Foundations of Artificial Intelligence
12. Making Simple Decisions under Uncertainty
Probability Theory, Bayesian Networks, Other Approaches

Joschka Boedecker and Wolfram Burgard and Bernhard Nebel

Albert-Ludwigs-Universität Freiburg

Contents

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

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
- $\rightarrow$ They imply different costs and different probabilities of actually achieving the goal.
- $\rightarrow$ 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.

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

→ This rule is *incorrect*! Better:

$$\forall p[Symptom(p, toothache) \Rightarrow$$
$$Disease(p, cavity) \lor Disease(p, gum\_disease) \lor \ldots]$$

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

Perhaps a *causal* rule is better?

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

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

Uncertainty in Logical 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!

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.

$\rightarrow$ From the fact that a landmark was perceived, we cannot conclude with certainty that the robot is at that location.

$\rightarrow$ The same is true when no landmark is perceived.

$\rightarrow$ Only the probability increases or decreases.

Degree of Belief and Probability Theory

- 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.

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

Decision-Theoretic Agent

---

function DT-AGENT(*percept*) **returns** an *action*
   persistent: *belief_state*, probabilistic beliefs about the current state of the world
               action, the agent's action

   update *belief_state* based on *action* and *percept*
   calculate outcome probabilities for actions,
      given action descriptions and current *belief_state*
   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.

Axiomatic Probability Theory

A function P of formulae from propositional logic in the set $[0,1]$ is a probability measure if for all propositions ϕ, ψ (whereby propositions are the equivalance classes fromed by logically equivalent formulae):

1. $0 \leq P(\phi) \leq 1$
2. $P(true) = 1$
3. $P(false) = 0$
4. $P(\phi \vee \psi) = P(\phi) + P(\psi) - P(\phi \wedge \psi)$

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

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

since $1 \overset{(2)}{=} P(\phi \vee \neg\phi) \overset{(4)}{=} P(\phi) + P(\neg\phi) - P(\phi \wedge \neg\phi) \overset{(3)}{=} P(\phi) + P(\neg\phi)$.

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?

$\rightarrow$ *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.

$\rightarrow$ If the beliefs are contradictory, then there exists a betting strategy (the dutch book) against the agent, where he will definitely loose!

Notation

We use random variable such as $Weather$ (capitalized word), which has a domain of ordered values. In our case that could be $sunny$, $rain$, $cloudy$, $snow$ (lower case words).

A proposition might then be: $Weather = cloudy$.

If the random variable is Boolean, e.g. $Headache$, we may write either $Headache = true$ or equivalently $headhache$ (lowercase!). Similarly, we may write $Headache = false$ or equivalently $\neg headache$.

Further, we can of course use Boolean connectors, e.g. $\neg headhache \wedge Weather = cloudy$.

Unconditional Probabilities (1)

$P(a)$ denotes the unconditional probability that it will turn out that $A = true$ *in the absence of any other information*, for example:

$$P(cavity) = 0.1$$

In case of non-Boolean random variables:

$$P(Weather = sunny) = 0.7$$
$$P(Weather = rain) = 0.2$$
$$P(Weather = cloudy) = 0.08$$
$$P(Weather = snow) = 0.02$$

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

$$\mathbf{P}(Headache) = \langle 0.1, 0.9 \rangle$$
$$\mathbf{P}(Weather) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$$

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

$\mathbf{P}(Headache, Weather)$ is a 4×2 table of probabilities of all combinations of the values of a set of random variables.

	$Headache = true$	$Headache = false$
$Weather = sunny$	$P(W = sunny \wedge headache)$	$P(W = sunny \wedge \neg headache)$
$Weather = rain$		
$Weather = cloudy$		
$Weather = snow$		

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!

$P(a \mid b)$ is the conditional or posterior probability of a given that *all we know* is b:

$$P(cavity \mid toothache) = 0.8$$

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

Conditional Probabilities (2)

$\mathbf{P}(Weather \mid Headache)$ is a 4×2 table of conditional probabilities of all combinations of the values of a set of random variables.

	$Headache = true$	$Headache = false$
$Weather = sunny$	$P(W = sunny \mid headache)$	$P(W = sunny \mid \neg headache)$
$Weather = rain$		
$Weather = cloudy$		
$Weather = snow$		

Conditional probabilities result from unconditional probabilities (if $P(b) > 0$) (by definition):

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

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

$$
\begin{aligned}
P(W = sunny \wedge headache) &= P(W = sunny \mid headache)P(headache) \\
P(W = rain \wedge headache) &= P(W = rain \mid headache)P(headache) \\
\dots &= \dots \\
P(W = snow \wedge \neg headache) &= P(W = snow \mid \neg headache)P(\neg headache)
\end{aligned}
$$

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

- Product rule: $P(a \wedge b) = P(a \mid b)P(b)$

- Similarly: $P(a \wedge b) = P(b \mid a)P(a)$

- a and b are independent if $P(a \mid b) = P(a)$
  (equiv. $P(b \mid a) = P(b)$).
  Then (and only then) it holds that $P(a \wedge b) = P(a)P(b)$.

The agent assigns probabilities to every proposition in the domain.

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

Example: Let X and Y be Boolean variables. Then we have the following 4 atomic events: $x \wedge y$, $x \wedge \neg y$, $\neg x \wedge y$, $\neg x \wedge \neg y$.

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

	toothache	*¬toothache*
cavity	0.04	0.06
¬cavity	0.01	0.89

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

Working with Joint Probability

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

Examples:

$$P(cavity \lor toothache) = P(cavity \land toothache)$$
$$+ P(\neg cavity \land toothache)$$
$$+ P(cavity \land \neg toothache)$$

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

$$P(cavity) = P(cavity \land toothache) + P(cavity \land \neg toothache)$$

$$P(cavity \mid toothache) = \frac{P(cavity \land toothache)}{P(toothache)} = \frac{0.04}{0.04 + 0.01} = 0.80$$

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.

$\rightarrow$ Difficult to represent

$\rightarrow$ Difficult to assess

Questions:

$\rightarrow$ Is there a more compact way of representing joint probabilities?

$\rightarrow$ 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.

Using the product rule $P(a \wedge b) = P(a \mid b) \, P(b)$, joint probabilites can be expressed as products of conditional probabilities.

$P(x_1, \ldots, x_n) = P(x_n, \ldots, x_1) = P(x_n \mid x_{n-1} \ldots, x_1) \, P(x_{n-1}, \ldots, x_1)$
$= P(x_n \mid x_{n-1} \ldots, x_1) \, P(x_{n-1} \mid x_{n-2} \ldots, x_1) \, P(x_{n-2}, \ldots, x_1)$
$= P(x_n \mid x_{n-1} \ldots, x_1) \, P(x_{n-1} \mid x_{n-2} \ldots, x_1) \, P(x_{n-2} \mid x_{n-3} \ldots, x_1)$
$P(x_{n-3}, \ldots, X_1)$
$= \ldots$
$= P(x_n \mid x_{n-1} \ldots, x_1) \, P(x_{n-1} \mid x_{n-2} \ldots, x_1) \ldots P(x_2 \mid x_1) \, P(x_1)$
$= \Pi_{i=1}^{n} P(x_i \mid x_{i-1} \ldots x_1)$

Bayes' Rule

We know (product rule):

$$P(a \wedge b) = P(a \mid b)P(b) \text{ and } P(a \wedge b) = P(b \mid a)P(a)$$

By equating the right-hand sides, we get

$$P(a \mid b)P(b) = P(b \mid a)P(a)$$

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

For multi-valued variables we get a set of equalities:

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

Generalization (conditioning on background evidence e):

$$\mathbf{P}(Y \mid X, e) = \frac{\mathbf{P}(X \mid Y, e)\mathbf{P}(Y \mid e)}{\mathbf{P}(X \mid e)}$$

Applying Bayes' Rule

$$P(toothache \mid cavity) = 0.4$$

$$P(cavity) = 0.1$$

$$P(toothache) = 0.05$$

$$P(cavity \mid toothache) = \frac{0.4 \times 0.1}{0.05} = 0.8$$

Why don't we try to assess $P(cavity \mid toothache)$ directly?

$P(toothache \mid cavity)$ (causal) is more robust than $P(cavity \mid toothache)$ (diagnostic):

- $P(toothache \mid cavity)$ is independent from the prior probabilities $P(toothache)$ and $P(cavity)$.
- If there is a cavity epidemic and $P(cavity)$ increases, $P(toothache \mid cavity)$ does not change, but $P(toothache)$ and $P(cavity \mid toothache)$ will change proportionally.

Relative Probability

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

$$P(toothache \mid gumdisease) = 0.7$$
$$P(gumdisease) = 0.02$$

Which diagnosis is more probable?

$$P(c \mid t) = \frac{P(t \mid c)P(c)}{P(t)} \text{ or } P(g \mid t) = \frac{P(t \mid g)P(g)}{P(t)}$$

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

$$\frac{P(c \mid t)}{P(g \mid t)} = \frac{P(t \mid c)P(c)}{P(t)} \times \frac{P(t)}{P(t \mid g)P(g)} = \frac{P(t \mid c)P(c)}{P(t \mid g)P(g)}$$
$$= \frac{0.4 \times 0.1}{0.7 \times 0.02} = 2.857$$

$\rightarrow$ Important for excluding possible diagnoses.

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 $\neg c$) and use the fact that $P(c \mid t) + P(\neg c \mid t) = 1$ (here Boolean variables):

$$P(c \mid t) = \frac{P(t \mid c)P(c)}{P(t)}$$

$$P(\neg c \mid t) = \frac{P(t \mid \neg c)P(\neg c)}{P(t)}$$

$$P(c \mid t) + P(\neg c \mid t) = \frac{P(t \mid c)P(c)}{P(t)} + \frac{P(t \mid \neg c)P(\neg c)}{P(t)}$$

$$P(t) = P(t \mid c)P(c) + P(t \mid \neg c)P(\neg c)$$

Normalization (2)

By substituting into the first equation:

$$P(c \mid t) = \frac{P(t \mid c)P(c)}{P(t \mid c)P(c) + P(t \mid \neg c)P(\neg c)}$$

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 for each value of X.

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

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?

$$P(d \mid t) = \frac{P(t \mid d)P(d)}{P(t)} = \frac{P(t \mid d)P(d)}{P(t \mid d)P(d) + P(t \mid \neg d)P(\neg d)}$$

$$P(d) = 0.0001 \quad P(t \mid d) = 0.99 \quad P(t \mid \neg d) = 0.01$$

$$P(d \mid t) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} = \frac{0.000099}{0.000099 + 0.009999}$$
$$= \frac{0.000099}{0.010088} \approx 0.01$$

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.

Multiple Evidence (1)

A dentist's probe catches ($Catch = true$) in the aching tooth ($Toothache = true$) of a patient. We already know that $P(cavity \mid toothache) = 0.8$. Furthermore, using Bayes' rule, we can calculate:

$$P(cavity \mid catch) = 0.95$$

But how does the combined evidence ($tooth \wedge catch$) help?

Using Bayes' rule, the dentist could establish:

$$P(cav \mid tooth \wedge catch) = \frac{P(tooth \wedge catch \mid cav) \times P(cav)}{P(tooth \wedge catch)}$$

$$P(cav \mid tooth \wedge catch) = \alpha P(tooth \wedge catch \mid cav) \times P(cav)$$

Multiple Evidence (2)

Problem: The dentist needs $P(tooth \land catch \mid cav)$, 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: $P(tooth \mid catch) \neq P(tooth)$ - 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:

$$P(tooth \mid catch, cav) = P(tooth \mid catch)$$

If one already knows, that there is a cavity, then the additional knowledge of the probe catches does not change the probabilty.

$$P(tooth \land catch \mid cav) =$$

$$P(tooth \mid catch, cav)P(catch \mid cav) = P(tooth \mid cav)P(catch \mid cav)$$

Conditional Independence

Thus our diagnostic problem turns into:

$$P(cav \mid tooth \wedge catch) = \alpha P(tooth \wedge catch \mid cav)P(cav)$$

$$= \alpha P(tooth \mid catch, cav)P(catch \mid cav)P(cav)$$

$$= \alpha P(tooth \mid cav)P(catch \mid cav)P(cav)$$

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

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

Eating icecream and observing sunshine is not independent

$$P(ice \mid sun) \neq P(ice)$$

The variables Ice and Sun are not independent.
But if the reason for eating icecream is simply that it is hot outside, then the additional observation of sunshine does not make a difference:

$$P(ice \mid sun, hot) = P(ice \mid hot)$$

The variables Ice and Sun are conditionally independent given that $Hot = true$ is observed.
The knowledge about indpendence often comes from insight of the domain and is part of the modelling of the problem. Conditional indpendence can often be exploited to make things simpler (see later).

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

$$\mathbf{P}(X \mid Z_1, Z_2) = \alpha \mathbf{P}(X)\mathbf{P}(Z_1 \mid X)\mathbf{P}(Z_2 \mid X)$$

where α is the normalization constant.

Generalization: Recursive Bayesian Updating

$$\mathbf{P}(X \mid Z_1, \ldots, Z_n) = \alpha \mathbf{P}(X)\prod_{i=1}^{n}\mathbf{P}(Z_i \mid X)$$

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.

Marginalization

For any sets of variables Y and Z we have

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

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**.

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

Bayesian Networks

Domain knowledge/ assumptions:

- Events *Burglary* and *Earthquake* are independent. (of course, to be discussed: a burglary does not cause an earthquake, but a burglar might use an earthquake to do the burglary. Then the independence assumption is not true. This is a 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.

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 $Alarm$
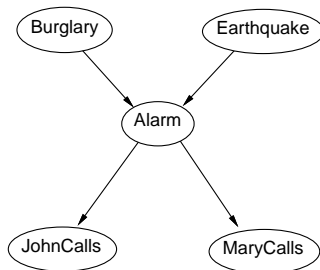
The Meaning of Bayesian Networks
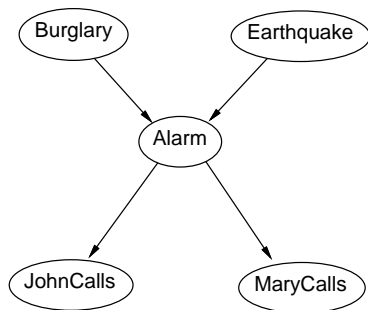
- $Alarm$ depends on $Burglary$ and $Earthquake$.
- $MaryCalls$ only depends on $Alarm$.
  $P(maryCalls \mid alarm, burglary) = P(maryCalls \mid alarm)$ or
  $P(maryCalls \mid alarm, burglary, johnCalls, earthquake) =$
  $P(maryCalls \mid alarm)$
$\rightarrow$ Bayesian Networks can be considered as sets of (conditional)
  independence assumptions.

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

Let all nodes $X_1, \ldots, X_n$ be ordered topologically according to the arrows in the network. Let $x_1, \ldots, x_n$ be the values of the variables. Then
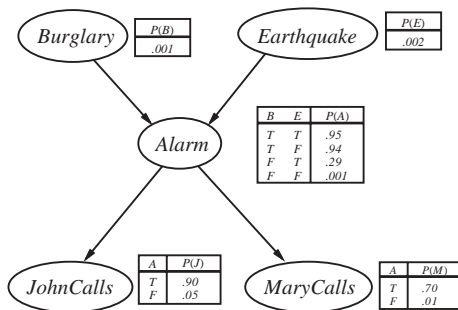
$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) \cdot \ldots \cdot P(x_2 \mid x_1)P(x_1)$$
$$= \prod_{i=1}^n P(x_i \mid x_{i-1}, \ldots, x_1)$$

From the independence assumption, this is equivalent to

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

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

Example

Only prob. for pos. events are given, negative: $P(\neg x) = 1 - P(x)$. Note: the size of the table depends on the number of parents!

$$P(j, m, a, \neg b, \neg e) =$$
$$P(j \mid m, a, \neg b, \neg e)P(m \mid a, \neg b, \neg e)P(la \mid \neg b, \neg e)P(\neg b \mid \neg e)P(\neg e)$$
$$= P(j \mid a)P(m \mid a)P(a \mid \neg b, \neg e)P(\neg b)P(\neg e)$$
$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.00062$$

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 $k = 5$

$\rightarrow$ $2^{20} = 1,048,576$ and $20 \times 2^5 = 640$ different explicitly-represented probabilities!

$\rightarrow$ In the worst case, a Bayesian network can become exponentially large, for example if every variable is directly influenced by all the others.

$\rightarrow$ 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.
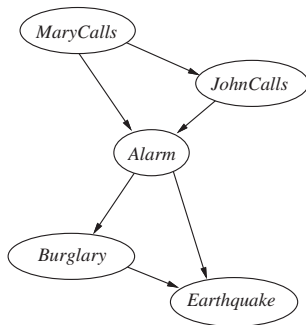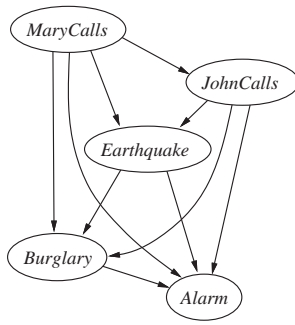
Example 1

M, J, A, B, E

Example 2

M, J, E, B, A

Example

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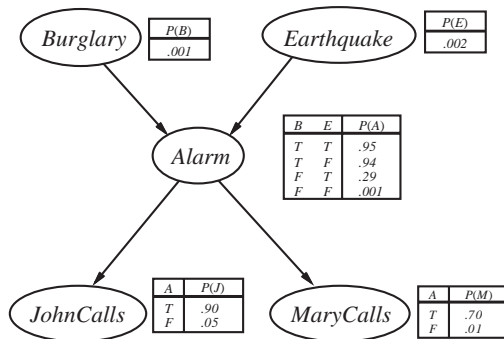

(a)   (b)

$\rightarrow$ Appears to be an 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.

Instantiating evidence variables and sending queries to nodes.

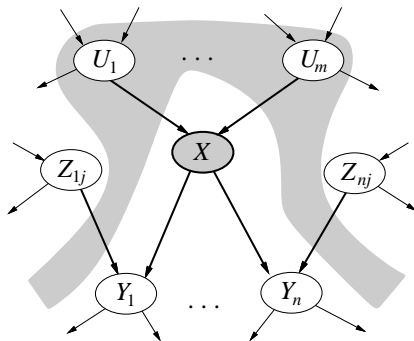| B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

	P(B)
	.001

	P(E)
	.002

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

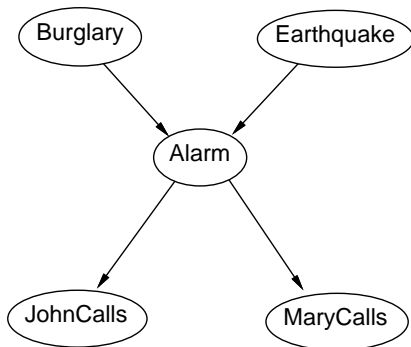| What is	$P(burglary \mid johncalls)$
or	$P(burglary \mid johnCalls, maryCalls)$?

Conditional Independence Relations in Bayesian Networks (1)

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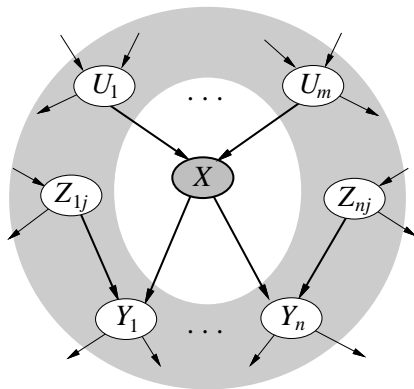
Example

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

# 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.

Example

Burglary is independent of *JohnCalls* and *MaryCalls*, given the values of *Alarm* and *Earthquake*, i.e.,
$P(Burglary \mid JohnCalls, MaryCalls, Alarm, Earthquake) = P(Burglary \mid Alarm, Earthquake)$
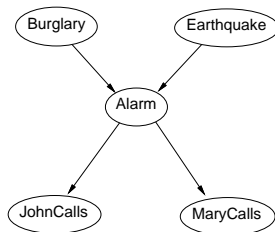
Exact Inference in Bayesian Networks

- 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 $P(X \mid e)$ where e are the observed values of E.

- In the remainder: X is a singleton

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

- $P(X \mid e) = \alpha P(X, e) = \sum_y \alpha P(X, e, y)$

- 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.

Example

- Consider $\mathbf{P}(Burglary \mid JohnCalls = true, MaryCalls = true)$

- The evidence variables are $JohnCalls$ and $MaryCalls$.

- The hidden variables are $Earthquake$ and $Alarm$.
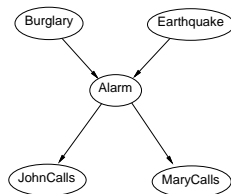
- We have: $\mathbf{P}(B \mid j, m) = \alpha \mathbf{P}(B, j, m)$
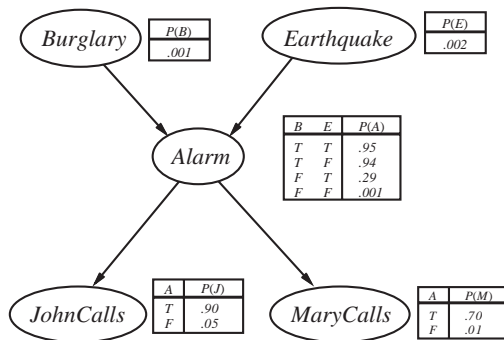$$= \alpha \sum_e \sum_a \mathbf{P}(B, j, m, e, a)$$

- If we consider the independence of variables, we obtain for $B = true$

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

- Reorganization of the terms yields:
$$P(b \mid j, m) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid e, b) P(j \mid a) P(m \mid a)$$

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

Evaluation of $P(b \mid j, m)$

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

$P(b)$
.001

$P(e)$
.002

$P(\neg e)$
.998

$P(a/b,e)$
.95

$P(\neg a/b,e)$
.05

$P(a/b,\neg e)$
.94

$P(\neg a/b,\neg e)$
.06

$P(j/a)$
.90

$P(j/\neg a)$
.05

$P(j/a)$
.90

$P(j/\neg a)$
.05

$P(m/a)$
.70

$P(m/\neg a)$
.01

$P(m/a)$
.70

$P(m/\neg a)$
.01
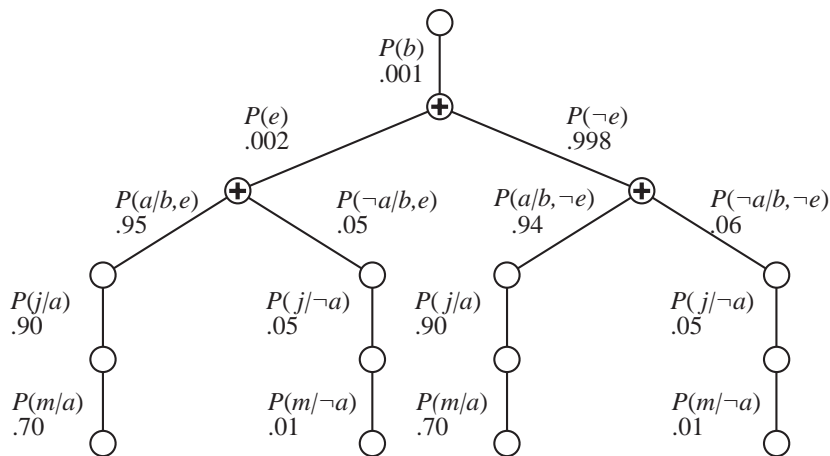
$$\mathbf{P}(B \mid j, m) = \alpha(0.0006, 0.0015) = (0.284, 0.716)$$

Enumeration Algorithm for Answering Queries on Bayesian Networks

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

  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X **do**
      $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(bn.VARS, $\mathbf{e}_{x_i}$)
          where $\mathbf{e}_{x_i}$ is $\mathbf{e}$ extended with $X = x_i$
  return NORMALIZE($\mathbf{Q}(X)$)

---

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

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 in the worst case, all terms must be evaluated for the two cases ("true" and "false")

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 (bottom-up) and to save results for later use.

- Additionally, unnecessary expressions can be removed.

Example

- Let us consider the query $P(JohnCalls \mid Burglary = true)$.

- The nested sum is

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

- Obviously, the rightmost sum equals 1 so that it can safely be dropped.

- general observation: variables, that are not query or evidence variables and not ancestor nodes of query or evidence variables can be removed. Variable elimination repeatedly removes these variables and this way speeds up computation.

- within example: $Alarm$ and $Earthquake$ are ancestor nodes of query variable $JohnCalls$ and can not be removed. $MaryCalls$ is neither a query nor an evidence variable and no ancestor node. Therefore it can be removed.

Complexity of Exact Inference

- If the network is singly connected or a polytree (at most one undirected path between two nodes in the graph), 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.
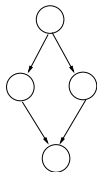
- There are approximate inference methods for multiply connected networks such as sampling techniques or Markov chain Monte Carlo.

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.

Other Approaches (2)

- Dempster-Shafer Theory

  - Allows the representation of *ignorance* as well as uncertainly.

  - Example: If a coin is fair, we assume $P(Heads) = 0.5$. But what if we don't know if the coin is fair? $\rightarrow Bel(Heads) = 0$, $Bel(Tails) = 0$. If the coin is 90% fair, 0.5×0.9, i.e. $Bel(Heads) = 0.45$.

$\rightarrow$ Interval of probabilities is $[0.45, 0.55]$ with the evidence, $[0, 1]$ without.

$\rightarrow$ The notion of utility is not yet well understood in Dempster-Shafer Theory.

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*.

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.