Lecture 11: Rational Decision Making with Uncertainty

Viliam Lisý & Branislav Bošanský

Artificial Intelligence Center
Department of Computer Science, Faculty of Electrical Eng.
Czech Technical University in Prague

viliam.lisy@fel.cvut.cz

April, 2021

Plan of today's lecture

1. Rationality as expected utility maximization
2. Reasoning with joint probability distribution
3. Bayesian networks
4. Decision networks

Acknowledgements

Slides are closely following AIMA 3rd edition (mainly Ch. 13, 14)

Further based on:

- Percy Liang's lecture
- Patrick Winston's lecture
- Roman Bartak's lecture

Rational agent

Rational agent chooses the actions that maximise its expected utility over all possible outcomes.

$$\begin{aligned} \textit{rational decisions} &= \textit{decision theory} \\ &= \textbf{utility}\ \textit{theory} + \textbf{probability}\ \textit{theory} \end{aligned}$$

Rationality and its limitation is often studied in the form of lotteries, e.g., Would you rather have:

- 20% chance of winning \$100 or
- 50% chance of winning \$20?

To build intelligent (rational) agents, we need to assess the utility and the probability of various events.

Typical AI Applications with Probabilistic Reasoning

Diagnosis support

- Medical, IT support, machinery service, etc.

Robotic localisation: what is the position of the robot given

- Noisy actuators
- Multiple noisy sensors

Natural language processing

- What is the topic of a text given its words?

Basic (discrete) probability recapitulation

Random variable: sunshine $S \in \{0, 1\}$, rain $R \in \{0, 1\}$, dice $D \in \{1, 2, 3, 4, 5, 6\}$.

Joint distribution:

$P(S, R) =$

s	r	$P(S = s, R = r)$
0	0	0.20
0	1	0.08
1	0	0.70
1	1	0.02

Marginal distribution:

$P(S) =$

s	$P(S = s)$
0	0.28
1	0.72

(sum rows)

Conditional distribution:

$P(S|R = 1) =$

s	$P(S = s\|R = 1)$
0	0.8
1	0.2

(select rows + normalize)

Basic probability statements

Random variables are exhaustive and mutually exclusive:

$$\sum_{a \in A} P(A = a) = 1$$

Inclusion-exclusion principle:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

Product rule:

$$P(A, B) = P(A|B)P(B)$$

Bayes' rule:

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

Inference Using Full Joint Distribution

Knowledge base:

	toothache		$\neg$*toothache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	0.108	0.012	0.072	0.008
$\neg$*cavity*	0.016	0.064	0.144	0.576

Updating belief based on evidence:

$$P(cavity|toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

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

The size of full joint distribution grows exponentially

Slightly more realistic example:

$Cavity \in \{true, false\}$	$Sex \in \{male, female\}$	$Hygene \in$
$Toothache \in \{true, false\}$	$Age \in \{child, teen, adult, senior\}$	$History \in$
$Catch \in \{true, false\}$	$Diet \in \{omnivore, vegetarian, vegan\}$	$\ldots$

The size of the joint probability table for problem with variables $X_1, \ldots X_n$ is:

$$\prod_{i=1}^{n} |X_i| \geq 2^n$$

We need to:

- Store the data in memory
- Iterate over large portions of them to answer queries
- Obtain a probability for each cell!

Absolute Independence

Assume variables:
$Cavity, Toothache, Catch, Weather \in \{cloudy, sunny, rain, snow\}$
The size of $P(Cavity, Toothache, Catch, Weather)$ is $2 \times 2 \times 2 \times 4 = \mathbf{32}$.
We know that

$$P(Cavity, Toothache, Catch, Weather) =$$
$$P(Weather|Cavity, Toothache, Catch)P(Cavity, Toothache, Catch)$$

Dental problems do not influence the weather, hence:

$$P(Weather|Cavity, Toothache, Catch) = P(Weather)$$

Therefore without loss of precision, we can represent

$$P(Cavity, Toothache, Catch, Weather) =$$
$$P(Cavity, Toothache, Catch) \quad \text{of size } 2 \times 2 \times 2 = 8$$
$$* P(Weather) \quad \text{of size } 4$$

The overall size of the representation is $8 + 4 = \mathbf{12}$.

Conditional Independence

Absolute independence is quite rare. We can use conditional independence to reduce the representation size further.
When one has cavity, does catch depend on toothache?

$$P(Catch|Toothache, Cavity) = P(Catch|Cavity)$$

Variables X and Y are independent given Z, if we any of the following holds:

$$P(X|Y,Z) = P(X|Z), P(Y|X,Z) = P(Y|Z), P(X,Y|Z) = P(X|Z)P(Y|Z)$$

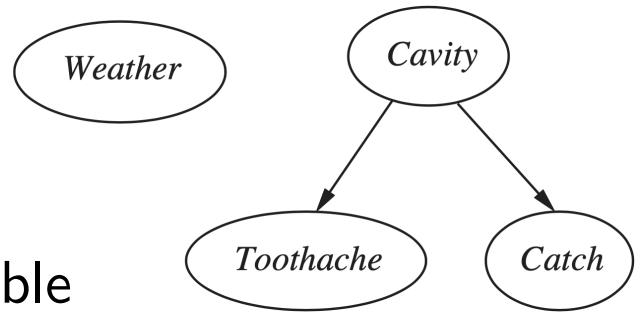
$$\begin{aligned} P(Toothache, Catch, Cavity) = \\ P(Toothache|Catch, Cavity)P(Catch|Cavity)P(Cavity) = \\ = P(Toothache|Cavity) \quad \text{of size } 2 \times 2 = 4 \\ *P(Catch|Cavity) \quad \text{of size } 2 \times 2 = 4 \\ *P(Cavity) \quad \text{of size } 2 \end{aligned}$$

It does not lead to savings here, but often does in large problems.

Bayesian Network

Formal framework for compact representation and inference in large joint distributions.

It specifies the **conditional independence relationships among random variables** and the corresponding necessary joint distributions.

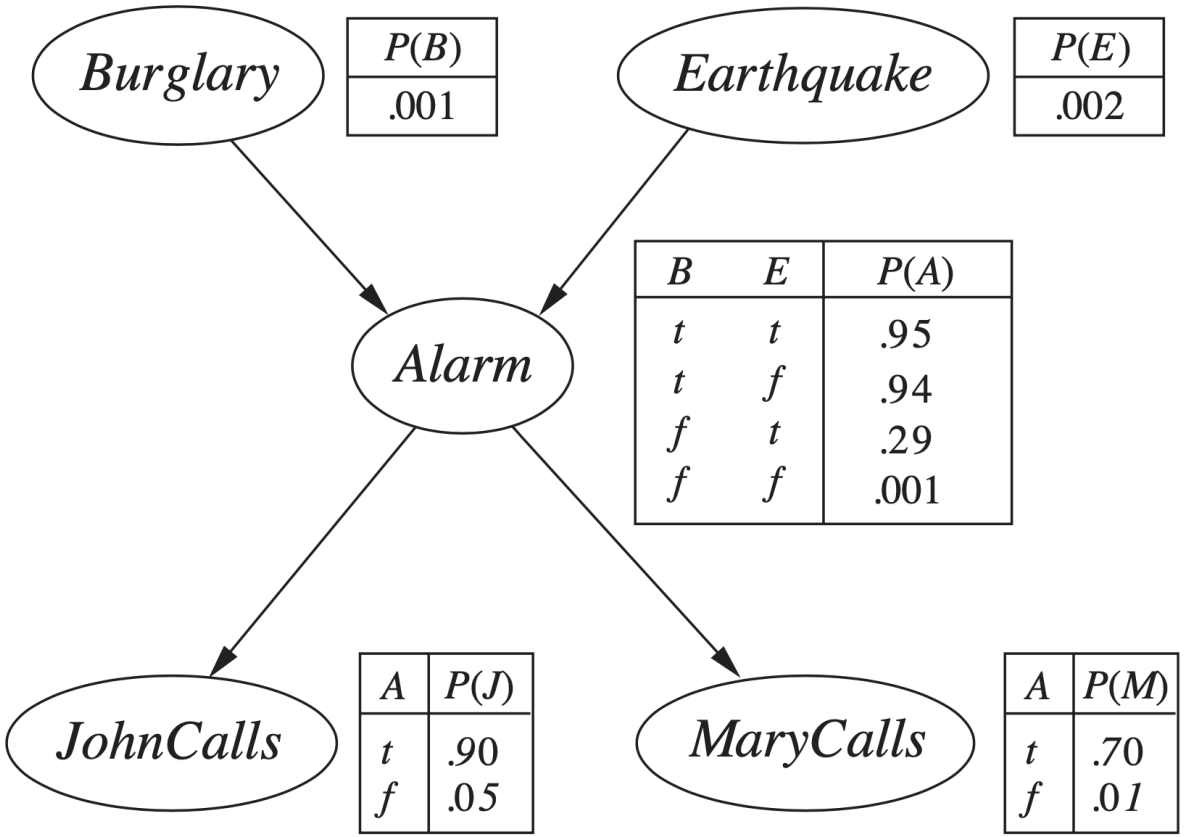


Bayesian network consists of:

- A graph node for each random variable
- Directed edges from parents to children represented direct influence of the children's values by the parent values. The edges form a Directed Acyclic Graph (DAG).
- For each node X_i, a conditional probability table $P(X_i|Parents(X_i))$

Bayesian Network Example

An **alarm** usually sounds when a **burglary** is in progress, but sometimes it is started by a minor **earthquake**. There are two neighbours which may here the alarm and call us and **John** is more likely to call than **Mary**.



How does a BN represent the joint distribution?

From the chain rule (iterative application of the product rule) we know:

$$P(J, M, A, B, E) = P(J|M, A, B, E) * P(M|A, B, E) * P(A|B, E) * P(B|E) * P(E)$$

From conditional independence of individual variables

$$= P(J|A) * P(M|A) * P(A|B, E) * P(B) * P(E)$$

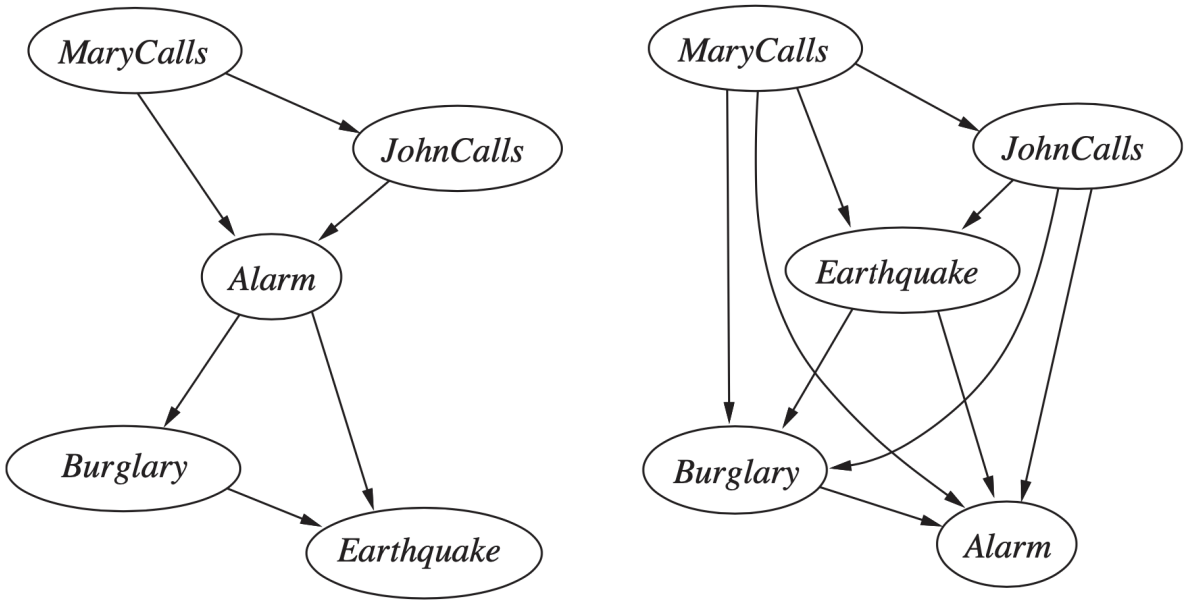
And these are exactly the tables included in the BN

$$= \prod_i P(X_i|Parents(X_i))$$

Since BN is a DAG, the topological ordering will always provide a correct ordering of the variables.

Other structures may also represent the distibution

While the edges are easy to think about as **causality**, it is not necessarily the case.



Conditional independence in BNs

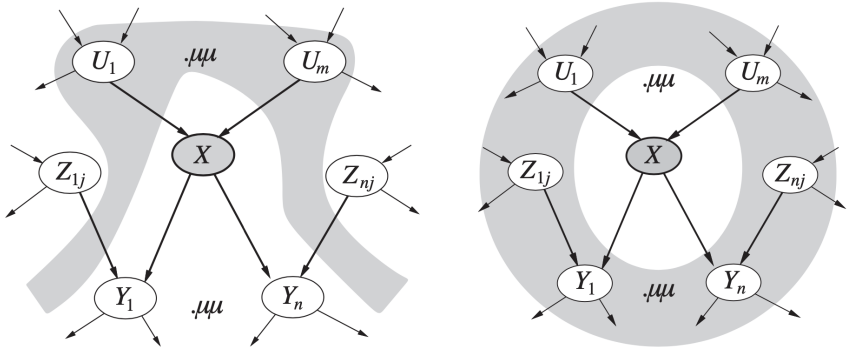
We required that only node's parents influence its value; hence

(1) a node is conditionally independent of its predecessors, given its parents.

The BN's topology captures other independence relationships:

(2) a node is conditionally independent of its non-descendants, given its parents;

(3) a node is conditionally independent of all other nodes, given its parents, children, and children's parents.



Understanding independence speeds up the inference algorithms.

Inference in BNs

Task: Compute the **posterior** probability distribution over a set of **query variables** ($\mathbf{X}$), given some observed assignments for **evidence variables** ($\mathbf{E} = \mathbf{e}$).

Let $\mathbf{Y}$ be the set of non-query and non-evidence variables, the task is

$$P(\mathbf{X}|\mathbf{e}) = \frac{P(\mathbf{X}, \mathbf{e})}{P(\mathbf{e})} = \alpha P(\mathbf{X}, \mathbf{e}) = \alpha \sum_{\mathbf{y} \in \mathbf{Y}} P(\mathbf{X}, \mathbf{e}, y),$$

where the last joint distribution is represented by the BN as the product

$$\prod_i P(X_i | Parents(X_i)).$$

Inference in BNs by enumeration

In the burglary example, assume that both John and Marry called and we are interested in the probability of the burglary.

$\mathbf{X} = \{Burglary\}$, $\mathbf{E} = \{MaryCalls, JohnCalls\}$, $\mathbf{Y} = \{Alarm, Earthquake\}$

Than from the previous

$$P(B|j,m) = \alpha P(B,j,m) = \alpha \sum_{e \in E} \sum_{a \in A} P(B,j,m,e,a)$$

For $Burglary = true$

$$P(b|j,m) = \alpha \sum_{e \in E} \sum_{a \in A} P(b)P(e)P(a|b,e)P(j|a)P(m|a)$$

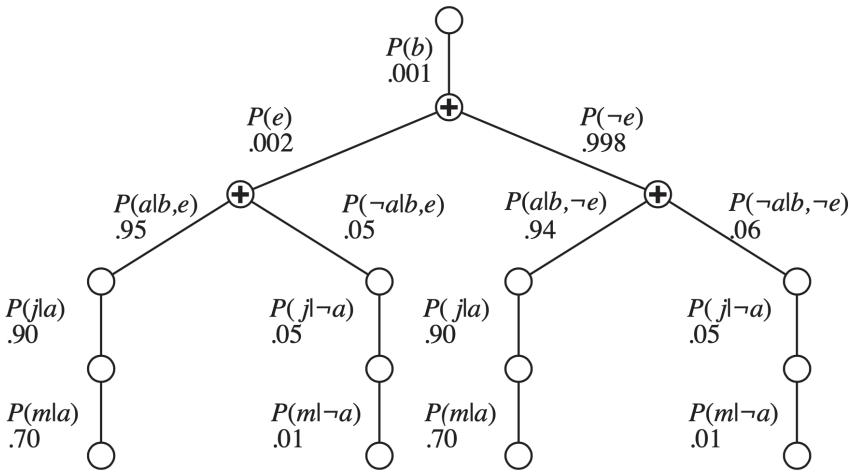
It works, but requires iterating all combinations of variables in $\mathbf{Y}$.

Inference in BNs by enumeration

$$P(b|j,m) = \alpha \sum_{e \in E} \sum_{a \in A} P(b)P(e)P(a|b,e)P(j|a)P(m|a)$$
$$= \alpha P(b) \sum_{e \in E} P(e) \sum_{a \in A} P(a|b,e)P(j|a)P(m|a)$$

The computation can be visualised in a tree.

- Multiplication by $P(b)$ only once
- Identical subtrees can be re-used



The complexity of exact inference in BNs

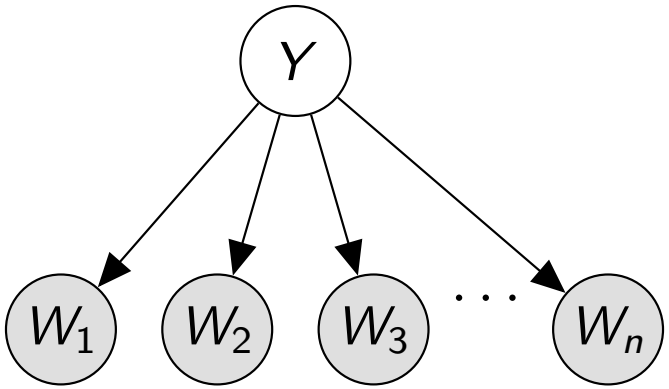
If the BN is a **polytree** (a tree disregarding the edge orientation), then the time and space complexity is linear in the overall number of conditional probability table entries $O(n.d^k)$.

For general BNs, the problem is **#P-hard** (harder than NP-complete).

Special types of BNs: Naïve Bayes classifier

Simple classifier, used for example for spam filtering.

An unobservable class $Y \in \{ham, spam\}$ influences the probability of occurrence of individual words, e.g., $W_1 = bargin$, $W_2 = socrates$, etc.



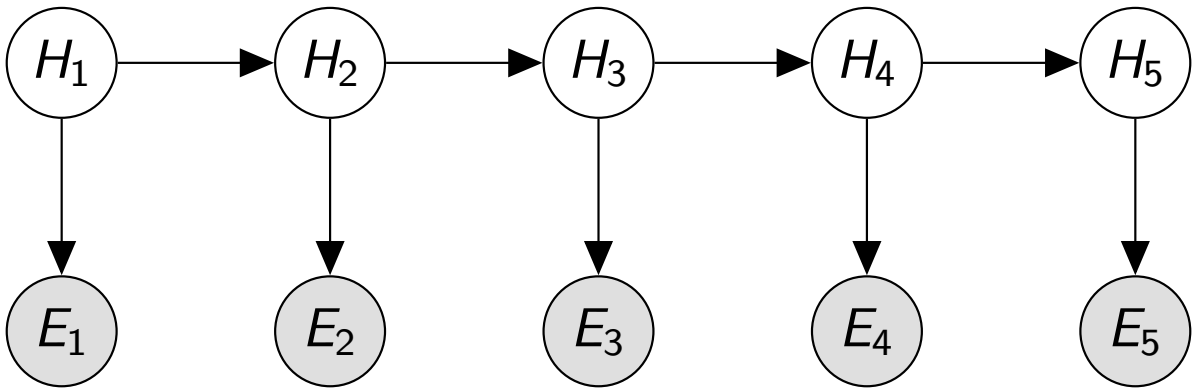
Given an email, what is the probability that it is spam?

$$P(spam|\mathbf{w}) = \alpha P(spam) \prod_{i=1}^{n} P(w_i|spam)$$

Special types of BNs: Hidden Markov moddel

Special case of **dynamic** Bayesian network, used for example for robotic localisation or speech recognition.

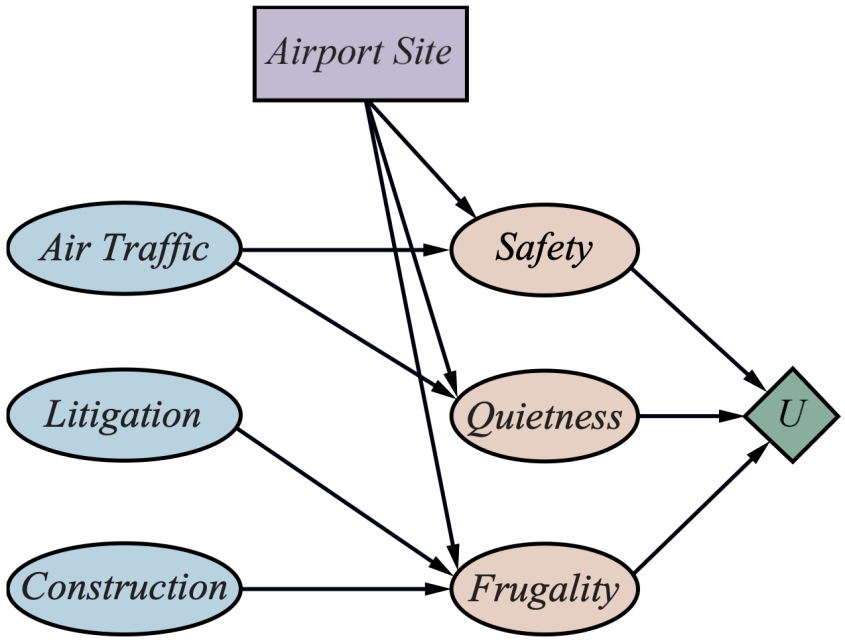
An unobservable true position H_t in time step t. Noisy sensor reading about the position E_t.



Given last sensor readings, what is the probability of some position?

Decision networks

Influence diagrams (aka. decision networks) combine Bayesian networks with additional nodes for decisions (rectangles) and utilities (diamonds).



In the simplest form, each decision is evaluated in the same way as a chance node with evidence and the decision that maximises the expected utility is selected.

Summary

Many AI problems require reasoning with uncertainty

Joint probability distribution is a powerful knowledge representation

- But they are growing exponentially, which causes problems

Bayesian networks compactly represent joint distributions

- Represent exactly the joint distribution in smaller space
- Inference is still exponential in general
- Inference is polynomial with special structure

Bayesian networks are a general framework with many specializations

- Naive Bayes, HMM

Influence diagrams introduce decisions and utilities and allow expected utility maximization