Introduction to Artificial Intelligence

Lecture 5: Probabilistic reasoning I

LIÈGE université

Announcement

Research seminar

"What does the Revolution in Artificial Intelligence Mean for Particle Physics?",

Prof. Kyle Cranmer, New York University.

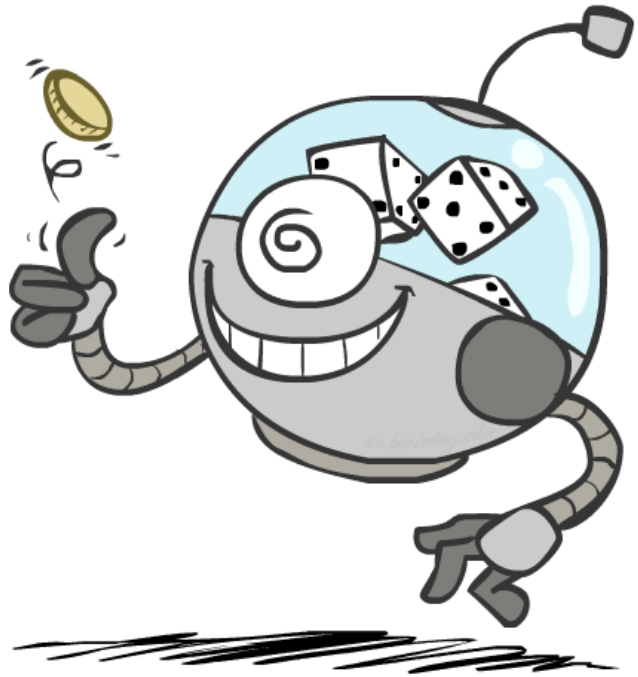
When/Where?

- Friday December 1, 2017
- 3:00 PM
- Room R7 / B28

Not mandatory, but you are welcome to attend!

Today

- Probability:
  - Random variables
  - Joint and marginal distributions
  - Conditional distributions
  - Product rule, Chain rule, Bayes' rule
  - Inference
- Bayesian networks:
  - Representing uncertain knowledge
  - Semantics
  - Construction

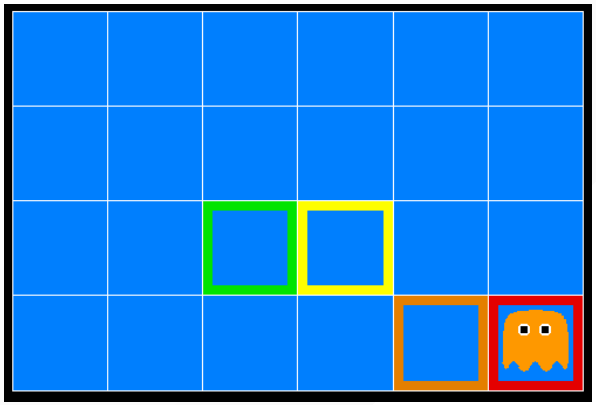


You will need these concepts a lot! (Now and in future courses)

Credits: UC Berkeley, CS188

Quantifying uncertainty

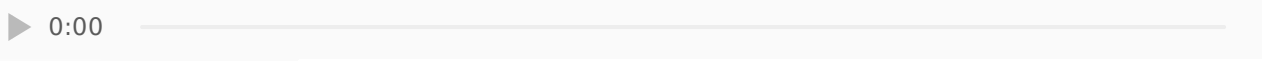
Inference in Ghostbusters



- A ghost is hidden in the grid somewhere.
- Sensor readings tell how close a square is to the ghost.
  - On the ghost: red
  - 1 or 2 away: orange
  - 3 away: yellow
  - 4+ away green
- Sensors are noisy, but we know $P(Color|Distance)$.

Credits: UC Berkeley, CS188

Inference in Ghostbusters



[Q] Could we use a logical agent for this game?

Credits: UC Berkeley, CS188

Uncertainty

- General situation:
  - Observed variables (evidence): agent knows certain things about the state of the world (e.g., sensor readings).
  - Unobserved variables: agent needs to reason about other aspects that are uncertain (e.g., where the ghost is).
  - (Probabilistic) model: agent knows or believes something about how the known variables relate to the unknown variables.
- How to handle uncertainty?
  - A purely logical approach either:
    - risks falsehood (because of ignorance about the world or laziness in the model), or
    - leads to conclusions that are too weak for decision making.
  - Probabilistic reasoning provides a framework for managing our knowledge and beliefs.

Probability

- Probabilistic assertions express the agent's inability to reach a definite decision regarding the truth of a proposition.
- Probabilities summarize effects of
  - laziness (failure to enumerate all world states)
  - ignorance (lack of relevant facts, initial conditions, correct model, etc).
- Bayesian (subjective) probabilities relate propositions to one's own state of knowledge.
  - e.g., $P(\text{ghost in } [3, 2]) = 0.02$
- These are not claims of a "probabilistic tendency" in the current situation (but might be learned from past experience of similar situations).

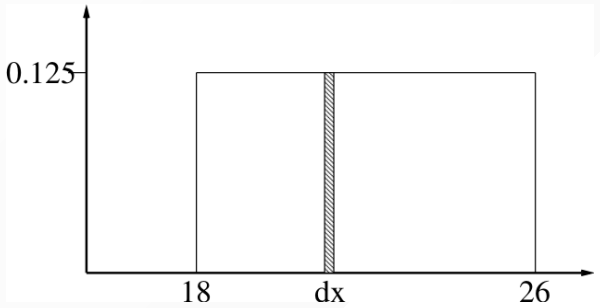
Probability basics

- Begin with a set Ω, the sample space.
  - e.g., 6 possible rolls of a die.
  - $\omega \in \Omega$ is a sample point, possible world or atomic event.
- A probability space is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ such that:
  - $0 \leq P(\omega) \leq 1$
  - $\sum_{\omega} P(\omega) = 1$
  - e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$

Random variables

- A random variable is a function $X : \Omega \rightarrow D_X$ from the sample space to some domain defining its outcomes.
  - e.g., $Odd(1) = true$ and $D_{Odd} = \{true, false\}$.
- P induces a probability distribution for any random variable X.
  - $P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$
  - e.g., $P(Odd = true) = P(1) + P(3) + P(5) = \frac{1}{2}$.
  - When clear from the context, we will denote $P(X = x_i)$ as $P(x_i)$.
- For discrete variables, the probability distribution can be encoded by a discrete list of the probabilities of the outcomes, known as the probability mass function.
- In practice, we will use random variables to represent aspects of the world about which we (may) have uncertainty.
  - R: Is it raining?
  - T: Is it hot or cold?
  - L: Where is the ghost?

Probability for continuous variables



- For continuous variables, the probability distribution can be described by a probability density function.
  - That is, the distribution is decribed by a parameterized function of value:
    - e.g., $P(X = x) = U[18, 26](x)$ for a uniform density between 18 and 26.
  - a density integrates to 1.
- That is, $P(X = 20.5) = 0.125$ really means

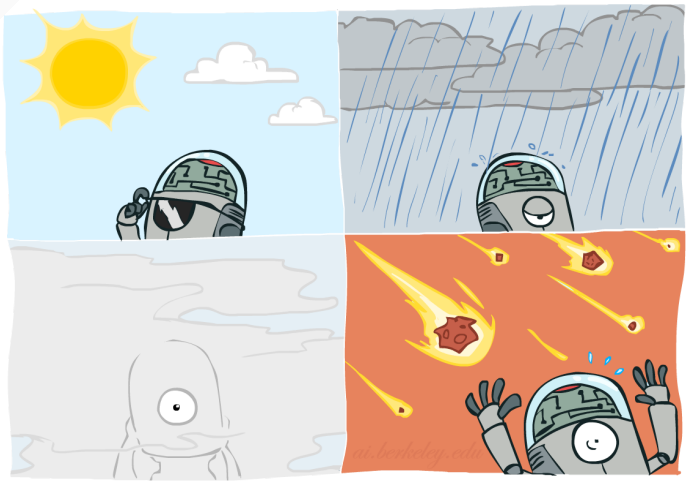
$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

Probability distributions

- Intuitively, one can think of the probability distribution of a random variable as a table that associates a probability value to each outcome of the variable.
- By construction, probability values are normalized (i.e., sum to 1).
- This table can be infinite!

$$P(W)$$

W	P
sun	0.6
$rain$	0.1
fog	0.3
$meteor$	0.0



Credits: UC Berkeley, CS188

Joint distributions

- A joint probability distribution over a set of random variables $X_1, \ldots, X_n$ specifies the probability of each (combined) outcome.

$$P(x_1, \ldots, x_n) = \sum_{\{\omega : X_1(\omega)=x_1, \ldots, X_n(\omega)=x_n\}} P(\omega)$$

- Example $P(T, W)$:

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

Events

- An event is a set E of outcomes.
  - $P(E) = \sum_{(x_1,...,x_n) \in E} P(x_1, ..., x_n)$
- From a joint distribution, the probability of any event can be calculated.
  - Probability that it is hot and sunny?
  - Probability that it is hot?
  - Probability that it is hot or sunny?
- Interesting events often correspond to partial assignments.
  - e.g., $P(T = hot)$

Marginal distributions

- The marginal distribution of a subset of a collection of random variables is the joint probability distribution of the variables contained in the subset.
- Intuitively, marginal distributions are sub-tables which eliminate variables.

$P(T, W)$

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

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

$$P(t) = \sum_w P(t, w) \qquad P(w) = \sum_t P(t, w)$$

[Q] To what events are marginal probabilities associated?

Conditional probability

- Prior or unconditional probabilities of an event correspond to some initial belief, prior to arrival of any evidence.
  - e.g., $P(W = sun) = 0.6$.
  - e.g., $P(\text{ghost in } [3, 2]) = 0.02$
- Posterior or conditional probabilities correspond to the probability of an event, given some observed evidence.
- Formally,

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

  - e.g., $P(W = sun|T = cold) = \frac{P(W=sun,T=cold)}{P(T=cold)} = \frac{0.2}{0.2+0.3} = 0.4$

Conditional distributions

- Conditional distributions are probability distributions over some variables, given fixed values for others.

$P(T, W)$

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

$P(W|T = hot)$

T	P
sun	0.8
$rain$	0.2

$P(W|T = cold)$

W	P
sun	0.4
$rain$	0.6

[Q] To what events are conditional probabilities associated?

[Q] Is a conditional distribution defined on the same sample and probability space than the joint?

Normalization trick

$P(T, W)$

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

$\rightarrow P(c, W)$

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

Select the joint probabilities matching the evidence c.

$\rightarrow P(W|c)$

W	P
sun	0.4
$rain$	0.6

Normalize the selection (make it sum to 1).

[Q] Why does this work?

Probabilistic inference (1)

- Probabilistic inference is the problem of computing a desired probability from other known probabilities (e.g., conditional from joint).
- We generally compute conditional probabilities.
  - e.g., $P(\text{on time}|\text{no reported accidents}) = 0.9$
  - These represent the agent's beliefs given the evidence.
- Probabilities change with new evidence:
  - e.g., $P(\text{on time}|\text{no reported accidents}, 5\text{AM}) = 0.95$
  - e.g., $P(\text{on time}|\text{no reported accidents}, \text{rain}) = 0.8$
  - e.g., $P(\text{ghost in } [3, 2]|\text{red in } [3, 2]) = 0.99$
  - Observing new evidence causes beliefs to be updated.

Credits: UC Berkeley, CS188

Probabilistic inference (2)

- General case:
  - Evidence variables: $E_1, ..., E_k = e_1, ..., e_k$
  - Query variables: Q
  - Hidden variables: $H_1, ..., H_r$
  - $(Q \cup E_1, ..., E_k \cup H_1, ..., H_r)$ = all variables $X_1, ..., X_n$
- We want to compute $P(Q|e_1, ..., e_k)$.

Inference by enumeration

1. Select the entries consistent with the evidence $E_1, ..., E_k = e_1, ..., e_k$.
2. Marginalize out the hidden variables to obtain the joint of the query and the evidence variables.

$$P(Q, e_1, ..., e_k) = \sum_{h_1,...,h_r} P(Q, h_1, ..., h_r, e_1, ..., e_k)$$

3. Normalize.

$$Z = \sum_q P(Q = q, e_1, ..., e_k)$$

$$P(Q|e_1, ..., e_k) = \frac{1}{Z} P(Q, e_1, ..., e_k)$$

Example

- $P(W)$?
- $P(W|winter)$?
- $P(W|winter, hot)$?

S	T	W	P
summer	*hot*	*sun*	0.3
summer	*hot*	*rain*	0.05
summer	*cold*	*sun*	0.1
summer	*cold*	*rain*	0.05
winter	*hot*	*sun*	0.1
winter	*hot*	*rain*	0.05
winter	*cold*	*sun*	0.15
winter	*cold*	*rain*	0.2

Complexity

- Inference by enumeration can be used to answer probabilistic queries for discrete variables (i.e., with a finite number of values).
- However, enumeration does not scale!
  - Assume a domain described by n variables taking at most d values.
  - Space complexity: $O(d^n)$
  - Time complexity: $O(d^n)$
- Can we reduce the size of the representation of the joint distribution?

The product rule

$$P(b)P(a|b) = P(a,b)$$

Example:

$P(W)$

W	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

$P(D,W)$

D	W	P
wet	sun	?
dry	sun	?
wet	$rain$	?
dry	$rain$	?

The chain rule

More generally, any joint distribution can always be written as an incremental product of conditional distributions.

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, ..., x_n) = \prod_i P(x_i|x_1, ..., x_{i-1})$$

[Q] Why is this always true?

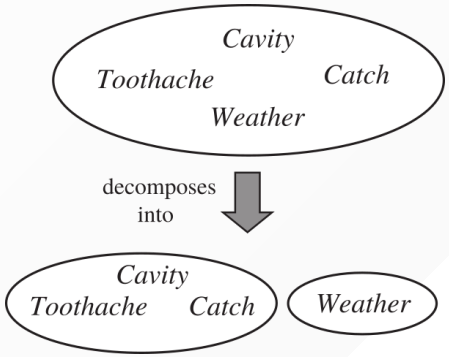
Independence

A and B are independent iff, for all $a \in D_A$ and $b \in D_B$,

- $P(a|b) = P(a)$, or
- $P(b|a) = P(b)$, or
- $P(a, b) = P(a)P(b)$

Independence is also written as $A \perp B$.

Independence example



$$P(\text{toothache}, \text{catch}, \text{cavity}, \text{weather}) = P(\text{toothache}, \text{catch}, \text{cavity})P(\text{weather})$$

- The original 32-entry table reduces to one 8-entry and one 4-entry table (assuming 4 values for $weather$ and boolean values otherwise).
- For n independent coin flips, the joint distribution can be fully factored and represented as the product of n 1-entry tables.
  - $2^n \rightarrow n$
- In practice, absolute independence is rare. What to do?

Conditional independence (1)

A and B are conditionally independent given C iff, for all $a \in D_A, b \in D_B$ and $c \in D_C$,

- $P(a|b,c) = P(a|c)$, or
- $P(b|a,c) = P(b|c)$, or
- $P(a,b|c) = P(a|c)P(b|c)$

Conditional independence is also written as $A \perp B|C$.

Conditional independence (2)

- Using the chain rule, the join distribution can be factored as a product of conditional distributions.
- Each conditional distribution may potentially be simplified by conditional independence.
- Conditional independence assertions can allow probabilistic models to scale up.

Conditional independence example

- Assume three random variables $\mathrm{Toothache}$, Catch and Cavity.
- Catch is conditionally independent of $\mathrm{Toothache}$, given Cavity.

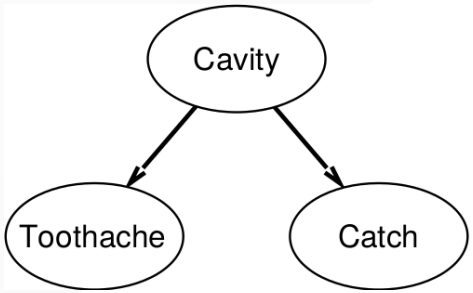
Therefore, we can write:

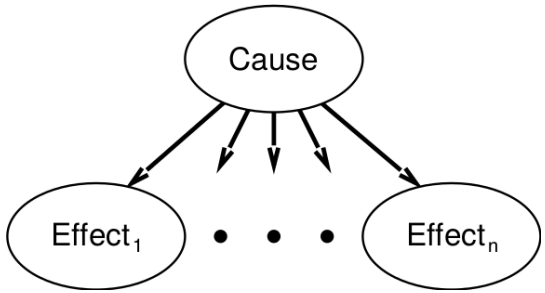
$$
\begin{aligned}
&P(\mathrm{toothache}, \mathrm{catch}, \mathrm{cavity}) \\
&= P(\mathrm{toothache}|\mathrm{catch}, \mathrm{cavity})P(\mathrm{catch}|\mathrm{cavity})P(\mathrm{cavity}) \\
&= P(\mathrm{toothache}|\mathrm{cavity})P(\mathrm{catch}|\mathrm{cavity})P(\mathrm{cavity})
\end{aligned}
$$

In this case, the representation of the joint distribution reduces to $2 + 2 + 1$ independent numbers (instead of $2^n - 1$).

Naive Bayes

- More generally, from the product rule, we have:
  - $P(\text{cause}, \text{effect}_1, ..., \text{effect}_n) = P(\text{effect}_1, ..., \text{effect}_n | cause) P(\text{cause})$
- Assuming pairwise conditional independence between the effects given the cause, it comes:
  - $P(\text{cause}, \text{effect}_1, ..., \text{effect}_n) = P(\text{cause}) \prod_i P(\text{effect}_i | cause)$
- This probabilistic model is called a naive Bayes model.
  - The complexity of this model is $O(n)$ instead of $O(2^n)$ without the conditional independence assumptions.
  - Naive Bayes can work surprisingly well in practice, even when the assumptions are wrong.





The Bayes' rule

- The product rule defines two ways to factor the joint distribution of two random variables.
  - $P(a, b) = P(a|b)P(b) = P(b|a)P(a)$
- Therefore,

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

- Why is this helpful?
  - Let us build one conditional from its reverse.
  - Often one conditional is tricky, but the other is simple.
  - This equation is the foundation of many AI systems.

Inference with Bayes' rule

Example: diagnostic probability from causal probability.

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

- $P(\text{effect}|\text{cause})$ quantifies the relationship in the causal direction.
- $P(\text{cause}|\text{effect})$ describes the diagnostic direction.

Let S=stiff neck and M=meningitis.

- Given
  - $P(s|m) = 0.7$
  - $P(m) = 1/50000$
  - $P(s) = 0.01$
- It comes:
  - $P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014$

Ghostbusters, revisited

- Let assume a random variable G for the ghost location and a set of random variables $R_{i,j}$ for the individual readings:
  - We define a prior distribution $P(G)$ over ghost locations.
    - Assume it is uniform.
  - We assume a sensor reading model $P(R_{i,j}|G)$.
    - That is, we know what the sensors do.
    - $R_{i,j}$ = reading color measured at $[i, j]$
    - e.g., $P(R_{1,1} = yellow|G = [1, 1]) = 0.1$
    - Two readings are conditionally independent, given the ghost position.
- We can calculate the posterior distribution $P(G|R_{i,j})$ using Bayes' rule:
  - $P(G|R_{i,j}) = \frac{P(R_{i,j}|G)P(G)}{P(R_{i,j})}$
- For the next reading $R_{i',j'}$, this posterior distribution becomes the prior distribution over ghost locations, which we update similarly.

Ghostbusters, revisited

Credits: UC Berkeley, CS188

Frequentism vs. Bayesianism

What do probability values represent?

- The objectivist frequentist view is that probabilities are real aspects of the universe.
  - i.e., propensities of objects to behave in certain ways.
  - e.g., the fact that a fair coin comes up heads with probability 0.5 is a propensity of the coin itself.
- The subjectivist Bayesian view is that probabilities are a way of characterizing an agent's beliefs.
  - i.e., probabilities do not have external physical significance.
  - This is the interpretation of probabilities that we will use!

Probabilistic reasoning

Representing knowledge

- The joint probability distribution can answer any question about the domain.
- However, its representation can be come intractably large as the number of variable grows.
- Independence and conditional independence reduce the number of probabilities that need to be specified in order to define the full joint distribution.
- These relationships can be represented explicitly in the form of a Bayesian network.

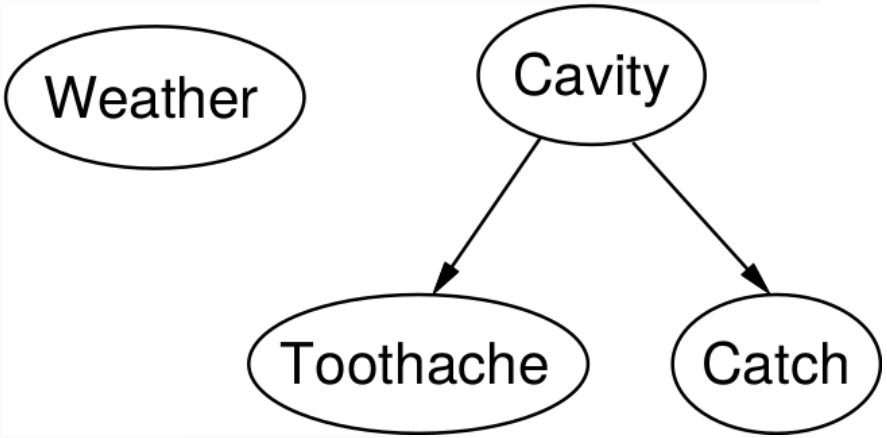
Bayesian networks

A Bayesian network is a directed graph in which:

- Each node corresponds to a random variable.
  - Can be observed or unobserved.
  - Can be discrete or continuous.
- Each edge indicates "direct influence" between variables.
  - If there is an arrow from node X to node Y, X is said to be a parent of Y.
  - The graph has no directed cycle (i.e., the graph is a DAG).
- Each node X_i is annotated with a conditional probability distribution $P(X_i|\text{parents}(X_i))$ that quantifies the effect of the parents on the node.
  - In the simplest case, conditional distributions are represented as conditional probability table (CTP).

Example (1)

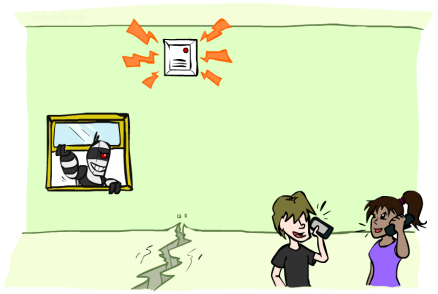
The topology of the network encodes conditional independence assertions:



- $\mathrm{Weather}$ is independent of the other variables.
- $\mathrm{Toothache}$ and Catch are conditionally independent given Cavity.

[Q] What is the BN for n independent coin flips?

Example (2a)

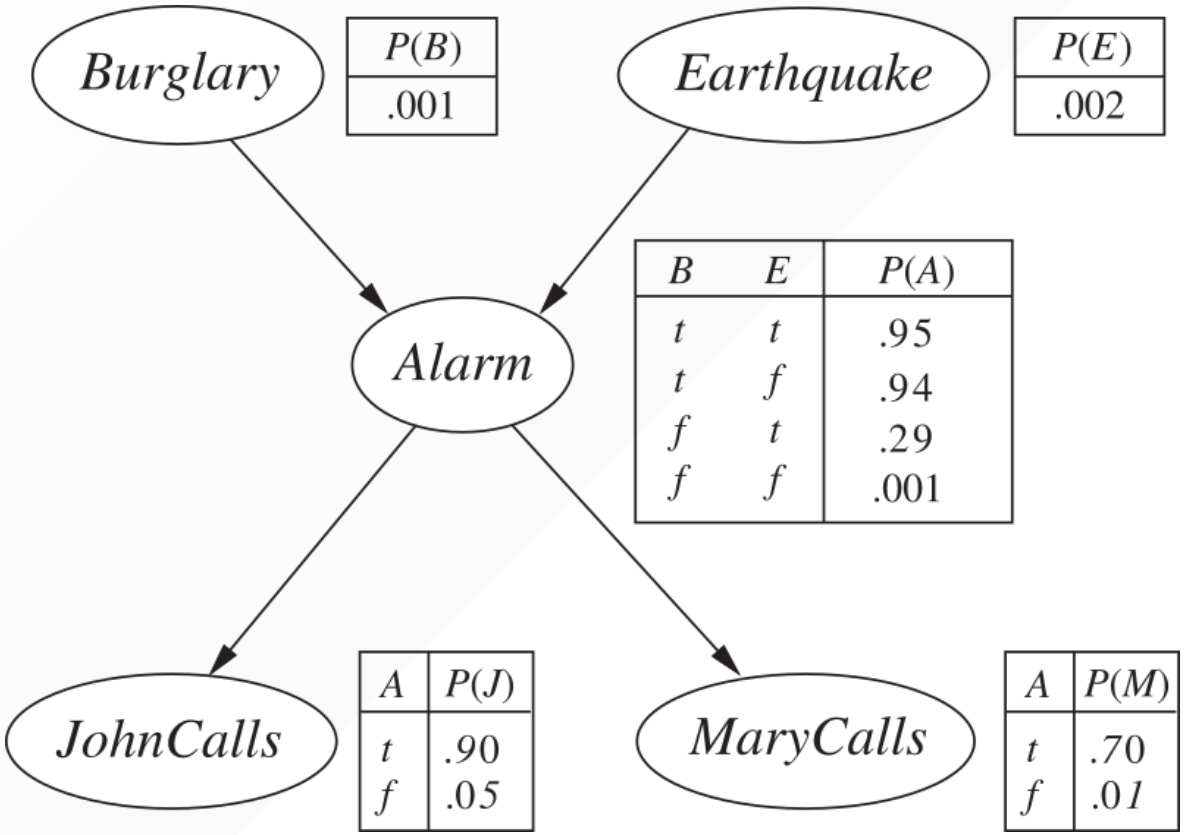


I am at work, neighbor John calls to say my alarm is ringing, but neighbor Mary does not call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: $\mathrm{Burglar}, \mathrm{Earthquake}, \mathrm{Alarm}, \mathrm{JohnCalls}, \mathrm{MaryCalls}$.
- Network topology from "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alaram off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Credits: UC Berkeley, CS188

Example (2b)


Burglary
P(B)
.001
Earthquake
P(E)
.002
Alarm
B E P(A)
t t .95
t f .94
f t .29
f f .001
JohnCalls
A P(J)
t .90
f .05
MaryCalls
A P(M)
t .70
f .01

Probabilities in BNs (1)

A Bayesian network implicitly encodes the full joint distribution as the product of the local distributions:

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

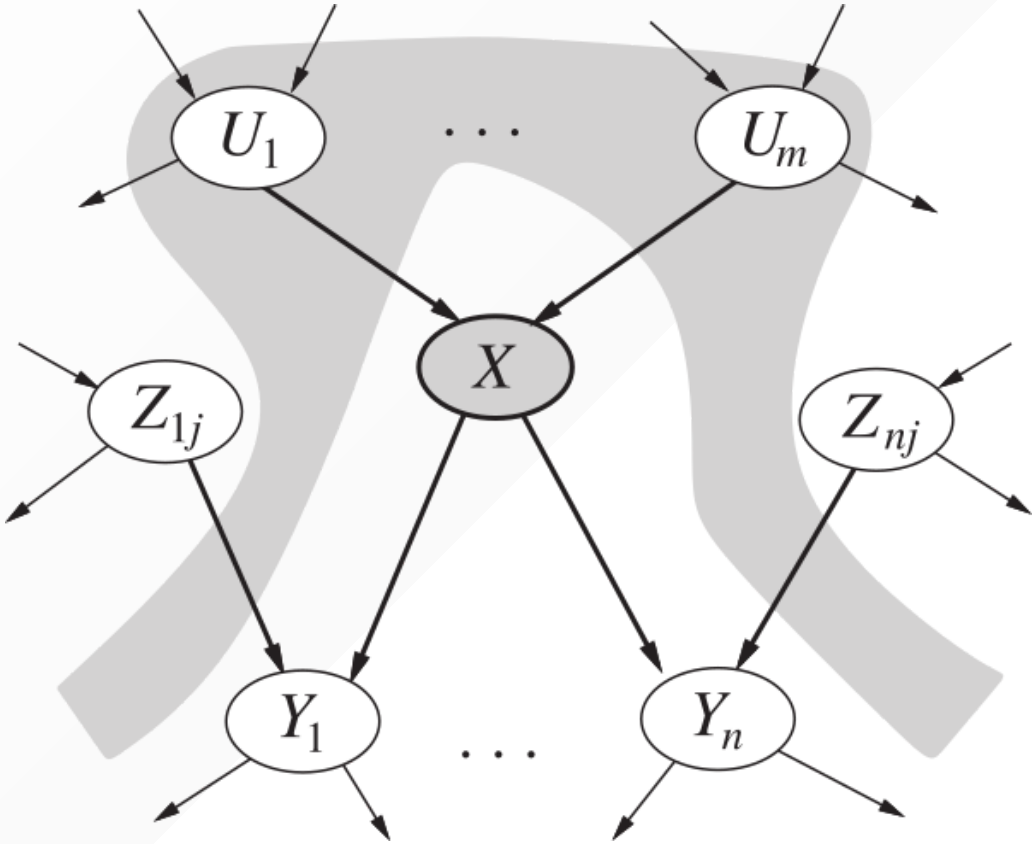
For example:

$$\begin{aligned} & P(j, m, a, \neg b, \neg e) \\ &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

Probabilities in BNs (2)

- Why does $\prod_{i=1}^{n} P(x_i|\text{parents}(X_i))$ result in a proper joint distribution?
- By the chain rule:
  - $P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|x_1, ..., x_{i-1})$
- Assume conditional independencies of X_i with its predecessors in the ordering given the parents, and provided $\text{parents}(X_i) \subseteq \{X_1, ..., X_{i-1}\}$:
  - $P(x_i|x_1, ..., x_{i-1}) = P(x_i|\text{parents}(X_i))$
- Therefore $P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i))$.

Local semantics



A node X is conditionally independent ot its non-descendants (the Z_{ij}) given its parents (the U_i').

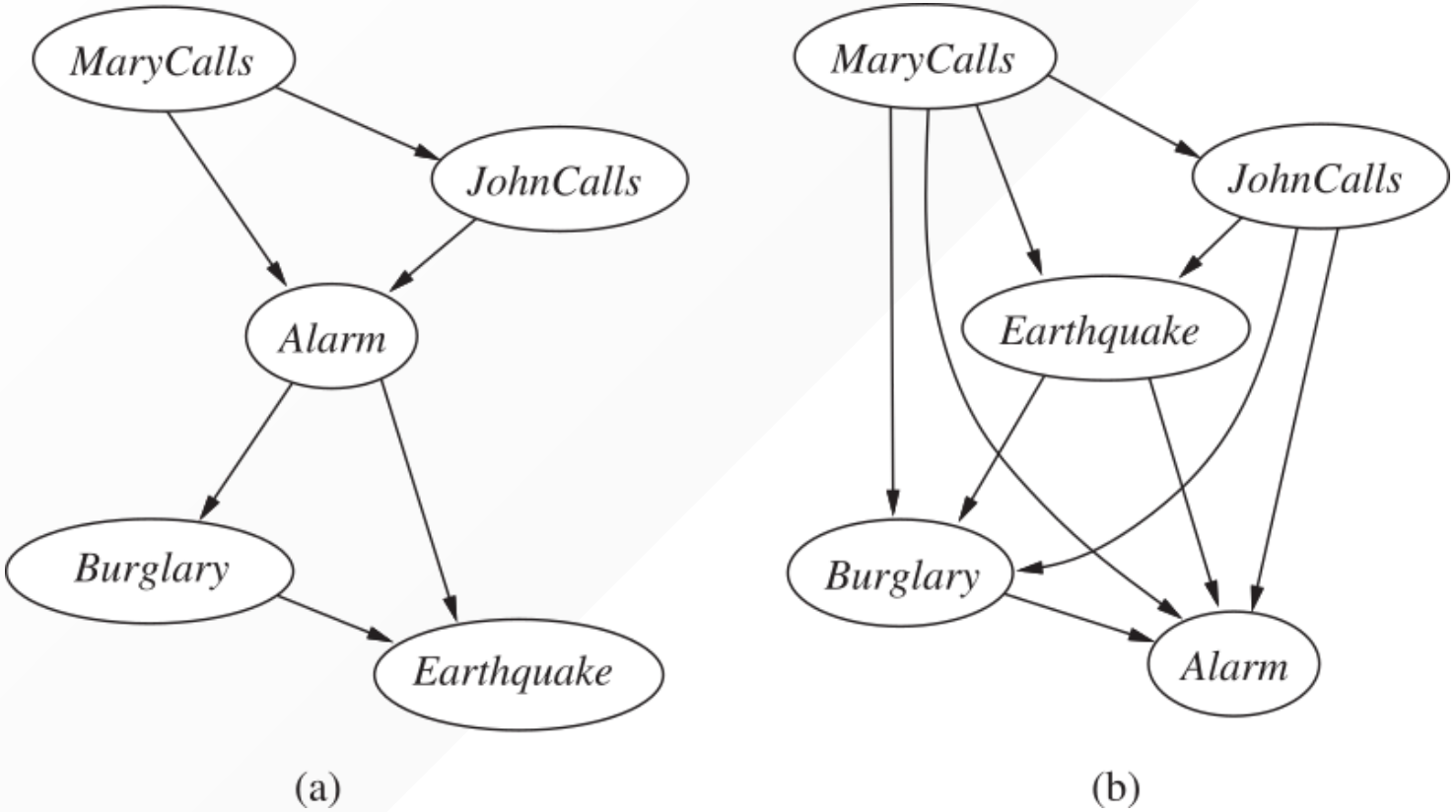
Constructing Bayesian Networks

BNs are correct representations of the domain only if each node is conditionally independent of its other predecessors in the node ordering, given its parents.

This suggests the following methodology for building Bayesian networks:

1. Choose an ordering of variables $X_1, ..., X_n$.
2. For $i = 1$ to n:
   - Add X_i to the network.
   - Select a minimal set of parents from $X_1, ..., X_n$ such that $P(x_i|x_1, ..., x_{i-1}) = P(x_i|\text{parents}(X_i))$.
   - For each parent, insert a link from the parent to X_i.
   - Write down the CPT.

Examples



These BNs are alternatives to Example 2b.

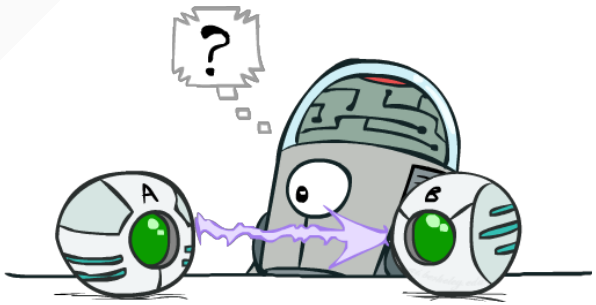
[Q] What do you think of these BNs?

Compactness

- A CPT for boolean X_i with k boolean parents has 2^k rows for the combinations of parent values.
- Each row requires one number p for $X_i = true$.
  - The number of $X_i = false$ is just $1 - p$.
- If each variable has no more than k parents, the complete network requires $O(n \times 2^k)$ numbers.
  - i.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution.
- For the burglary net, we need $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$).
- Compactness depends on the node ordering.
  - Compare the three networks representing the same joint distribution.

Causality?

- When the network reflects the true causal patterns:
  - Often more compact (nodes have fewer parents).
  - Often easier to think about.
  - Often easier to elicit from experts.
- But, Bayesian networks need not be causal.
  - Sometimes no causal network exists over the domain (e.g., if variables are missing).
  - Edges then reflect correlation, not causation.
- What do the edges really mean then?
  - Topology may happen to encode causal structure.
  - Topology really encodes conditional independence.



Credits: UC Berkeley, CS188

Summary

- Uncertainty arises because of laziness and ignorance. It is inescapable in complex non-deterministic or partially observable environments.
- Probabilistic reasoning provides a framework for managing our knowledge and beliefs.
- Bayesian networks are DAGs whose nodes correspond to random variables; each node has a conditional distribution for the node, given its parents.
- A Bayesian Network specifies a full joint distribution.
    - They are often exponentially smaller than an explicitly enumerated joint distribution.