Artificial Intelligence Week 4: Probabilities, Uncertainties, Bayesian Networks

Debasis Ganguly
University of Glasgow, Glasgow, UK

October 13, 2025

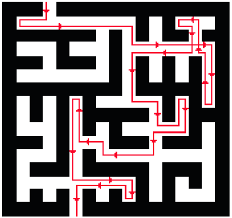
Overview

1. Probabilistic Approach to address Uncertainties
2. Bayesian Networks - probability for knowledge representation and reasoning
3. Inference on Bayesian Nets

Learning Objectives

- Why **representation of uncertainty in agents**, and why **probability theory** helps.
- Understand the **basics of probability theory** including sample space, events, joint distributions, conditionals.
- Understand, explain and be able to apply **Bayes Theorem** to various problems.
- Understand and explain how **Bayesian networks** can **represent knowledge**.
- Explain and apply **exact inference** for simple networks.
- Appreciate the role of sampling methods for **approximate inference**.

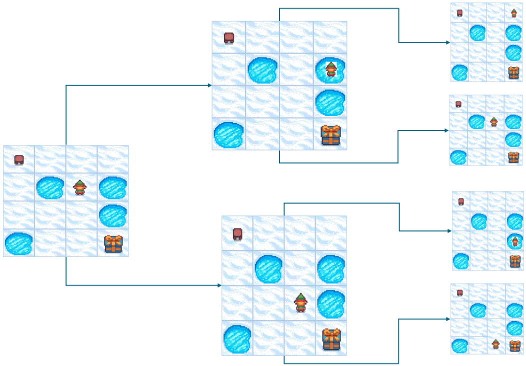
Partial Observability and Uncertain Outcomes



Have studied how a rational agents should function for:

- Fully observable (full access to state-space at anytime, noise free)
- Known (environment and rules)
- Deterministic outcomes (no uncertainty about where you end up after taking an action)
- Static (e.g. the world does not change while we think about what to do next).

Actions in Stochastic Environment



- Easiest to think of this with the "**many world**" interpretation.
- Imagine the agent can be in multiple states.
- We can then continue processing from each of these states with a deterministic algorithm like A*.
- Importantly, somewhere down the line the best among the "many worlds" has to **collapse** into the one real world.

Logic-based Approach to Tackle Uncertainty

Rule-based Planning

- A decision is a hard-coded function of several Boolean attributes.

- Example: An agent that drives a taxi to the airport.
  - A_t: Agent decides to leave your home t minutes before departure.
  - Will A_t ensure that you catch you flight?
- Issues with a rule-based planning:
  - Define the Boolean attributes: "flat tire", "icy roads", "other traffic", "enough fuel".
  - Yes, A_{90} will get there in time if: i) "no flat tire", "no icy roads", "no other traffic", "enough fuel".
  - But are all these really Boolean (0/1) attributes?
  - **Too many rules** for practical decision making!

Rational Decision in Stochastic Environment

- What about A_{180}? Almost certainly gets in time but you waste a lot of time in the airport.
- Rational decision making:
  - **Relative importance** of various goals - Utility Theory (Next week).
  - **Likelihood** that they will be achieved - Probability Theory (Now).

Summarizing Uncertainty

Logic Rules

- Enumerate all possibilities (cause-effects)
- Not always easy! For example:
  - Toothache $\implies$ Cavity (not always correct!)
  - Toothache $\implies$ Cavity or Gum Problem or ...
  - Cavity $\implies$ Toothache (not always correct either!)
- Make a (contingency) plan for ALL possible eventualities (e.g. all possible sensor outcomes $\rightarrow$ **grows arbitrarily large**)

Limitations

- Need a more generic framework than rules connected by logic operators.
- Need a language to describe and represent uncertainty of a belief state

Probability

- Describes the degree of belief in a current state of the world (possibly as explained by the evidence)
- For example:
  - A_{90} will get me there with probability 0.85 (given all the known and unknown factors in the world)
  - If a patient has a toothache there is a 0.8 probability that he has a cavity (e.g. based on previous experience)

Probability Introduction

- An experiment (or trial) is an occurrence with an uncertain outcome.
  - E.g., we don't know the outcome before rolling a dice.
  - For example: the outcome of a dice throw is 2.
- **Sample space**: A set Ω specifies all possible world states (exhaustively enumerates all possible worlds states).
  - For a dice there are 6 atomic events / sample points $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$.
- ω is a sample point /atomic event in Ω.
  - E.g. ⚀.

Classical Definition of Probability

- Probability of an atomic event A is the number of times we observe A (e.g., outcome of an experiment) from the **sample space** out of the total number of outcomes in the sample space.
  - $P(A) = n_A/n$
- Not always easy to count these (one must be careful! It's easy to make mistakes). Look at the following example:
- After rolling two die, find the probability that the sum is 7.
  - Total number of sums $(n) = \{2, \ldots, 12\}$. $P(7) = 1/11$ (Correct?)
  - Count the favourable pairs - {(⚂, ⚃), (⚄, ⚁), (⚅, ⚀)}; $P(7) = 6/36 = 1/6$ (Correct?).

Axiomatic Definition of Probability

- $P(A) > 0$, where $A \subset \Omega$
- $P(A) < 1 \; \forall A \subset \Omega$
- $P(A) = 0$ if $A = \emptyset$
- $P(A \cup B) = P(A) + P(B)$ if A and B are **mutually exclusive** events
- With the above axioms you can derive that: $P(AB) = P(A)P(B)$ if A and B are **independent**.

Probabilistic Thinking isn't Natural to Humans!

Linda Problem

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
Which one is more probable?

1. Linda is a **bank teller**.
2. Linda is a **bank teller** and is **active in the feminist movement**.

Probabilistic Thinking isn't Natural to Humans!

Linda Problem

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
Which one is more probable?

1. Linda is a **bank teller**.
2. Linda is a **bank teller** and is **active in the feminist movement**.

- A: Event that Linda is a bank teller.
- B: Event that Linda active in feminist movement.
- A and B are independent? Yes.
- $P(AB) = P(A)P(B) < P(A)$. So, option 1 is more probable.
- This is an example of **stereotype bias in humans**.

Probabilistic Thinking isn't Natural to Humans!

- Consider a bin: $\{{\color{green}\bullet}, {\color{green}\bullet}, {\color{green}\bullet}, {\color{green}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}\}$
- Consider the results of the following three "**sampling with replacement**" trials.
- Which one is more likely?
  1. ${\color{red}\bullet}, {\color{green}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}$
  2. ${\color{green}\bullet}, {\color{red}\bullet}, {\color{green}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}$
  3. ${\color{green}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}, {\color{red}\bullet}$

Probabilistic Thinking isn't Natural to Humans!

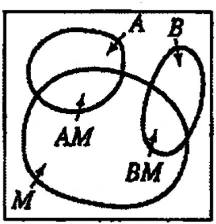
- Consider a bin: $\{\textcolor{green}{\bullet}, \textcolor{green}{\bullet}, \textcolor{green}{\bullet}, \textcolor{green}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}\}$
- Consider the results of the following three "**sampling with replacement**" trials.
- Which one is more likely?
  1. $\textcolor{red}{\bullet}, \textcolor{green}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}$
  2. $\textcolor{green}{\bullet}, \textcolor{red}{\bullet}, \textcolor{green}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}$
  3. $\textcolor{green}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}$
- '2' can't be more likely than '1' because the probability of observing a particular sequence of length 6 has to be less than that of a length of 5.
- Work out the probabilities of the sequence yourselves as an exercise.

Sample Numerical Problem

A box contains m white balls and n black balls. If balls are drawn **at random without replacement**, find the probability of seeing a white ball by the k-th draw.

- $W_k =$ White ball is drawn by the k-th draw.
  $\{\circ\}, \{\bullet, \circ\}, \{\bullet, \bullet, \circ\} \ldots, \{\bullet, \ldots, \bullet, \circ\}$.
- $X_i = \{i$ black balls followed by a white ball is drawn$\}$
- $W_k = X_0 \cup X_1 \cup X_{k-1}$ – These are all **mutually exclusive** events.
- By probability axiom: $P(W_k) = \sum_{i=0}^{k-1} P(X_i)$
- $P(X_0) = m/(m+n)$; $P(X_1) = n/(m+n) \times m/(m+n-1)$ and so on.

Conditional Probability



- Probability that A is observed given that M is already observed: $P(A|M) = \frac{P(AM)}{P(M)}$.
  - Example: $P(⚁|\text{an even number is seen}) = P(⚁)/P(\text{even}) = \frac{1}{6}/\frac{1}{2} = 1/3$.
- If $A \subset M$, $P(A|M) \geq P(A)$ (Why?)

Probability axioms hold true for any conditional M

- $P(A|M) > 0$
- $P(S|M) = 1$ $(M \subset S)$
- $P(A \cup B|M) = \frac{P(AM)+P(BM)}{P(M)}$

Numerical Example

A box contains 3 white balls $\{w_1, w_2, w_3\}$ and 2 red balls $\{r_1, r_2\}$. What is the probability that a white ball **gets removed** (draws w/o replacement) before a red one?

Solution w/o conditional probabilities

- Space of all ordered pairs: (w_1, w_2), (w_1, r_1) and so on.
- $\#$pairs $= 5 \times 4 = 20$ Why?
- Favourable pairs $= 6/20 = 3/10$.

Solution w/ conditional probabilities (more elegant)

- $P(W_1) = 3/5$ (event: white ball first).
- $P(R_2|W_1) = 2/4$
- $P(W_1 R_2) = P(R_2|W_1) \times P(W_1) = 2/4 \times 3/5 = 3/10$

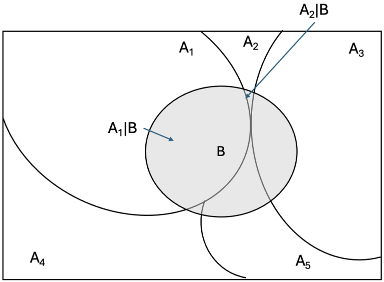
Bayes Theorem (Overview)

- A summarized view:

$$P(cause|effect) \propto P(effect|cause)P(cause)$$

- Used to estimate the probabilities from the **causal direction** (aka **priors**) to the **diagnostic direction** (aka **posteriors**).
- Note that the posterior is a function of two different types of priors - one conditional: $P(effect|cause)$ and the other unconditional: $P(cause)$.
- The conditional needs to be look into associations of effects and causes from the past data.

Bayes Theorem (More Formal Description)



- A_i - Hypotheses/Causes (forms a partition over the set of all possibilities)
- B - Evidence/Effect, i.e., one that is observed.
- Bayes Theorem - The **most likely cause** that has led to this **observation**.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

$$= \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

- Depends on:
  - Prior associations of how a cause can lead to an effect.
  - Prior likelihood of a cause itself.

A Visual Illustration

- Two identical looking bins: A ($\color{red}\bullet$, $\color{green}\bullet$, $\color{green}\bullet$, $\color{red}\bullet$, $\color{red}\bullet$), and B ($\color{green}\bullet$, $\color{green}\bullet$, $\color{green}\bullet$, $\color{red}\bullet$, $\color{green}\bullet$)
- You are blind-folded and asked to select a ball from a bin (you don't know which bin that is).
- **Question**: You observe a $\color{red}\bullet$ ball. What is the likelihood that it came from bin B?

A Visual Illustration

- Two identical looking bins: A (${\color{red}\bullet}$, ${\color{green}\bullet}$, ${\color{green}\bullet}$, ${\color{red}\bullet}$, ${\color{red}\bullet}$), and B (${\color{green}\bullet}$, ${\color{green}\bullet}$, ${\color{green}\bullet}$, ${\color{red}\bullet}$, ${\color{green}\bullet}$)
- You are blind-folded and asked to select a ball from a bin (you don't know which bin that is).
- **Question**: You observe a ${\color{red}\bullet}$ ball. What is the likelihood that it came from bin B?
- Compute the priors:
  - $P({\color{red}\bullet}|A) = 3/5$, $P({\color{green}\bullet}|A) = 2/5$, $P({\color{red}\bullet}|B) = 1/5$, $P({\color{green}\bullet}|B) = 4/5$
  - P(A) = P(B) = 1/2 (No other information is given)
  - $P(B|{\color{red}\bullet}) = P({\color{red}\bullet}|B)P(B)/\mathrm{P}({\color{red}\bullet})$
  - $= P({\color{red}\bullet}|B)P(B)/\ (P({\color{red}\bullet}|A)P(A) + P({\color{red}\bullet}|B)P(B))$
  - $= \frac{1/5\times 1/2}{3/5\times 1/2 + 1/5\times 1/2} = \frac{1/10}{3/10+1/10} = 1/4.$

Another numerical problem

The curious case of a cab - A common psychological test

A cab was involved in an accident. Two cab companies - the • and the • operate in the city with g and b number of cabs, respectively. A witness identified the cab involved in the accident as • (chance of error in the testimony due to poor light conditions is α). Probability that the cab involved in the accident was •?

- Most people just go with the witness guessing that the probability is close to $1 - \alpha$.
- We consider the following random variables.
  - $C \in \{G, B\}$: true color of the cab that was involved in the accident.
  - $O \in \{G, B\}$: observed color by the witness.

  Need to compute $P(C = G|O = B) = P(O = B|C = G)P(C = G)/(P(O = B|C = G)P(C = G) + P(O = B|C = B)P(C = B))$.
- $P(C = G|O = B) = \frac{\alpha g/(g+b)}{\alpha(g/(g+b))+(1-\alpha)b/(g+b)}$
- What happens when g increases? What happens when α decreases?

Inference by Enumeration

Three Boolean variables

- Toothache
- Cavity
- Catch (the dentist's nasty steel probe to remove a tooth)

The full joint distribution is a 2×2×2 table:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

For any proposition ϕ relating to the values of these random variables, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Inference by Enumeration (Working example)

	$toothache$		$\neg\, toothache$	
	$catch$	$\neg\, catch$	$catch$	$\neg\, catch$
$cavity$	.108	.012	.072	.008
$\neg\, cavity$	.016	.064	.144	.576

$P(cavity) = 0.108 + 0.012 + .0072 + 0.008 = 0.2$

Also called the marginal probability distribution of $Cavity = true$.

Inference by Enumeration (Working example)

Example:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by Enumeration (Working example)

Example:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

$$P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by Enumeration (Working example)

Determine $P(\neg cavity|toothache) =??$

Hint: Product rule or Bayes rule.

Solution: The product rule states: $P(\neg a|b) = P(\neg a \wedge b)/P(b)$

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Note: this is effectively an indirect use of Bayes theorem

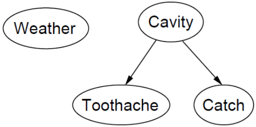
Bayesian Network

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions.
- Syntax:
  - a set of nodes, one per random variable a directed, acyclic graph (link $\rightarrow$ "directly influences")
  - a conditional distribution for each node given its parents:

$$P(X_i|Parents(X_i))$$

  - In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each possible state of the parent variables.

Bayesian Network

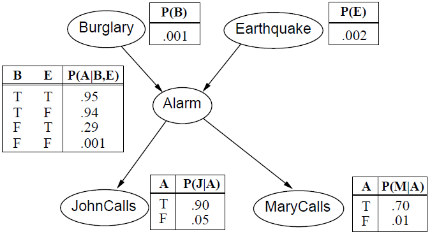


- The topology of the network encodes conditional independence assertions.
- Weather is independent of the other variables.
- *Toothache* and *Catch* are conditionally independent given *Cavity*

$$P(Toothache, Catch, Cavity, Weather) = P(Toothache \mid Cavity) P(Catch \mid Cavity) P(Cavity) P(Weather)$$

Burglary Example by Judea Pearl

- Problem Statement: I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes the alarm is set off by a minor earthquake. Is there a burglar?
- Identify the **Variables**: Burglar, Earthquake, Alarm, JohnCalls, MaryCalls
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off (rare event)
  - An earthquake can set the alarm off (rare event)
  - The alarm can cause Mary to call you (not very reliable)
  - The alarm can cause John to call you (fairly reliable)



P(B)
.001

P(E)
.002

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

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

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

Burglary Example (Contd.)

- Each node is conditionally independent of its nondescendants given its parents.
  - j is independent of b, and e, given the value of a.
- Full joint distribution as the product of the local conditional distributions:

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

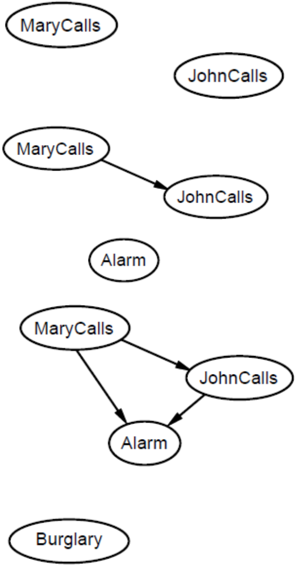
John and Mary both called, Alarm sounded, but no Burglary and no Earthquake

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

Bayesian Network Construction

- Construct the network such that a series of locally testable assertions of conditional independence guarantees the required global semantics.
- **Nodes**: Choose an ordering of variables $X_1, \ldots, X_n$
- Any will do but more compact if the **causes precedes effects**.
- **Connections**:
  - For $i = 1 \ldots n$
  - Add X_i to the network
  - Select minimal set of parents from $X_1, ..., X_{i-1}$ such that $P(X_i | Parents(X_i)) = P(X_i | X_1, \ldots, X_{i-1})$
  - Add link from parent(s) to X_i
  - Write down the CPT such that $P(X_i | Parents(X_i))$

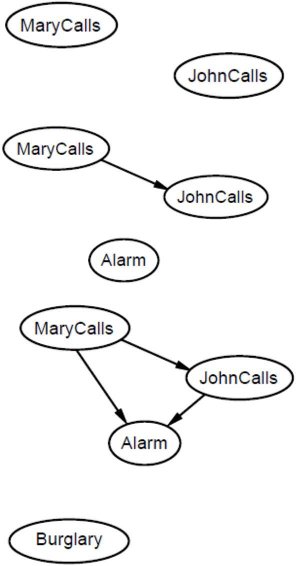
Bayesian Network Construction



Suppose we choose the ordering M, J, A, B, E

- Step 1: Add *MaryCalls* (no parents)
- Step 2: Add *JohnCalls*
  - Check $P(J|M) = P(J)$? No
  - If Mary calls then it is likely that the alarm has gone off and John will also call.

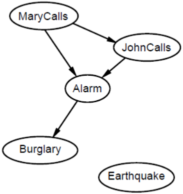
Bayesian Network Construction

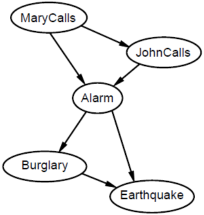


Suppose we choose the ordering M, J, A, B, E

- Step 1: Add *MaryCalls* (no parents)
- Step 2: Add *JohnCalls*
  - Check $P(J|M) = P(J)$? No
  - If Mary calls then it is likely that the alarm has gone off and John will also call.
- Step 3: Add *Alarm*
  - $P(A|J, M) = P(A|M)$? No
  - $P(A|J, M) = P(A|J)$? No
  - $P(A|J, M) = P(A)$? No
  - If both call, it is more likely that the alarm has gone off than if just one or neither calls, so we need both *MaryCalls* and *JohnCalls* as parents.
- Step 3: Add *Burglary*

Bayesian Network Construction



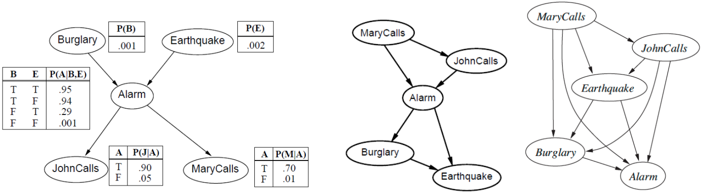


- Step 4: Add *Burglary* (no parents)
  - $P(B|A, J, M) = P(B|A)$? Yes
  - If we know the alarm state, we don't need to rely on the more uncertain events of J or M calling.
  - $P(B|A, J, M) = P(B)$? No

    Here we don't know the alarm state!
    Alarm gives us information about whether there is a burglary.
- Step 5: Add *Earthquake*
  - $P(E|B, A, J, M) = P(E|A)$? No

    Alarm ON $\rightarrow$ the likelihood of an earthquake changes if there has been a burglary or not.
    Implies the probability of an earthquake would only be **slightly above normal**.
  - Hence, we need **both Alarm and Burglary as parents**.

Different ordering leads to different BNs



- Left: **Causal model**. Easier to explain the arrows.
- Middle: **Diagnostic model**. More dependencies introduced, e.g., the arrow between *Burglary* and *Earthquake*.
- Right: Bad node ordering $\rightarrow$ Yet more complex and 'difficult to explain' model.
- They all represent the **same joint distribution**.

Inference on Bayesian Nets

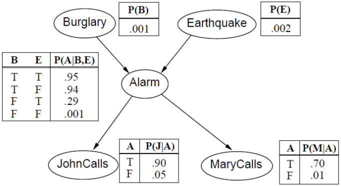
Types of inference methods

- Exact inference by enumeration

Not examinable

- Approximate inference by Markov Chain Monte Carlo (MCMC)

Naive Enumeration



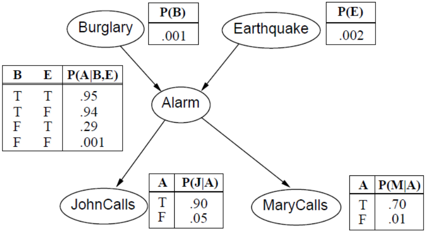
Use the basic rules of probability/Bayes and sum across relevant elements.

$$P(B|j,m) = \frac{P(B,j,m)}{P(j,m)}$$
$$= \alpha P(B,j,m)$$
$$= \alpha \sum_e \sum_a P(B,e,a,j,m)$$

Naive Enumeration

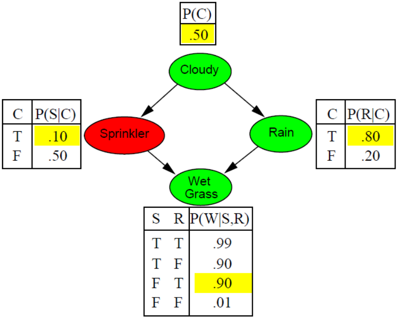
- Rewrite full joint entries using product of CPT entries:

$$
\begin{aligned}
P(B|j,m) &= \alpha \sum_e \sum_a P(B)P(e)P(a|B,e)P(j|a)P(m|a) \\
&= \alpha P(B) \sum_e P(e) \sum_a P(a|B,e)P(j|a)P(m|a)
\end{aligned} \tag{1}
$$



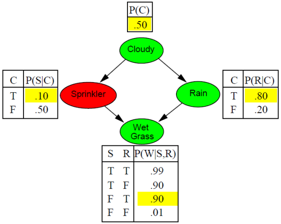
- **Exact**: Yes
- **Space**: $\mathcal{O}(n)$
- **Time complexity**: $\mathcal{O}(2^n)$ for a Boolean network
- In general: polynomial time on general trees (NP-hard on general graphs)
- **Issue**: Inefficient, since repeated computation e.g., computes $P(j|a)P(m|a)$ for each value of a.

Sampling-based



- Sample from $P(Cloudy) =< 0.5, 0.5 >$, value is true.
- Sample from $P(Sprinkler \mid Cloudy = true) =< 0.1, 0.5 >$, value is false.
- Sample from $P(Rain \mid Cloudy = true) =< 0.8, 0.2 >$, value is true.
- Sample from $P(WetGrass \mid Sprinkler = false, Rain = true) =< 0.9, 0.1 >$, value is true.
- Sampled events $[true, false, true, true]$.

Sampling-based



- Probability that the procedure generates a particular event $S_{PS}(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i)) = P(x_1, \dots, x_n)$
- In general, let $N_{PS}(x_1, \dots, x_n)$ be the number of samples generated for event $x_1, \dots, x_n$

$$\begin{aligned} \lim_{N\to\infty} P'(x_1 \dots x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \dots, x_n)/N \\ &= S_{PS}(x_1, \dots, x_n) \\ &= P(x_1 \dots x_n) \end{aligned} \tag{2}$$

- i.e., $P'(x_1, \dots, x_n) \approx P(x_1 \dots, x_n)$.
- **Issue**: Need a huge number of samples.

Summary

- Probabilistic reasoning: advantages over logical reasoning when there is not enough information to be sure actions will work.
- Belief networks / Bayesian Networks
  - Data structures for representing dependence among variables
  - Joint probability distribution
  - Cause-effect relationships
  - Inference: computing the p.d.f. of a subset of variables, given a set of evidence variables.
- Next Week:
  - Study Utility Theory
  - Combine utility with probabilistic reasoning for **decision making under uncertainty**.