Reasoning under Uncertainty

Introduction to Artificial Intelligence

G. Lakemeyer

Winter Term 2016/17

© G. Lakemeyer

Uncertainty

Toothache (T) and Cavity (C)

Would like to say:

If T then C is likely (T $\Rightarrow$ C)

There are various ways to model this. Here: statistical interpretation

"T $\Rightarrow$ C has probabilty 0.8."
(80% of those with T have C.)

$\Rightarrow$ **Probability Theory**

We only need basic discrete Probability Theory (suffices for most AI purposes).

We often abbreviate probability as P.

© G. Lakemeyer

Probability Distributions

P(A) is the probability that proposition A holds, where A is a Boolean combination ($\wedge, \vee, \neg$) of atomic propositions.

We also allow $X = n$ as an atomic proposition, where X is called a random variable and n is taken from a discrete domain.

Example:

Random variable *weather* with values from the sequence $\langle$ sunny, rain, cloudy, snow $\rangle$.

$$\begin{array}{lcl} P(weather = sunny) & = & 0.7 \\ P(weather = rain) & = & 0.2 \\ P(weather = cloudy) & = & 0.08 \\ P(weather = snow) & = & 0.02 \end{array}$$

P(weather) = (0.7; 0.2; 0.08; 0.02)

stands for the **probability distribution** of the random variable weather.

© G. Lakemeyer

Probability Theory and Decisions

Let A_1, A_2 und A_3 be plans to get to the airport ontime. Let $\mathrm{P}(A_i)$ be the probability that executing A_i allows us to get to the airport ontime.

$$\begin{array}{lcll} \mathrm{P}(A_1) & = & 0.9 & \text{(Leave 2 hour before departure)} \\ \mathrm{P}(A_2) & = & 0.96 & \text{(Leave 3 hours before departure)} \\ \mathrm{P}(A_3) & = & 0.9999 & \text{(Leave 12 hours before departure)} \end{array}$$

Note: The maximum probability need not be optimal. One needs to consider the utility of actions as well.

Decision Theory = Probability Theory $+$ Utility Theory.

© G. Lakemeyer

The Axioms of Probability Theory

1. P is a real number between 0 and 1. ($0 \leq P(A) \leq 1$)
2. P(true) = 1, P(false) = 0
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

© G. Lakemeyer

Why Are These Axioms Reasonable?

de Finetti (1930s) said that someone who believes P(A) should be willing to bet on P(A)!

“Theorem” (de Finetti):

Someone who does not follow the axioms of probability theory will lose his or her bet!

Example:

Player 1		Player 2	
Proposition	Belief	Bets on	
A	0.4	A	4 to 6
B	0.3	B	3 to 7
$A \vee B$	0.8	$\neg(A \vee B)$	2 to 8

© G. Lakemeyer

Conditional Probabilities (1)

Rolling dice

P(roll=3) = 1/6. Let E = “Roll is divisible by 3.”

Then we obtain the **conditional probability**:

$$P(\text{roll}=3 \mid E) = 1/2.$$

E is also called the evidence.

E often plays the role of background knowledge (similar to a propositional knowledge base).

Prior:	Probability before evidence.
Posterior:	Probability after the evidence.

© G. Lakemeyer

Conditional Probabilities (2)

$P(C \mid T) = 0.8$ is read as "the prob. to have C when T is given is 0.8."

Nonmonotonicity:

$P(\text{Flies} \mid \text{Bird}) = 0.99$
$P(\text{Flies} \mid \text{Bird} \wedge \text{Antarctica}) = 0.4$
$P(\text{Flies} \mid \text{Bird} \wedge \text{Antarctica} \wedge \text{Albatross}) = 0.999$

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \quad \text{or} \quad P(A \wedge B) = P(A \mid B) \cdot P(B)$$

(Product Rule)

$\mathbf{P}(X,Y) = \mathbf{P}(X \mid Y) \cdot \mathbf{P}(Y)$ stands for a system of equations of the form:

$$P(X{=}x_i \wedge Y = y_j) = P(X{=}x_i \mid Y = y_j) \cdot P(Y = y_j)$$

for all x_i, y_j of the domains of X and Y.

© G. Lakemeyer

Joint Distributions

Let $X_1, X_2, \ldots X_n$ be random variables. An atomic event is an assignment of values to all variables X_i.

The joint distribution $\mathbf{P}(X_1, X_2, \ldots X_n)$ assigns a probability to all atomic events.

Toothache-Cavity Example:

	T	¬T
C	0.04	0.06
¬C	0.01	0.89

- Atomic events exclude one another.
- $\sum_{x_1,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n) = 1$
- From the table one can read off all probabilities.
- Problem: Table grows exponentially in the number of variables. Thus probability-based systems work directly with conditional probs.

© G. Lakemeyer

Bayes Rule

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

Thus we have:

$$P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A),$$

from which we obtain:

Bayes Rule:

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

For variables X and Y we write

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

which again corresponds to a system of equations.
Often there is additional evidence E:

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

© G. Lakemeyer

Normalisation

For

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

$1/\mathbf{P}(X)$ only plays the role of a normalising constant so that the right-hand side sums to 1 over all values of Y.

In the literature one therefore often finds the following form:

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

In practice one usually calculates the unnormalised case first, and then looks for an appropriate α.

© G. Lakemeyer

Combining Evidence

How does one combine evidence consisting of several variables??

Example: (A = “Catch”)

How do we get from $P(C \mid T)$ to $P(C \mid T \wedge A)$, that is, how does the probability of cavity change if one also finds out that there is a catch.

According to Bayes we have:

$$P(C \mid T \wedge A) = \frac{P(T \wedge A \mid C) \cdot P(C)}{P(T \wedge A)}.$$

The term $P(T \wedge A \mid C)$ is problematic. For *n* variables we would need to compute 2^n combinations. If there are many variables as evidence, we have exponential growth!

© G. Lakemeyer

Bayesian Update

An elegant and efficient solution is possible when certain conditional independence assumptions can be made:

$$P(A \mid C \wedge T) = P(A \mid C) \qquad (**)$$

"If C is given, then T and A are independent of each other."

In that case, the evidence can be added one by one using a simple iterative method. According to Bayes (slightly reformulated):

$$P(C \mid T \wedge A) = P(C) \cdot \frac{P(T \mid C)}{P(T)} \cdot \frac{P(A \mid T \wedge C)}{P(A \mid T)}$$

Together with (**) we then obtain

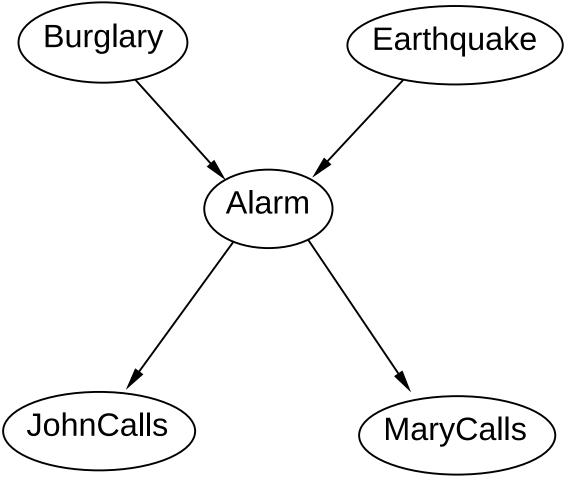
$$P(C \mid T \wedge A) = P(C \mid T) \cdot \frac{P(A \mid C)}{P(A \mid T)}$$

In general, for multi-valued X, Y, Z:

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

© G. Lakemeyer

Belief Networks (1)

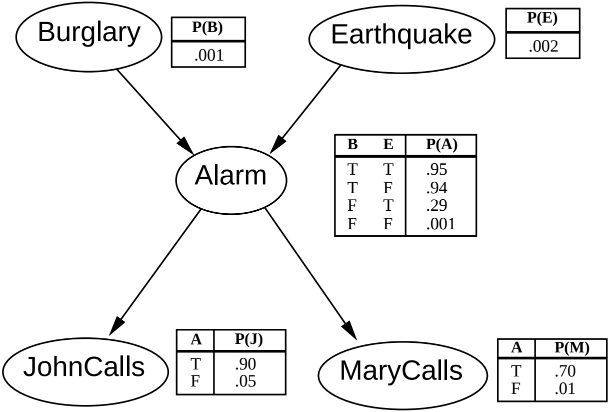


Idea: Only represent causal connections. Surprisingly simple in many applications!

© G. Lakemeyer

Belief Networks (2)

Same Example with labelled nodes $\mathbf{P}(X \mid Parents(X))$:



P(B)
.001

P(E)
.002

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

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

© G. Lakemeyer

Belief Networks in General

A belief network is an acyclic graph where

- the nodes represent random variables;
- each node X is labelled with the conditional probabilities

$$\mathbf{P}(X \mid \mathit{Parents}(X)),$$

where Y is in $\mathit{Parents}(X)$ if there is an edge from Y to X.
(The label is called a Conditional Probability Table (CPT).)

The tolopology of the network should be chosen in such a way that for each edge from Y to X, the parent node Y has direct causal influence on X.

© G. Lakemeyer

Belief Networks and Joint Distributions

Let $X_1, \ldots, X_n$ be random variabls. We abbreviate $P(X_1 = x_1 \wedge \cdots \wedge X_n = x_n)$ as $P(x_1, \ldots, x_n)$.

We can rewrite the joint distribution in the following way:

$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) \times P(x_{n-1}, \ldots, x_1)$$

Applying this rewriting recursively we get

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

A Belief network is a correct representation of a joint distribution if

$$\mathbf{P}(X_i \mid X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i \mid Parents(X_i)) \text{ and } Parents(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}.$$

In other words, each node must be conditionally independent of its predecessors given its parents.

© G. Lakemeyer

Constructing a Belief Network

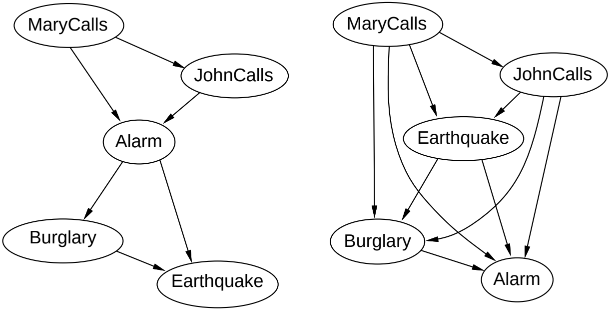
1. Choose the relevant random variables that descibe your application.
2. Choose an ordering $X_1, X_2, \ldots, X_n$ for those variables.
3. **While** there are still variables to consider **do**
   - Choose the least X_i and create a node in the network
   - Parents(X_i) := minimal set so that the following holds:

$$\mathbf{P}(X_i \mid X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i \mid Parents(X_i))$$

   - Create a CPT for X_i

© G. Lakemeyer

The Ordering of Nodes Matters



A bad choice in the ordering of the variables leads to large networks.

Orderings in the examples:

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

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

© G. Lakemeyer

d-Separation (1)

A set of nodes **E** is said to d-separate the sets of nodes **X** and **Y** iff every undirected path from a node in **X** to a node in **Y** is blocked by **E**. Blocking means that there is a node Z on this path such that one of the following conditions hold:

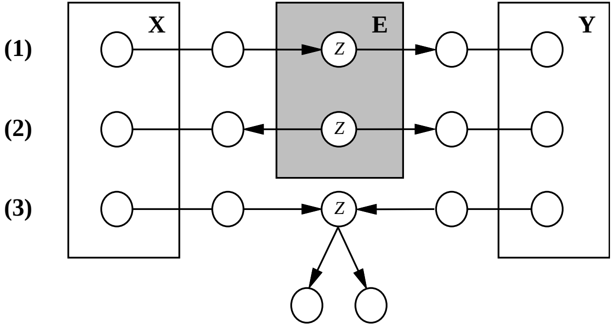
1. $Z \in \mathbf{E}$ and one directed edge on the path leads into Z and another points away from Z.
2. $Z \in \mathbf{E}$ and both edges point away from Z.
3. Neither Z nor any of its successors are in E and both edges connected to Z on the path lead into Z.

© G. Lakemeyer

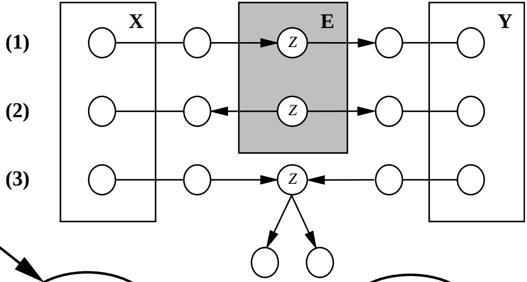
d-Separation (2)

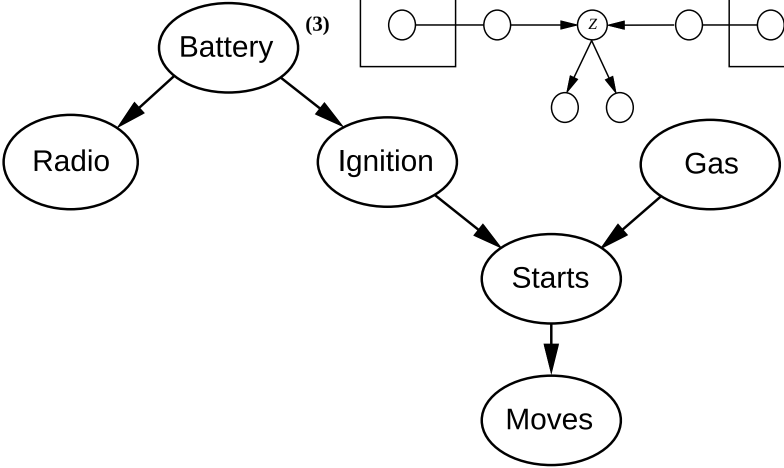
© G. Lakemeyer

3 ways of blocking paths from X to Y:



Examples for d-Separation





© G. Lakemeyer

Why d-Separation is Important

Theorem (Judea Pearl):

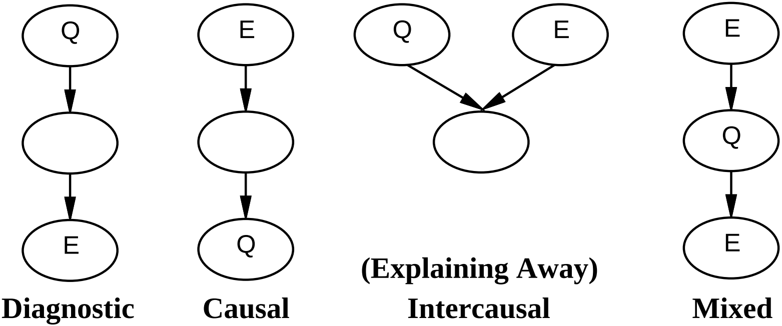
If **E** d-separates **X** from **Y**, then **X** is independent of **Y** given **E**.

Note: d-separation

- can be computed in polynomial time;
- is incomplete, that is, not every conditional independence is detected;
- is nevertheless sufficient for a number of inference algorithms.

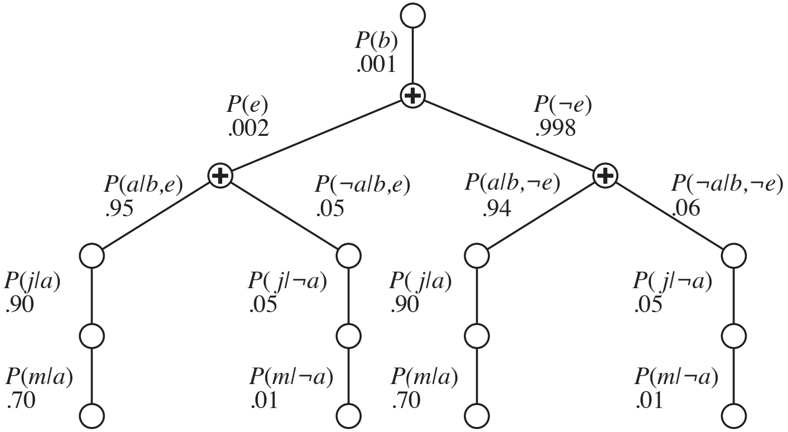
© G. Lakemeyer

Kinds of Inferences in Belief Networks



© G. Lakemeyer

Example Computation



© G. Lakemeyer

Computational Complexity

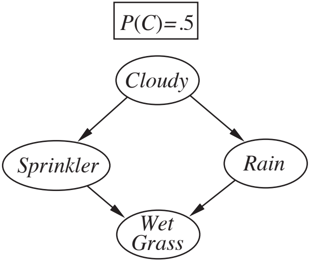
The problem is NP-hard for multiply connected networks.

(Actually, the problem is at least as hard as enumerating all satisfying assignments of a propositional formula (#P-hard), which is strictly harder than NP-completeness.)

It is linear in the case of singly connected networks.

© G. Lakemeyer

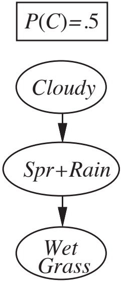
Clustering



C	$P(S)$
t	.10
f	.50

C	$P(R)$
t	.80
f	.20

S	R	$P(W)$
t	t	.99
t	f	.90
f	t	.90
f	f	.00



C	$P(S{+}R{=}x)$			
	$t\,t$	$t\,f$	$f\,t$	$f\,f$
t	.08	.02	.72	.18
f	.10	.40	.10	.40

$S{+}R$	$P(W)$
$t\ t$	.99
$t\ f$	.90
$f\ t$	.90
$f\ f$	.00

© G. Lakemeyer