

Foundations of AI

7. Making Simple Decisions under Uncertainty

Probability Theory, Bayesian Networks,
Other Approaches

*Wolfram Burgard, Andreas Karwath,
Bernhard Nebel, and Martin Riedmiller*

Motivation

- In many cases, our knowledge of the world is **incomplete** (not enough information) or **uncertain** (sensors are unreliable).
- Often, rules about the domain are incomplete or even incorrect – in the *qualification problem*, for example, what are the preconditions for an action?
- We have to act in spite of this!
- Drawing conclusions under uncertainty

Contents

- Motivation
- Foundations of Probability Theory
- Probabilistic Inference
- Bayesian Networks
- Alternative Approaches

Example

- **Goal:** Be in Freiburg at 9:15 to give a lecture.
- There are several **plans** that achieve the goal:
 - P_1 : Get up at 7:00, take the bus at 8:15, the train at 8:30, arrive at 9:00 ...
 - P_2 : Get up at 6:00, take the bus at 7:15, the train at 7:30, arrive at 8:00 ...
 - ...
- All these plans are correct, but
 - They imply different **costs** and different **probabilities** of actually achieving the goal.
 - P_2 eventually is the plan of choice, since giving a lecture is very important, and the success rate of P_1 is only 90-95%.

Uncertainty in Logical Rules (1)

Example: Expert dental diagnosis system.

$$\forall p[\text{Symptom}(p, \text{toothache}) \Rightarrow \text{Disease}(p, \text{cavity})]$$

→ This rule is *incorrect!* Better:

$$\forall p[\text{Symptom}(p, \text{toothache}) \Rightarrow \text{Disease}(p, \text{cavity}) \vee \text{Disease}(p, \text{gum_disease}) \vee \dots]$$

... but we don't know all the causes.

Perhaps a *causal* rule is better?

$$\forall p[\text{Disease}(p, \text{cavity}) \Rightarrow \text{Symptom}(p, \text{toothache})]$$

→ Does not allow to reason from symptoms to causes & is still wrong!

13/5

Uncertainty in Facts

Let us suppose we wanted to support the localization of a robot with (constant) landmarks. With the availability of landmarks, we can narrow down on the area.

Problem: **Sensors** can be imprecise.

- From the fact that a landmark was perceived, we cannot conclude with certainty that the robot is at that location.
- The same is true when no landmark is perceived.
- Only the **probability increases or decreases**.

7/7

Uncertainty in Rules (2)

- We cannot enumerate all possible causes, and even if we could...
- We do not know how correct the rules are (in medicine)
- ... and even if we did, there will always be uncertainty about the patient (the coincidence of having a toothache and a cavity that are unrelated, or the fact that not all tests have been run)
- Without perfect knowledge, logical rules do not help much!

7/6

Degree of Belief and Probability Theory (1)

- We (and other agents) are convinced by facts and rules only up to a certain degree.
- One possibility for expressing the **degree of belief** is to use **probabilities**.
- The agent is 90% (or 0.9) convinced by its sensor information = in 9 out of 10 cases, the information is correct (the agent believes).
- Probabilities sum up the “uncertainty” that stems from lack of knowledge.
- Probabilities are not to be confused with vagueness. The predicate *tall* is **vague**; the statement, “A man is 1.75–1.80m tall” is **uncertain**.

7/8

Uncertainty and Rational Decisions

- We have a choice of **actions** (or plans).
- These can lead to different solutions with different **probabilities**.
- The **actions** have different (subjective) **costs**.
- The **results** have different (subjective) **utilities**.
- It would be rational to choose the action with the maximum **expected total utility**!

Decision Theory = Utility Theory + Probability Theory

7/9

Unconditional Probabilities (1)

$P(A)$ denotes the **unconditional** probability or **prior** probability that A will appear *in the absence of any other information*, for example:

$$P(\text{Cavity}) = 0.1$$

Cavity is a proposition. We obtain prior probabilities from statistical analysis or general rules.

7/11

Decision-Theoretic Agent

```
function DT-AGENT(percept) returns an action
  static: a set probabilistic beliefs about the state of the world

  calculate updated probabilities for current state based on
    available evidence including current percept and previous action
  calculate outcome probabilities for actions,
    given action descriptions and probabilities of current states
  select action with highest expected utility
    given probabilities of outcomes and utility information
  return action
```

Decision theory: An agent is rational exactly when it chooses the **action** with the **maximum expected utility** taken over all results of actions.

7/10

Unconditional Probabilities (2)

In general, a **random variable** can take on *true* and *false* values, as well as other values:

$$\begin{aligned} P(\text{Weather}=\text{Sunny}) &= 0.7 \\ P(\text{Weather}=\text{Rain}) &= 0.2 \\ P(\text{Weather}=\text{Cloudy}) &= 0.08 \\ P(\text{Weather}=\text{Snow}) &= 0.02 \\ P(\text{Headache}=\text{TRUE}) &= 0.1 \end{aligned}$$

- Propositions can contain equations over random variables.
- Logical connectors can be used to build propositions, e.g. $P(\text{Cavity} \wedge \neg \text{Insured}) = 0.06$.

7/12

Unconditional Probabilities (3)

$\mathbf{P}(x)$ is the **vector of probabilities** for the (ordered) domain of the random variable X :

$$\mathbf{P}(\text{Headache}) = \langle 0.1, 0.9 \rangle$$

$$\mathbf{P}(\text{Weather}) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$$

define the probability distribution for the random variables *Headache* and *Weather*.

$\mathbf{P}(\text{Headache}, \text{Weather})$ is a 4x2 table of probabilities of all combinations of the values of a set of random variables.

	Headache = TRUE	Headache = FALSE
Weather = Sunny	$P(W = \text{Sunny} \wedge \text{Headache})$	$P(W = \text{Sunny} \wedge \neg \text{Headache})$
Weather = Rain		
Weather = Cloudy		
Weather = Snow		

7/13

Conditional Probabilities (2)

$\mathbf{P}(\text{Weather} | \text{Headache})$ is a 4x2 table of conditional probabilities of all combinations of the values of a set of random variables.

	Headache = TRUE	Headache = FALSE
Weather = Sunny	$P(W = \text{Sunny} \text{Headache})$	$P(W = \text{Sunny} \neg \text{Headache})$
Weather = Rain		
Weather = Cloudy		
Weather = Snow		

Conditional probabilities result from unconditional probabilities (if $P(B) > 0$) (**per definition**):

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

13/15

Conditional Probabilities (1)

New information can change the probability.

Example: The probability of a cavity increases if we know the patient has a toothache.

If additional information is available, we can no longer use the **prior** probabilities!

$P(A|B)$ is the **conditional** or **posterior** probability of A given that *all we know* is B :

$$P(\text{Cavity} | \text{Toothache}) = 0.8$$

$\mathbf{P}(X|Y)$ is the table of all conditional probabilities over all values of X and Y .

7/14

Conditional Probabilities (3)

$\mathbf{P}(X,Y) = \mathbf{P}(X|Y) \mathbf{P}(Y)$ corresponds to a system of equations:

$$\begin{aligned}
 P(W = \text{Sunny} \wedge \text{Headache}) &= P(W = \text{Sunny} | \text{Headache}) P(\text{Headache}) \\
 P(W = \text{Rain} \wedge \text{Headache}) &= P(W = \text{Rain} | \text{Headache}) P(\text{Headache}) \\
 &= \dots \\
 P(W = \text{Snow} \wedge \neg \text{Headache}) &= P(W = \text{Snow} | \neg \text{Headache}) P(\neg \text{Headache})
 \end{aligned}$$

7/16

Conditional Probabilities (4)

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

- Product rule: $P(A \wedge B) = P(A | B)P(B)$
- Similarly: $P(A \wedge B) = P(B | A)P(A)$
- A and B are **independent** if $P(A | B) = P(A)$
(equiv. $P(B | A) = P(B)$)
Then (and only then) it holds that
 $P(A \wedge B) = P(A)P(B)$

7/17

Why are the Axioms Reasonable?

- If P represents an *objectively* observable probability, the axioms clearly make sense.
 - But why should an agent respect these axioms when it models its own degree of belief?
- *Objective* vs. *subjective* probabilities

The axioms limit the set of beliefs that an agent can maintain.

One of the most convincing arguments for why subjective beliefs should respect the axioms was put forward by de Finetti in 1931. It is based on the connection between actions and degree of belief.

- If the beliefs are contradictory, then the agent will fail in its environment in the long run!

7/19

Axiomatic Probability Theory

A function P of formulae from propositional logic in the set $[0,1]$ is a probability measure if for all propositions A, B :

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

All other properties can be derived from these axioms, for example:

$$P(\neg A) = 1 - P(A)$$

follows from $P(A \vee \neg A) = 1$ and $P(A \wedge \neg A) = 0$.

7/18

Joint Probability

The agent assigns probabilities to every proposition in the domain.

An **atomic event** is an assignment of values to all random variables X_1, \dots, X_n (= complete specification of a state).

Example: Let X and Y be bool. $X \wedge Y, X \wedge \neg Y, \neg X \wedge Y, \neg X \wedge \neg Y$ the following 4 atomic events:

The **joint probability distribution** $P(X_1, \dots, X_n)$ assigns a probability to every atomic event.

	Toothache	\neg Toothache
Cavity	0.04	0.06
\neg Cavity	0.01	0.89

Since all atomic events are disjoint, the sum of all fields is 1 (disjunction of events). The conjunction is necessarily false.

7/20

Working with Joint Probability

All relevant probabilities can be computed using the joint probability by expressing them as a disjunction of atomic events.

Examples:

$$\begin{aligned} P(\text{Cavity} \vee \text{Toothache}) &= P(\text{Cavity} \wedge \text{Toothache}) \\ &+ P(\neg \text{Cavity} \wedge \text{Toothache}) \\ &+ P(\text{Cavity} \wedge \neg \text{Toothache}) \end{aligned}$$

We obtain unconditional probabilities by adding across a row or column:

$$P(\text{Cavity}) = P(\text{Cavity} \wedge \text{Toothache}) + P(\text{Cavity} \wedge \neg \text{Toothache})$$

$$P(\text{Cavity} | \text{Toothache}) = \frac{P(\text{Cavity} \wedge \text{Toothache})}{P(\text{Toothache})} = \frac{0.04}{0.04+0.01} = 0.80$$

7/21

Bayes' Rule

We know (product rule):

$$P(A \wedge B) = P(A|B) P(B) \quad \text{and} \quad P(A \wedge B) = P(B|A) P(A)$$

By equating the right-hand sides, we get

$$\begin{aligned} P(A|B) P(B) &= P(B|A) P(A) \\ \Rightarrow P(A|B) &= \frac{P(B|A) P(A)}{P(B)} \end{aligned}$$

For multi-valued variables (set of equalities):

$$P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}$$

Generalization (conditioning on background evidence E):

$$P(Y|X,E) = \frac{P(X|Y,E) P(Y|E)}{P(X|E)}$$

7/23

Problems with Joint Probabilities

We can easily obtain all probabilities from the joint probability.

The joint probability, however, involves k^n values, if there are n random variables with k values.

- Difficult to represent
- Difficult to assess

Questions:

- Is there a more compact way of representing joint probabilities?
- Is there an efficient method to work with this representation?

Not in general, but it can work in many cases. Modern systems work directly with conditional probabilities and make assumptions on the independence of variables in order to simplify calculations.

7/22

Applying Bayes' Rule

$$P(\text{Toothache} | \text{Cavity}) = 0.4$$

$$P(\text{Cavity}) = 0.1$$

$$P(\text{Toothache}) = 0.05$$

$$P(\text{Cavity} | \text{Toothache}) = \frac{0.4 \times 0.1}{0.05} = 0.8$$

Why don't we try to assess $P(\text{Cavity} | \text{Toothache})$ directly?

$P(\text{Toothache} | \text{Cavity})$ (**causal**) is more robust than

$P(\text{Cavity} | \text{Toothache})$ (**diagnostic**):

- $P(\text{Toothache} | \text{Cavity})$ is independent from the prior probabilities $P(\text{Toothache})$ and $P(\text{Cavity})$.
- If there is a cavity epidemic and $P(\text{Cavity})$ increases, $P(\text{Toothache} | \text{Cavity})$ does not change, but $P(\text{Toothache})$ and $P(\text{Cavity} | \text{Toothache})$ will change proportionally.

7/24

Relative Probability

Assumption: We would also like to consider the probability that the patient has gum disease.

$$\begin{aligned} P(\text{Toothache} \mid \text{Gum Disease}) &= 0.7 \\ P(\text{Gum Disease}) &= 0.02 \end{aligned}$$

Which diagnosis is more probable?

$$P(C \mid T) = \frac{P(T \mid C) P(C)}{P(T)} \quad \text{or} \quad P(G \mid T) = \frac{P(T \mid G) P(G)}{P(T)}$$

If we are only interested in the **relative probability**, we need not assess $P(T)$:

$$\begin{aligned} \frac{P(C \mid T)}{P(G \mid T)} &= \frac{P(T \mid C) P(C)}{P(T)} \times \frac{P(T)}{P(T \mid G) P(G)} = \frac{P(T \mid C) P(C)}{P(T \mid G) P(G)} \\ &= \frac{0.4 \times 0.1}{0.7 \times 0.02} = 2.857 \end{aligned}$$

→ Important for excluding possible diagnoses.

7/25

Normalization (2)

By substituting into the first equation:

$$P(C \mid T) = \frac{P(T \mid C) P(C)}{P(T \mid C) P(C) + P(T \mid -C) P(-C)}$$

For random variables with multiple values:

$$\mathbf{P}(Y \mid X) = \alpha \mathbf{P}(X \mid Y) \mathbf{P}(Y)$$

where α is the normalization constant needed to make the entries in $\mathbf{P}(Y \mid X)$ sum to 1.

Example: $\alpha (.1, .1, .3) = (.2, .2, .6)$.

7/27

Normalization (1)

If we wish to determine the absolute probability of $P(C \mid T)$ and we do not know $P(T)$, we can also carry out a **complete case analysis** (e.g. for C and $-C$) and use the fact that $P(C \mid T) + P(-C \mid T) = 1$ (here boolean variables):

$$\begin{aligned} P(C \mid T) &= \frac{P(T \mid C) P(C)}{P(T)} \\ P(-C \mid T) &= \frac{P(T \mid -C) P(-C)}{P(T)} \\ P(C \mid T) + P(-C \mid T) &= \frac{P(T \mid C) P(C)}{P(T)} + \frac{P(T \mid -C) P(-C)}{P(T)} \\ P(T) &= P(T \mid C) P(C) + P(T \mid -C) P(-C) \end{aligned}$$

7/26

Example

Your doctor tells you that you have tested positive for a serious but rare (1/10000) disease. This test (T) is correct to 99% (1% false positive & 1% false negative results).

What does this mean for you?

$$\begin{aligned} P(D \mid T) &= \frac{P(T \mid D) P(D)}{P(T)} = \frac{P(T \mid D) P(D)}{P(T \mid D) P(D) + P(T \mid -D) P(-D)} \\ P(D) &= 0.0001 \quad P(T \mid D) = 0.99 \quad P(T \mid -D) = 0.01 \\ P(D \mid T) &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} = \frac{0.000099}{0.000099 + 0.009999} \\ &= \frac{0.000099}{0.010098} \approx 0.01 \end{aligned}$$

Moral: If the test imprecision is much greater than the rate of occurrence of the disease, then a positive result is not as threatening as you might think.

7/28

Multiple Evidence (1)

A dentist's probe catches in the aching tooth of a patient. Using Bayes' rule, we can calculate:

$$P(\text{Cavity} \mid \text{Catch}) = 0.95$$

But how does the combined evidence help? Using Bayes' rule, the dentist could establish:

$$P(\text{Cav} \mid \text{Tooth} \wedge \text{Catch}) = \frac{P(\text{Tooth} \wedge \text{Catch} \mid \text{Cav}) \times P(\text{Cav})}{P(\text{Tooth} \wedge \text{Catch})}$$

$$P(\text{Cav} \mid \text{Tooth} \wedge \text{Catch}) = \alpha \cdot P(\text{Tooth} \wedge \text{Catch} \mid \text{Cav}) \times P(\text{Cav})$$

7/29

Conditional Independence

The general definition of conditional independence of two variables X and Y given a third variable Z is:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

Thus our diagnostic problem turns into:

$$P(\text{Cav} \mid \text{Tooth} \wedge \text{Catch}) = \alpha \cdot P(\text{Tooth} \mid \text{Cav}) P(\text{Catch} \mid \text{Cav}) P(\text{Cav})$$

7/31

Multiple Evidence (2)

Problem: The dentist needs $P(\text{Tooth} \wedge \text{Catch} \mid \text{Cav})$, i.e., diagnostic [knowledge of all combinations of symptoms](#) in the general case.

It would be [nice](#) if [Tooth](#) and [Catch](#) were independent but [they are not](#): if a probe catches in the tooth, it probably has cavity which probably causes toothache.

[They are independent given we know whether the tooth has cavity](#):

$$P(\text{Tooth} \wedge \text{Catch} \mid \text{Cav}) = P(\text{Tooth} \mid \text{Cav}) P(\text{Catch} \mid \text{Cav})$$

Each is directly caused by the cavity but neither has a direct effect on the other.

7/30

Recursive Bayesian Updating

Multiple evidence can be reduced to prior probabilities and conditional probabilities (assuming conditional independence).

The general combination rule, if Z_1 and Z_2 are independent given X is

$$P(X \mid Z_1, Z_2) = \alpha \cdot P(X) P(Z_1 \mid X) P(Z_2 \mid X)$$

where α is the normalization constant.

Generalization: [Recursive Bayesian Updating](#)

$$P(X \mid Z_1, \dots, Z_n) = \alpha \cdot P(X) \prod_{i=1..n} P(Z_i \mid X)$$

7/32

Types of Variables

- Variables can be **discrete** or **continuous**:
- **Discrete variables**
 - Weather: sunny, rain, cloudy, snow
 - Cavity: true, false (boolean)
- **Continuous variables**
 - Tomorrow's maximum temperature in Berkeley
 - Domain can be the entire real line or any subset.
 - Distributions for continuous variables are typically given by **probability density functions**.

7/33

Summary

- **Uncertainty** is unavoidable in complex, dynamic worlds in which agents are ignorant.
- **Probabilities** express the agent's inability to reach a definite decision. They summarize the agent's beliefs.
- **Conditional** and **unconditional** probabilities can be formulated over propositions.
- If an agent disrespects the theoretical probability **axioms**, it is likely to demonstrate irrational behaviour.
- **Bayes' rule** allows us to calculate known probabilities from unknown probabilities.
- **Multiple evidence** (assuming independence) can be effectively incorporated using **recursive Bayesian updating**.

7/35

Marginalization and Normalization

For any sets of variables Y and Z we have

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y} | \mathbf{z}) \mathbf{P}(\mathbf{z})$$

Let X be a random variable and **e** be the observed value of a variable E.

$$\mathbf{P}(X | \mathbf{e}) = \mathbf{P}(X, \mathbf{e}) / \mathbf{P}(\mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{z}} \mathbf{P}(X, \mathbf{e}, \mathbf{z})$$

Since **e** is known, the factor $1 / \mathbf{P}(\mathbf{e})$ is the same for all values of X.

7/34

Bayesian Networks

Example domain: I am at work. My neighbour John calls me to tell me, that my alarm is ringing. My neighbour Mary doesn't call. Sometimes, the alarm is started by a slight earthquake.

Question: Is there a burglary?

Variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls.

7/36

Bayesian Networks

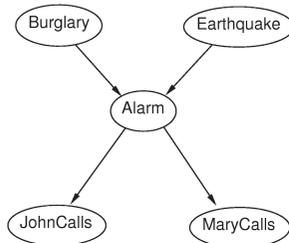
Domain knowledge/ assumptions:

- Events Burglary and Earthquake are independent. (of course, to be discussed: a burglary doesn't cause an earthquake, but a burglar might use an earthquake to do the burglary. Then the independence assumption is not true. Design decision!)
- Alarm might be activated by burglary or earthquake
- John calls if and only if he heard the alarm. His call probability is not influenced by the fact, that there is an earthquake at the same time. Same for Mary.

How to model this domain efficiently? Goal: Answer questions.

7/37

The Meaning of Bayesian Networks



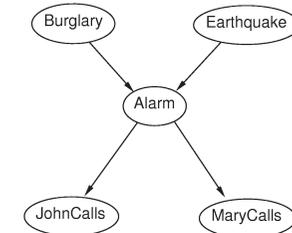
- *Alarm* depends on *Burglary* and *Earthquake*.
 - *MaryCalls* only depends on *Alarm*.
 $P(\text{MaryCalls} \mid \text{Alarm}, \text{Burglary}) = P(\text{MaryCalls} \mid \text{Alarm})$
- Bayesian Networks can be considered as sets of independence assumptions.

7/39

Bayesian Networks

(also belief networks, probabilistic networks, causal networks)

- The *random variables* are the **nodes**.
- **Directed edges** between nodes represent *direct influence*.
- A **table of conditional probabilities** (CPT) is associated with every node, in which the effect of the **parent** nodes is quantified.
- The graph is **acyclic** (a DAG).



Remark: Burglary and Earthquake are denoted as the **parents** of Alarm

7/38

Bayesian Networks and the Joint Probability

Bayesian networks can be seen as a more compact representation of joint probabilities.

Let all nodes X_1, \dots, X_n be ordered topologically according to the the arrows in the network. Let x_1, \dots, x_n be the values of the variables. Then

$$\begin{aligned} P(x_1, \dots, x_n) &= P(x_n \mid x_{n-1}, \dots, x_1) \cdot \dots \cdot P(x_2 \mid x_1) P(x_1) \\ &= \prod_{i=1}^n P(x_i \mid x_{i-1}, \dots, x_1) \end{aligned}$$

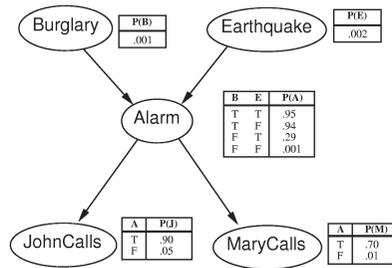
From the independence assumption, this is equivalent to

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{parents}(x_i))$$

We can **calculate the joint probability from the network topology and the CPTs!**

7/40

Example



Only the probabilities for positive events are given. The negative probabilities can be found using $P(\neg X) = 1 - P(X)$.

$$\begin{aligned}
 P(J, M, A, \neg B, \neg E) &= P(J|A) P(M|A) P(A|\neg B, \neg E) P(\neg B) P(\neg E) \\
 &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\
 &= 0.00062
 \end{aligned}$$

7/41

7/42

Compactness of Bayesian Networks

- For the explicit representation of Bayesian networks, we need a table of size 2^n where n is the number of variables.
- In the case that every node in a network has at most k parents, we only need n tables of size 2^k (assuming boolean variables).
- Example:* $n = 20$ and $k = 5$
 - $2^{20} = 1\,048\,576$ and $20 \times 2^5 = 640$ different explicitly-represented probabilities!
 - In the worst case, a Bayesian network can become exponentially large, for example if every variable is directly influenced by all the others.
 - The size depends on the application domain (local vs. global interaction) and the skill of the designer.

Naive Design of a Network

- Order all variables
- Take the first from those that remain
- Assign all direct influences from nodes already in the network to the new node (Edges + CPT).
- If there are still variables in the list, repeat from step 2.

7/43

Example 1

M, J, A, B, E

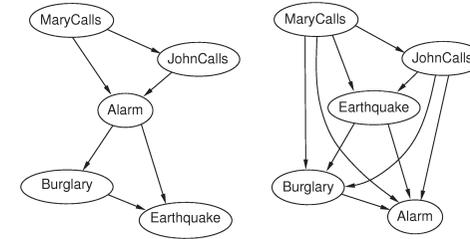
7/44

Example 2

M, J, E, B, A

Example

left = M, J, A, B, E, right = M, J, E, B, A



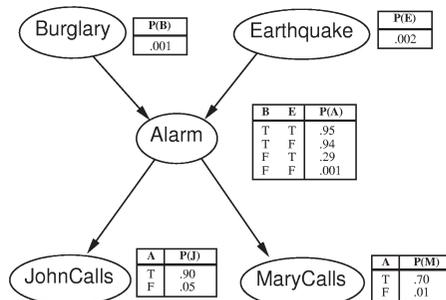
→ Attempt to build a diagnostic model of symptoms and causes, which always leads to dependencies between causes that are actually independent and symptoms that appear separately.

7/45

7/46

Inference in Bayesian Networks

Instantiating evidence variables and sending queries to nodes.



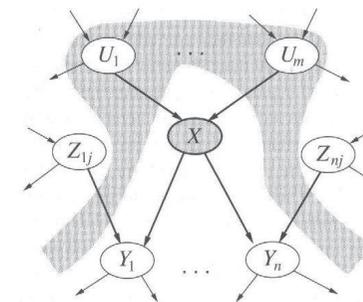
What is
or

$P(\text{Burglary} \mid \text{JohnCalls})$
 $P(\text{Burglary} \mid \text{JohnCalls}, \text{MaryCalls})?$

7/47

Conditional Independence Relations in Bayesian Networks (1)

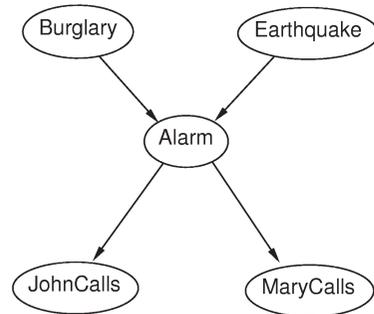
A node is conditionally independent of its non-descendants given its parents.



7/48

Example

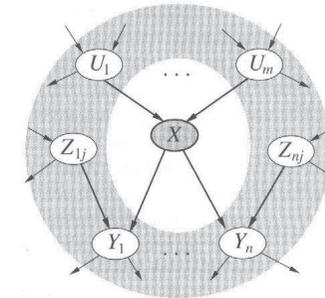
JohnCalls is independent of Burglary and Earthquake given the value of Alarm.



7/49

Conditional Independence Relations in Bayesian Networks (2)

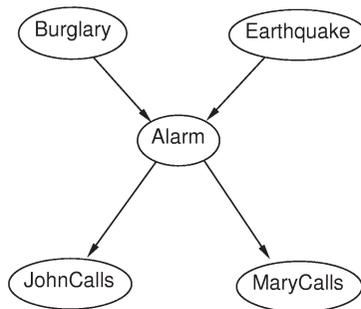
A node is conditionally independent of all other nodes in the network given the **Markov blanket**, i.e., its parents, children and children's parents.



7/50

Example

Burglary is independent of JohnCalls and MaryCalls, given the values of Alarm and Earthquake.



7/51

Exact Inference in Bayesian Networks

- Compute the **posterior probability** distribution for a **set of query variables X** given an observation, i.e., the values of a **set of evidence variables E**.
- **Complete set of variables is $X \cup E \cup Y$**
- **Y** are called the **hidden variables**
- Typical query $P(X | e)$ where e are the observed values of E .
- In the remainder: X is a singleton

Example:

$$P(\text{Burglary} \mid \text{JohnCalls} = \text{true}, \text{MaryCalls} = \text{true}) = (0.284, 0.716)$$

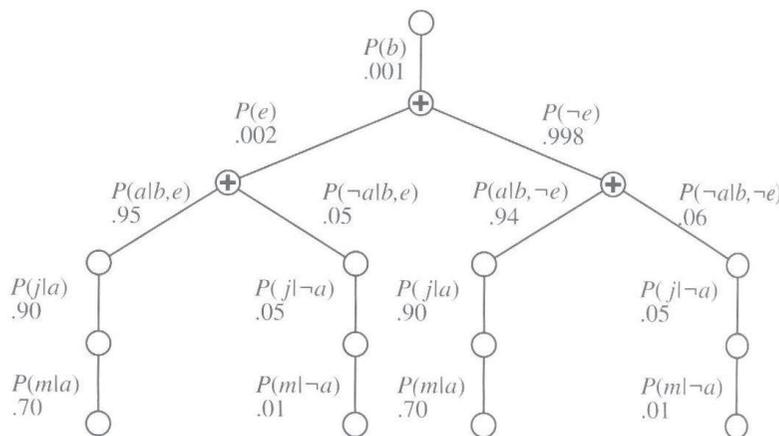
7/52

Inference by Enumeration

- $P(X|e) = \alpha P(X,e) = \sum_y \alpha P(X,e,y)$
- The network gives a complete representation of the full joint distribution.
- A query can be answered using a Bayesian network by computing sums of products of conditional probabilities from the network.
- We sum over the hidden variables.

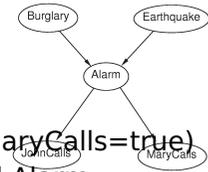
7/53

Evaluation of $P(b|j,m)$



13/55

Example



- Consider $\mathbf{P}(\text{Burglary} \mid \text{JohnCalls} = \text{true}, \text{MaryCalls} = \text{true})$
- The hidden variables are Earthquake and Alarm.
- We have: $\mathbf{P}(B \mid j, m) = \alpha \mathbf{P}(B, j, m)$
 $= \alpha \sum_e \sum_a \mathbf{P}(B, j, m, e, a)$
- If we consider the independence of variables, we obtain for $B=b$
- $P(b \mid j, m) = \alpha \sum_e \sum_a P(j|a) P(m|a) P(a|e,b) P(e) P(b)$
- Reorganization of the terms yields
- $P(b \mid j, m) = \alpha P(b) \sum_e P(e) \sum_a P(a|e,b) P(j|a) P(m|a)$
- As a result we obtain:
 $\mathbf{P}(b \mid j, m) = \alpha (0.00059224, 0.0014919)$
 $\approx (0.284, 0.716)$

7/54

Enumeration Algorithm for Answering Queries on Bayesian Networks

```

function ENUMERATION-ASK( $X, e, bn$ ) returns a distribution over  $X$ 
  inputs:  $X$ , the query variable
          $e$ , observed values for variables  $E$ 
          $bn$ , a Bayes net with variables  $\{X\} \cup E \cup Y$  /*  $Y = \text{hidden variables}$  */

   $Q(X) \leftarrow$  a distribution over  $X$ , initially empty
  for each value  $x_i$  of  $X$  do
    extend  $e$  with value  $x_i$  for  $X$ 
     $Q(x_i) \leftarrow$  ENUMERATE-ALL(VARS[ $bn$ ],  $e$ )
  return NORMALIZE( $Q(X)$ )

function ENUMERATE-ALL( $vars, e$ ) returns a real number
  if EMPTY?( $vars$ ) then return 1.0
   $Y \leftarrow$  FIRST( $vars$ )
  if  $Y$  has value  $y$  in  $e$ 
    then return  $P(y \mid \text{parents}(Y)) \times$  ENUMERATE-ALL(REST( $vars$ ),  $e$ )
  else return  $\sum_y P(y \mid \text{parents}(Y)) \times$  ENUMERATE-ALL(REST( $vars$ ),  $e_y$ )
    where  $e_y$  is  $e$  extended with  $Y = y$ 
  
```

7/56

Properties of the Enumeration-Ask Algorithm

- The Enumeration-Ask algorithm evaluates the trees in a depth-first manner.
- Space complexity is linear in the number of variables.
- Time complexity for a network with n boolean variables is $O(2^n)$,
- since sub-expressions are repeatedly evaluated.

7/57

Example

- Let us consider the query $P(\text{JonCalls} | \text{Burglary} = \text{true})$.
- The nested sum is
$$P(j,b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(j,a) \sum_m P(m|a)$$
- Obviously, the rightmost sum equals 1 so that it can safely be dropped.
- Variable elimination repeatedly removes leaf nodes that are not query or evidence variables or non-ancestor nodes of query or evidence variables and this way speeds up computation.

7/59

Variable Elimination

- The enumeration algorithm can be improved significantly by eliminating repeating or unnecessary calculations.
- The key idea is to evaluate expressions from right to left and to save results for later use.
- Additionally, unnecessary expressions can be removed.

7/58

Complexity of Exact Inference

- If the network is singly connected or a polytree, the time and space complexity of exact inference is linear in the size of the network.
- The burglary example is a typical singly connected network.
- For multiply connected networks inference in Bayesian Networks is NP-hard.
- There are approximate inference methods for multiply connected networks such as sampling techniques or Markov chain Monte Carlo.

7/60

Other Approaches (1)

- Rule-based methods with “certainty factors”.
 - Logic-based systems with weights attached to rules, which are combined using inference.
 - Had to be designed carefully to avoid undesirable interactions between different rules.
 - Might deliver **incorrect** results through overcounting of evidence.
 - Their use is no longer recommended.

7/61

Other Approaches (2)

- Dempster-Shafer Theory
 - Allows the representation of *ignorance* as well as uncertainty.
 - Example: If a coin is fair, we assume $P(\text{Heads}) = 0.5$. But what if we don't know if the coin is fair? $\rightarrow Bel(\text{Heads})=0, Bel(\text{Tails})=0$. If the coin is 90% fair, 0.5×0.9 , i.e. $Bel(\text{Heads}) = 0.45$.
- Interval of probabilities is $[0.45, 0.55]$ with the evidence, $[0,1]$ without.
- The notion of utility is not yet well understood in Dempster-Shafer Theory.

7/62

Other Approaches (3)

- Fuzzy logic and fuzzy sets
 - A means of representing and working with **vagueness**, not uncertainty.
 - Example: The car is *fast*.
 - Used especially in control and regulation systems.
 - In such systems, it can be interpreted as an *interpolation technique*.

7/63

Summary

- Bayesian Networks allow a **compact representation** of joint probability distribution.
- Bayesian Networks provide a concise way to represent **conditional independence** in a domain.
- Inference in Bayesian networks means **computing the probability distribution of a set of query variables, given a set of evidence variables**.
- **Exact inference algorithms such as variable elimination are efficient for poly-trees**.
- In **complexity of belief network inference depends on the network structure**.
- In general, **Bayesian network inference is NP-hard**.

7/64