heads) = Bel(¬Heads) = 0
- If we are given that the coin is fair with 90% certainty, then:
    - Bel(Heads) = 0.9 X 0.5 = 0.45
    - Bel(¬Heads) = 0.9 X 0.5 = 0.45
- Note that we still have a gap of 0.1 that is not accounted for by the evidence

Fuzzy Logic

- Fuzzy set theory is a means of specifying how well an object satisfies a vague description
  - Truth is a value between 0 and 1
  - Uncertainty stems from lack of evidence, but given the dimensions of a man concluding whether he is fat has no uncertainty involved
- The rules for evaluating the fuzzy truth, T, of a complex sentence are:
  - $T(A \wedge B) = \min(T(A), T(B))$
  - $T(A \vee B) = \max(T(A), T(B))$
  - $T(\neg A) = 1 - T(A)$

Summary

- Representation
  - Bayes nets compactly encode joint distributions (by making use of conditional independences!)
  - Guaranteed independencies of distributions can be deduced from BN graph structure
  - D-separation gives precise conditional independence guarantees from graph alone
  - A Bayes net's joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution
- Conditional Independences
- Probabilistic Inference
- Enumeration (exact, exponential complexity)
  - Variable elimination (exact, worst-case exponential complexity, often better)
  - Probabilistic inference is NP-complete
  - Sampling (approximate)
- Learning Bayes' Nets from Data

Summary

- Probabilistic reasoning is an integral part of many domains of AI. We intend to study the following in future –
  - Probabilities reasoning in state machines (Markov Chains)
    - Good for modeling dynamical systems, recurrent behavior
    - Reinforcement Learning methods work with Markov Decision processes
- You may also look up some of these for further reading –
  - Bayesian optimization is an advanced method for automated problem solving under limited knowledge of the state space
  - Bayesian learning methods are gaining in popularity for making classifiers more important
    - Uncertainty needs to be factored into classifiers, so that the classifier can separate out lack of knowledge as one of the outcomes
    - For example, if a ML classifier is trained to separate wolves from huskies, it should be able to say "I don't know" if presented with the picture of a cat
- Structures like Stochastic AND/OR Graphs are being conceived for explainable AI (XAI)

Home Work

- Three candidates run for an election as a major in a city. According to a public opinion poll, their chances to win are 0.25, 0.35 und 0.40. The chances that they build a bridge after they have been elected are 0.60, 0.90 and 0.80. What is the probability that the bridge will be built after the election?
- On an airport all passengers are checked carefully. Let T with t ∈ {0, 1} be the random variable indicating whether somebody is a terrorist (t = 1) or not (t = 0). Let A with a ∈ {0, 1} be the variable indicating arrest. A terrorist shall be arrested with probability P(A = 1|T = 1) = 0.98, a non-terrorist with probability P(A = 1|T = 0) = 0.001. One in a lakh passengers is a terrorist, P(T = 1) = 0.00001. What is the probability that an arrested person actually is a terrorist?

Home Work

A smell of sulphur (S) can be caused either by rotten eggs (E) or as a sign of the doom brought by the Mayan Apocalypse (M). The Mayan Apocalypse also causes the oceans to boil (B). The Bayesian network and corresponding conditional probability tables for this situation are shown below. For each part, you should give a numerical answer (e.g. 0.81). We use the short-form E' to denote the negation of E. No marks will be awarded unless intermediate steps are shown

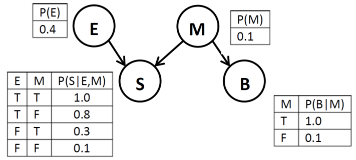
(a) Compute P(E'S'M'B')

(b) What is the probability that the oceans boil?

(c) What is the probability that the Mayan Apocalypse is occurring, given that the oceans are boiling?

(d) What is the probability that the Mayan Apocalypse is occurring, given that there is a smell of sulphur the oceans are boiling, and there are rotten eggs?

(e) What is the probability that rotten eggs are present, given that the Mayan Apocalypse is occurring?



P(E)
0.4

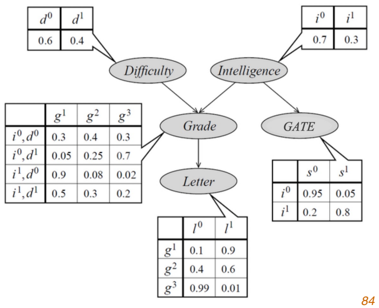
P(M)
0.1

E	M	P(S|E,M)
T	T	1.0
T	F	0.8
F	T	0.3
F	F	0.1

M	P(B|M)
T	1.0
F	0.1

Exercise

- P(i1, d0, g2, s1, l0)
  = P(i1)P(d0)P(g2 | i1, d0)P(s1 | i1)P(l0 | g2)
  = 0.3*0.6*0.08*0.8*0.4
  = 0.004608.



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Module 6: Reasoning under Uncertainty

- PART 6.1 : Quantifying Uncertainty
  - Basic of Probability
- PART 6.2 : Probablistic Reasoning
  - Bayes Rule
  - Bayesian Network
- PART 6.3 : Fuzzy Logic
- PART 6.4 : Probabilistic Reasoning over time
  - Hidden Markov Model
  - Kalman filter
  - Markov Chain Monte Carlo
- PART 6.5 : Decisions Theory
  - Utility Function
  - Decision Network
  - Markov Decision Proces

References

- ***Artificial Intelligence*** *by Elaine Rich & Kevin Knight*, Third Ed, Tata McGraw Hill
- ***Artificial Intelligence and Expert System*** *by Patterson*
- http://www.cs.rmit.edu.au/AI-Search/Product/
- http://aima.cs.berkeley.edu/demos.html (for more demos)
- ***Artificial Intelligence and Expert System*** *by Patterson*
- Slides adapted from CS188 Instructor: Anca Dragan, University of California, Berkeley
- Slides adapted from CS60045 ARTIFICIAL INTELLIGENCE