CS-E4800 Artificial Intelligence

Jussi Rintanen

Department of Computer Science
Aalto University

February 13, 2020

Today's Lecture

- Logic vs. Probabilities
- Bayesian Networks
- Probabilistic reasoning with BN

Motivation

- Uncertainty is inherent in many applications
  - Uncertainty in causality: causes $\rightarrow$ effects
  - Medical diagnosis: illness – symptoms
- Logic and related methods good at handling incomplete information
- Quantifying the incompleteness/uncertainty can be critical
  - Many types of decision-making need probabilities
  - Example: Justify (expensive) diagnostic test by its impact on probabilities

Probabilistic vs. Logical Reasoning

Deductive (logical) reasoning:

- $(a \to b) \wedge (b \to c) \models a \to c$
- "All cats are felines", "All felines are mammals" entail "All cats are mammals"?

Probabilistic reasoning:

- "All Finns are Europeans", "Most Europeans live south of the Baltic Sea" entails "Most Finns live south of the Baltic Sea"?
- $P(B|A) = 0.999, P(C|B) = 0.999$ entails $P(C|A) =$???

Probabilistic vs. Logical Reasoning

Deductive (logical) reasoning:

- $(a \to b) \wedge (b \to c) \models a \to c$
- "All cats are felines", "All felines are mammals" entail "All cats are mammals"?

Probabilistic reasoning:

- "All Finns are Europeans", "Most Europeans live south of the Baltic Sea" entails "Most Finns live south of the Baltic Sea"?
- $P(B|A) = 0.999, P(C|B) = 0.999$ entails $P(C|A) =$???

Valuations vs. Probability Distributions

Truth of a formula under all valuations

a	b	c	ϕ
0	0	0	1
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

Logic: something is possible, or it is not possible

Probability theory: quantifying the possibility

Valuations vs. Probability Distributions

Probabilities of all valuations = full joint distribution

a	b	c	P
0	0	0	0.05
0	0	1	0.05
0	1	0	0
0	1	1	0.1
1	0	0	0.2
1	0	1	0
1	1	0	0.6
1	1	1	0

Logic: something is possible, or it is not possible
Probability theory: quantifying the possibility

Where to Obtain Probability Distributions?

Construct from raw data:

100111	001101	010001	110110	001110
111000	010110	101111	110111	110111
001101	011001	100010	010111	100100
110100	001100	001110	110110	010101
110001	110111	011100	010110	100000
010000	001000	001100	000111	000101
110001	100110	000100	011111	111101
111000	110101	001010	010001	110101
110000	010110	000010	010111	010101
000000	100010	011010	001101	001101
111110	110001	010011	111111	011001
101111	100001	110000	101101	010001
010110	111000	110000	000001	010001
000011	011100	000100	110101	000111
011011	011000	010000	011110	100110
110110	011011	111111	000010	110000
001010	110101	010001	110110	101101
100001	101001	111011	010101	011001

Represent by statements:

$$P(A) = 0.3$$
$$P(B|C) = 0.9$$
$$P(A \vee C) = 0.1$$
$$\ldots$$

Represent as a Bayesian network

Other graphical models

Which option? Depends on the application:

- Probability of a certain purchase combination in the local supermarket
- Probability of a certain fault combination in a nuclear power plant

Where to Obtain Probability Distributions?

Construct from raw data:

100111	001101	010001	110110	001110
111000	010110	101111	110111	110111
001101	011001	100010	010111	100100
110100	001100	001110	110110	010101
110001	110111	011100	010110	100000
010000	001000	001100	000111	000101
110001	100110	000100	011111	111101
111000	110101	001010	010001	110101
110000	010110	000010	010111	010101
000000	100010	011010	001101	001101
111110	110001	010011	111111	011001
101111	100001	110000	101101	010001
010110	111000	110000	000001	010001
000011	011100	000100	110101	000111
011011	011000	010000	011110	100110
110110	011011	111111	000010	110000
001010	110101	010001	110110	101101
100001	101001	111011	010101	011001

Represent by statements:

$$P(A) = 0.3$$
$$P(B|C) = 0.9$$
$$P(A \vee C) = 0.1$$
$$\ldots$$

Represent as a Bayesian network

Other graphical models

Which option? Depends on the application:

- Probability of a certain purchase combination in the local supermarket
- Probability of a certain fault combination in a nuclear power plant

Representing a Probability Distribution

1. Enumerate probabilities of all valuations:

$$P(A, B, C, D, E) = 0.02$$
$$P(A, B, C, D, \neg E) = 0.01$$
$$P(A, B, C, \neg D, E) = 0.1$$
$$P(A, B, C, \neg D, \neg E) = 0.03$$
$$\cdots P(\neg A, \neg B, \neg C, \neg D, \neg E) = 0.01$$

   2^n statements for n variables: not feasible for high n

2. Represent some facts about probabilities explicitly:

$$P(A) = 0.3 \quad P(B) = 0.1 \quad P(C|A, B) = 0.5 \quad P(D|B) = 0.1$$

   Most dependencies left open: multiple possible distributions

3. Is there a more convenient way of representing joint distributions?

Representing a Probability Distribution

1. Enumerate probabilities of all valuations:

$$P(A, B, C, D, E) = 0.02$$
$$P(A, B, C, D, \neg E) = 0.01$$
$$P(A, B, C, \neg D, E) = 0.1$$
$$P(A, B, C, \neg D, \neg E) = 0.03$$
$$\cdots P(\neg A, \neg B, \neg C, \neg D, \neg E) = 0.01$$

   2^n statements for n variables: not feasible for high n

2. Represent some facts about probabilities explicitly:

$$P(A) = 0.3 \quad P(B) = 0.1 \quad P(C|A, B) = 0.5 \quad P(D|B) = 0.1$$

   Most dependencies left open: multiple possible distributions

3. Is there a more convenient way of representing joint distributions?

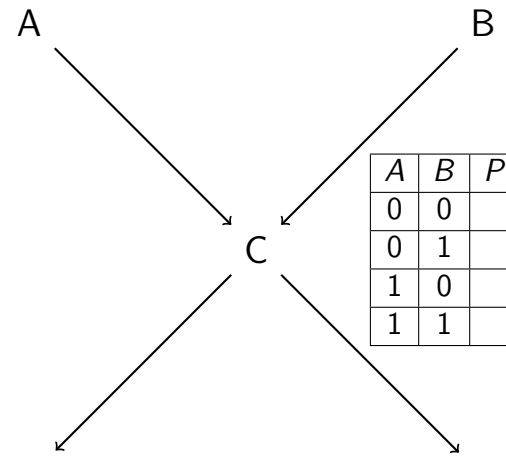
Bayesian Networks

- Compact representation of probability distributions
- Directed acyclic graphs, with variables as nodes
- Arcs indicate dependencies
- Paths indicate indirect dependencies
- Independence implicit in representation

Bayesian Networks: Example

$P(A)$
0.3

$P(B)$
0.9

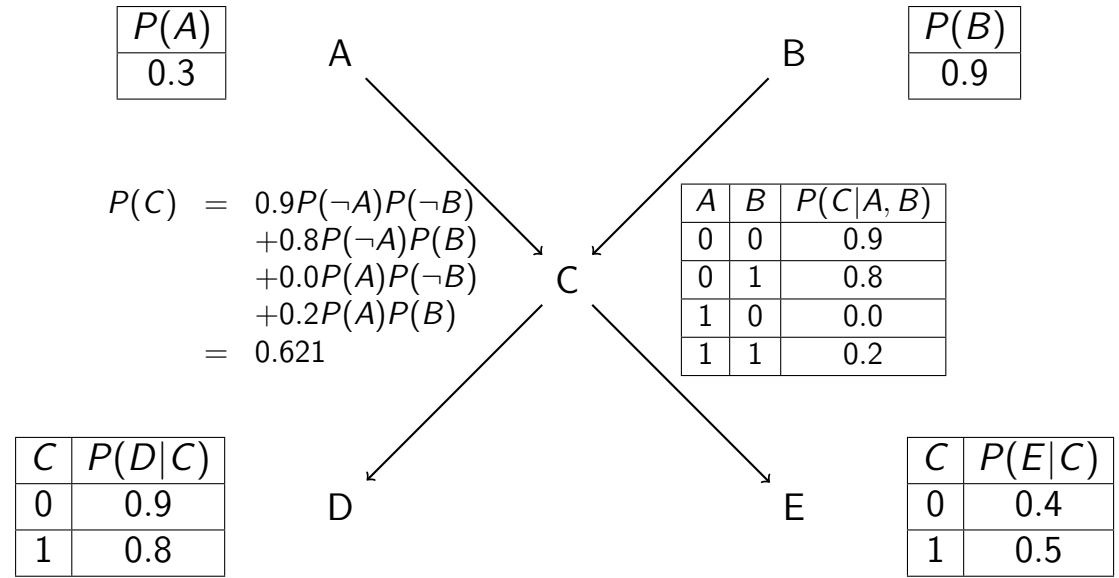


A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

C	$P(D\|C)$
0	0.9
1	0.8

C	$P(E\|C)$
0	0.4
1	0.5

Bayesian Networks: Example



$P(A)$
0.3

$P(B)$
0.9

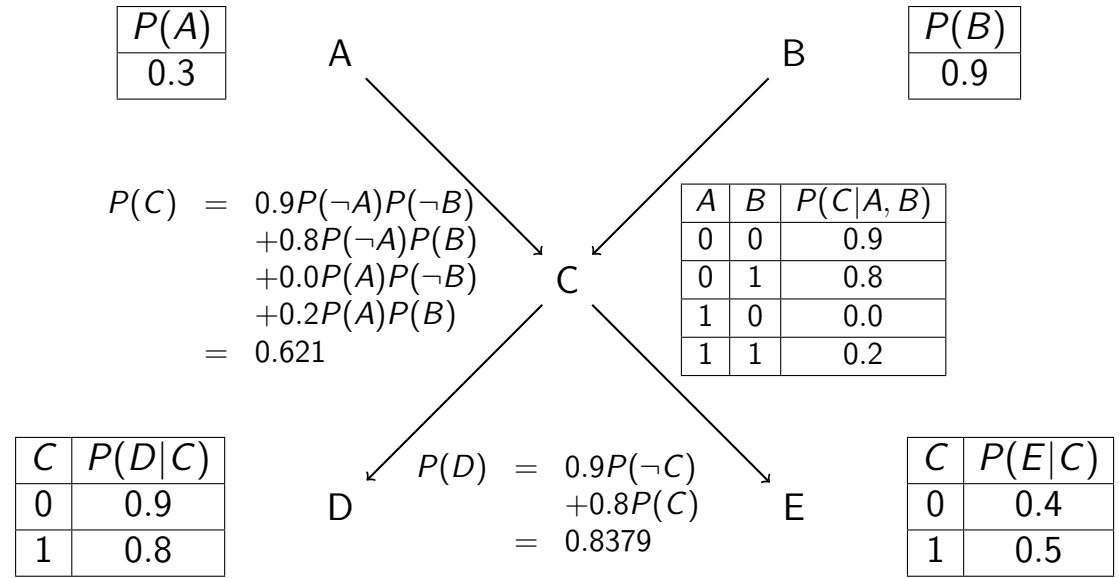
$$
\begin{aligned}
P(C) &= 0.9P(\neg A)P(\neg B) \\
&\quad +0.8P(\neg A)P(B) \\
&\quad +0.0P(A)P(\neg B) \\
&\quad +0.2P(A)P(B) \\
&= 0.621
\end{aligned}
$$

A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

C	$P(D\|C)$
0	0.9
1	0.8

C	$P(E\|C)$
0	0.4
1	0.5

Bayesian Networks: Example



$P(A)$
0.3

$P(B)$
0.9

$$
\begin{aligned}
P(C) &= 0.9P(\neg A)P(\neg B) \\
&\quad +0.8P(\neg A)P(B) \\
&\quad +0.0P(A)P(\neg B) \\
&\quad +0.2P(A)P(B) \\
&= 0.621
\end{aligned}
$$

A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

C	$P(D\|C)$
0	0.9
1	0.8

$$
\begin{aligned}
P(D) &= 0.9P(\neg C) \\
&\quad +0.8P(C) \\
&= 0.8379
\end{aligned}
$$

C	$P(E\|C)$
0	0.4
1	0.5

Notation

On the previous slide:

A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

really meant

A	B	C	$P(C\|A,B)$
0	0	0	0.1
0	0	1	0.9
0	1	0	0.2
0	1	1	0.8
1	0	0	1.0
1	0	1	0.0
1	1	0	0.8
1	1	1	0.2

Common notational short-cut: $P(x_1, \ldots, x_n)$ often really means $P(x_1 = b_1, \ldots, x_n = b_n)$ for particular values $b_1, \ldots, b_n$

Notation

On the previous slide:

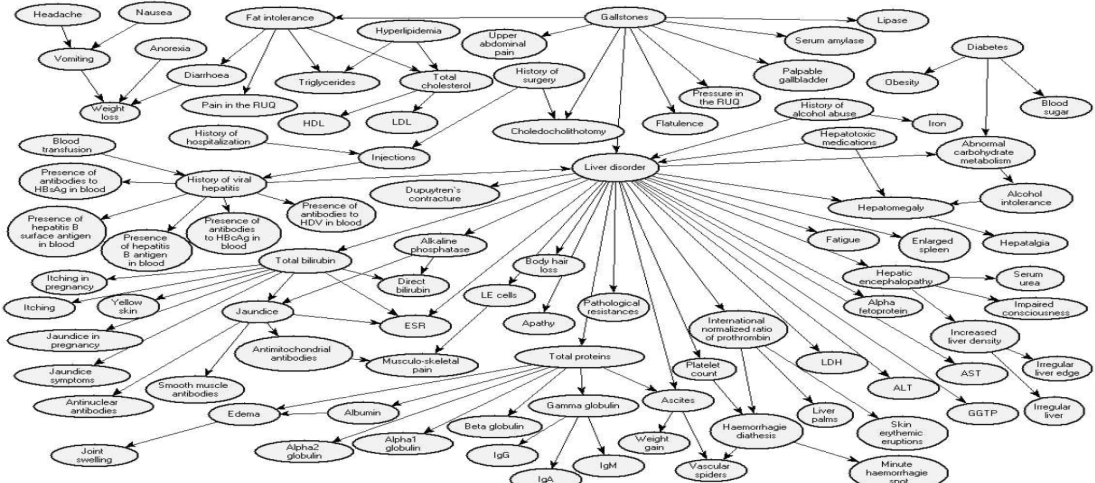
A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

really meant

A	B	C	$P(C\|A,B)$
0	0	0	0.1
0	0	1	0.9
0	1	0	0.2
0	1	1	0.8
1	0	0	1.0
1	0	1	0.0
1	1	0	0.8
1	1	1	0.2

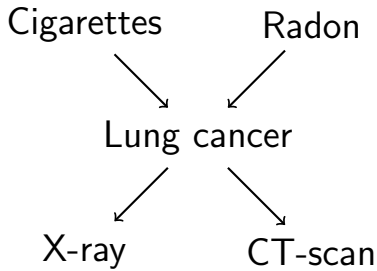
Common notational short-cut: $P(x_1, \ldots, x_n)$ often really means $P(x_1 = b_1, \ldots, x_n = b_n)$ for particular values $b_1, \ldots, b_n$

Bayesian Networks: Example



Bayesian network for liver disorder (Onisko et al. 1999)

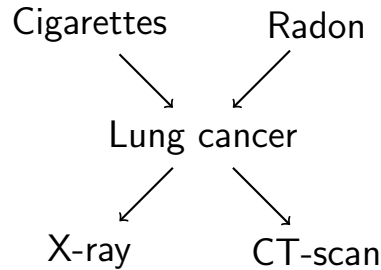
Bayesian Networks: Example



Diagnosis, with causes (known, unknown), evidence (known):

1. $P(Cancer|\neg Cigarettes \wedge \neg Radon)$
2. $P(Cancer|\neg XRay \wedge \neg Radon)$
3. $P(Cancer|\neg XRay \wedge \neg CTscan)$

Bayesian Networks: Example



Diagnosis, with causes (known, unknown), evidence (known):

1. $P(Cancer|\neg Cigarettes \wedge \neg Radon)$
2. $P(Cancer|\neg XRay \wedge \neg Radon)$
3. $P(Cancer|\neg XRay \wedge \neg CTscan)$

Independence and Conditional Independence

Independence

Variables X and Z are independent iff $P(X, Z) = P(X)P(Z)$.

Conditional independence

Variables X are conditionally independent of Z given Y iff $P(X|Y, Z) = P(X|Y)$. (Notation: $X \perp\!\!\!\perp Z \mid Y$)

Bayesian Networks: Explanation

Chain rule of probability

Let $x_1, \ldots, x_n$ be some ordering of the variables.

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

Example: $P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$

$$\begin{aligned}
P(x_1 \wedge x_2 \wedge x_3) &= P(x_1)P(x_2|x_1)P(x_3|x_1 \wedge x_2) \\
P(x_1 \wedge x_2 \wedge \neg x_3) &= P(x_1)P(x_2|x_1)P(\neg x_3|x_1 \wedge x_2) \\
P(x_1 \wedge \neg x_2 \wedge x_3) &= P(x_1)P(\neg x_2|x_1)P(x_3|x_1 \wedge \neg x_2) \\
P(x_1 \wedge \neg x_2 \wedge \neg x_3) &= P(x_1)P(\neg x_2|x_1)P(\neg x_3|x_1 \wedge \neg x_2) \\
P(\neg x_1 \wedge x_2 \wedge x_3) &= P(\neg x_1)P(x_2|\neg x_1)P(x_3|x_1 \wedge x_2) \\
P(\neg x_1 \wedge x_2 \wedge \neg x_3) &= P(\neg x_1)P(x_2|\neg x_1)P(\neg x_3|x_1 \wedge x_2) \\
P(\neg x_1 \wedge \neg x_2 \wedge x_3) &= P(\neg x_1)P(\neg x_2|\neg x_1)P(x_3|\neg x_1 \wedge \neg x_2) \\
P(\neg x_1 \wedge \neg x_2 \wedge \neg x_3) &= P(\neg x_1)P(\neg x_2|\neg x_1)P(\neg x_3|\neg x_1 \wedge \neg x_2)
\end{aligned}$$

Bayesian Networks: Explanation

Chain rule of probability

Let $x_1, \ldots, x_n$ be some ordering of the variables.

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

Example: $P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$

$$\begin{aligned}
P(x_1 \wedge x_2 \wedge x_3) &= P(x_1)P(x_2|x_1)P(x_3|x_1 \wedge x_2) \\
P(x_1 \wedge x_2 \wedge \neg x_3) &= P(x_1)P(x_2|x_1)P(\neg x_3|x_1 \wedge x_2) \\
P(x_1 \wedge \neg x_2 \wedge x_3) &= P(x_1)P(\neg x_2|x_1)P(x_3|x_1 \wedge \neg x_2) \\
P(x_1 \wedge \neg x_2 \wedge \neg x_3) &= P(x_1)P(\neg x_2|x_1)P(\neg x_3|x_1 \wedge \neg x_2) \\
P(\neg x_1 \wedge x_2 \wedge x_3) &= P(\neg x_1)P(x_2|\neg x_1)P(x_3|x_1 \wedge x_2) \\
P(\neg x_1 \wedge x_2 \wedge \neg x_3) &= P(\neg x_1)P(x_2|\neg x_1)P(\neg x_3|x_1 \wedge x_2) \\
P(\neg x_1 \wedge \neg x_2 \wedge x_3) &= P(\neg x_1)P(\neg x_2|\neg x_1)P(x_3|\neg x_1 \wedge \neg x_2) \\
P(\neg x_1 \wedge \neg x_2 \wedge \neg x_3) &= P(\neg x_1)P(\neg x_2|\neg x_1)P(\neg x_3|\neg x_1 \wedge \neg x_2)
\end{aligned}$$

Bayesian Networks: Explanation

Represent the joint distribution more compactly:

- Can represent every x_i in terms of $x_1, \ldots, x_{i-1}$
- But x_i often independent of some of $x_1, \ldots, x_{i-1}$
- And there are 2^{i-1} valuations of $x_1, \ldots, x_{i-1}$ (list all of them?)
- Choose $X_i' = \{x_1', \ldots, x_{n_i'}'\} \subseteq \{x_1, \ldots, x_{i-1}\}$ such that
  - $P(x_i|X_i') = P(x_i|x_1, \ldots, x_{i-1})$,
  - or equivalently $x_i \perp\!\!\!\perp (\{x_1, \ldots, x_{i-1}\} - X_i') \mid X_i'$.

  X_i' are the parents of x_i in the BN

BN a good representation if there are few dependencies!

Bayesian Networks: Explanation

Represent the joint distribution more compactly:

- Can represent every x_i in terms of $x_1, \ldots, x_{i-1}$
- But x_i often independent of some of $x_1, \ldots, x_{i-1}$
- And there are 2^{i-1} valuations of $x_1, \ldots, x_{i-1}$ (list all of them?)
- Choose $X_i' = \{x_1', \ldots, x_{n_i'}'\} \subseteq \{x_1, \ldots, x_{i-1}\}$ such that
  - $P(x_i|X_i') = P(x_i|x_1, \ldots, x_{i-1})$,
  - or equivalently $x_i \perp\!\!\!\perp (\{x_1, \ldots, x_{i-1}\} - X_i') \mid X_i'$.

  X_i' are the parents of x_i in the BN

BN a good representation if there are few dependencies!

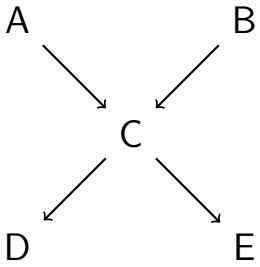
Bayesian Networks: Definition

Definition

A Bayesian network (BN) is a directed acyclic graph $G = \langle \{x_1, \ldots, x_n\}, A \rangle$ where

1. the nodes $x_1, \ldots, x_n$ are variables and
2. set of arcs $A \subseteq \{x_1, \ldots, x_n\}^2 = \{\langle x_i, x_j \rangle \mid 1 \leq i \leq n, 1 \leq j \leq n\}$.
3. Node x_i is associated with a conditional probability table, that maps values of n_i parent nodes to a probability : $\{0, 1\}^{n_i} \rightarrow \mathbb{R}$.

Bayesian Networks: Example



$P(A)$
0.3

$P(B)$
0.9

C	$P(D\|C)$
0	0.9
1	0.8

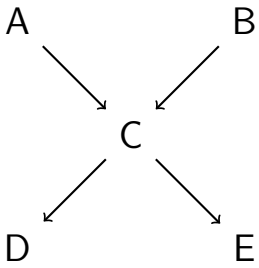
A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

C	$P(E\|C)$
0	0.4
1	0.5

$$P(A \wedge \neg B \wedge \neg C \wedge \neg D \wedge E) = 0.3 \cdot (1-0.9) \cdot (1-0.0) \cdot (1-0.9) \cdot 0.4 = 0.012$$

General fact: $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{Parents}(x_i))$

Bayesian Networks: Example



$P(A)$
0.3

$P(B)$
0.9

C	$P(D\|C)$
0	0.9
1	0.8

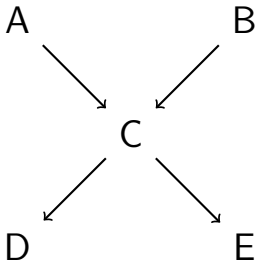
A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

C	$P(E\|C)$
0	0.4
1	0.5

$$P(A\wedge\neg B\wedge\neg C\wedge\neg D\wedge E) = 0.3\cdot(1-0.9)\cdot(1-0.0)\cdot(1-0.9)\cdot 0.4 = 0.012$$

General fact: $P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(x_i))$

Bayesian Networks: Example



$P(A)$
0.3

$P(B)$
0.9

C	$P(D\|C)$
0	0.9
1	0.8

A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

C	$P(E\|C)$
0	0.4
1	0.5

$$P(A\wedge\neg B\wedge\neg C\wedge\neg D\wedge E) = 0.3\cdot(1-0.9)\cdot(1-0.0)\cdot(1-0.9)\cdot 0.4 = 0.012$$

General fact: $P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(x_i))$

Bayesian Networks: Connection to Logic

- If all probabilities in a CPT for x are 0 or 1, it is expressible as $x \leftrightarrow \Phi(x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are the parents.
- In some applications, many probabilities are 0 or 1.
  0-1 reasoning with BN addressed by logic-based methods for BN

Constructing a Bayesian Network from Data

1. Structure is given: Construct the CPT for each variable separately
   - x with parents $x_1, \ldots, x_n$: calculate $P(x|x_1, ..., x_n)$
   - Count data items that satisfy x and $\neg x$ for given values for $x_1, \ldots, x_n$
   - Scales up reasonably well, close to linear time in the size of the BN
2. Structure is learned: Find structure & CPTs that best match data
   - NP-hard combinatorial problem
   - (Not discussed here)

Probabilistic Queries

Compute $P(X = x \mid E_1 = e_1, \ldots, E_m = e_m)$ for query variable X given values of evidence variables $E_1, \ldots, E_m$. Rest of the variables $Y_1, \ldots, Y_k$ are hidden variables.

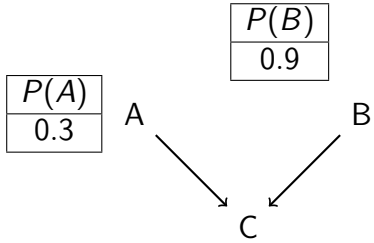
Problem is #P-complete (best algorithms are $O(2^n)$ time)
Several alternative methods:

1. Brute-force enumeration
2. Sampling methods
3. Weighted model-counting (propositional logic)

Inference by Enumeration: Procedure Outline

- Values of query and evidence variables are fixed
- Case analysis over possible values for hidden variables $\rightarrow$ tree search
- Go through all valuations (value combinations)
- Compute probability of each by $\prod_{x \in X} P(x|\text{Parents}(x))$

Inference by Enumeration: Illustration



A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

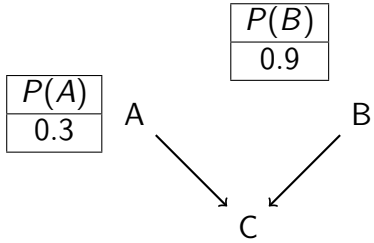
Compute $P(B|C)$ with A hidden:

$$
\begin{aligned}
P(B|C) &= \alpha P(B{=}1, C{=}1) \\
&= \alpha \sum_{a\in\{0,1\}} P(A{=}a, B{=}1, C{=}1) \\
&= \alpha \sum_{a\in\{0,1\}} P(A{=}a)P(B{=}1)P(C{=}1|A{=}a, B{=}1) \\
&= \begin{array}{l}\alpha(P(A{=}0)P(B{=}1)P(C{=}1|A{=}0, B{=}1) \\ \quad +P(A{=}1)P(B{=}1)P(C{=}1|A{=}1, B{=}1))\end{array} \\
&= \alpha(0.7{\cdot}0.9{\cdot}0.8{+}0.3{\cdot}0.9{\cdot}0.2) \\
&= \alpha{\cdot}0.558 \\
P(\neg B|C) &= \alpha P(B{=}0, C{=}1){=}\ldots \\
&= \alpha(0.7{\cdot}0.1{\cdot}0.9{+}0.3{\cdot}0.1{\cdot}0.0){=}\alpha{\cdot}0.063
\end{aligned}
$$

(Here $\alpha = \frac{1}{P(C)}$ is a corrective term ...)

Hence $P(B|C) = 0.899$, $P(\neg B|C) = 0.101$

Inference by Enumeration: Illustration



A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

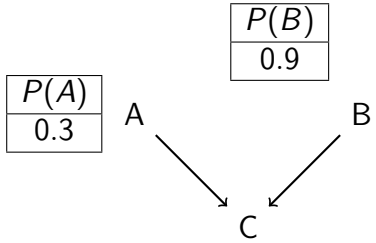
Compute $P(B|C)$ with A hidden:

$$
\begin{aligned}
P(B|C) &= \alpha P(B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a, B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a) P(B{=}1) P(C{=}1|A{=}a, B{=}1) \\
&= \begin{array}{l} \alpha (P(A{=}0) P(B{=}1) P(C{=}1|A{=}0, B{=}1) \\ \quad + P(A{=}1) P(B{=}1) P(C{=}1|A{=}1, B{=}1)) \end{array} \\
&= \alpha (0.7 \cdot 0.9 \cdot 0.8 + 0.3 \cdot 0.9 \cdot 0.2) \\
&= \alpha \cdot 0.558 \\
P(\neg B|C) &= \alpha P(B{=}0, C{=}1) = \ldots \\
&= \alpha (0.7 \cdot 0.1 \cdot 0.9 + 0.3 \cdot 0.1 \cdot 0.0) = \alpha \cdot 0.063
\end{aligned}
$$

(Here $\alpha = \frac{1}{P(C)}$ is a corrective term ...)

Hence $P(B|C) = 0.899$, $P(\neg B|C) = 0.101$

Inference by Enumeration: Illustration



A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

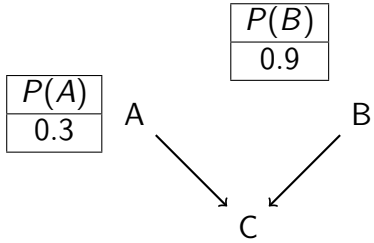
Compute $P(B|C)$ with A hidden:

$$
\begin{aligned}
P(B|C) &= \alpha P(B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a, B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a) P(B{=}1) P(C{=}1|A{=}a, B{=}1) \\
&= \alpha (P(A{=}0) P(B{=}1) P(C{=}1|A{=}0, B{=}1) \\
&\quad + P(A{=}1) P(B{=}1) P(C{=}1|A{=}1, B{=}1)) \\
&= \alpha (0.7 \cdot 0.9 \cdot 0.8 + 0.3 \cdot 0.9 \cdot 0.2) \\
&= \alpha \cdot 0.558 \\
P(\neg B|C) &= \alpha P(B{=}0, C{=}1) = \ldots \\
&= \alpha (0.7 \cdot 0.1 \cdot 0.9 + 0.3 \cdot 0.1 \cdot 0.0) = \alpha \cdot 0.063
\end{aligned}
$$

(Here $\alpha = \frac{1}{P(C)}$ is a corrective term ...)

Hence $P(B|C) = 0.899$, $P(\neg B|C) = 0.101$

Inference by Enumeration: Illustration



A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

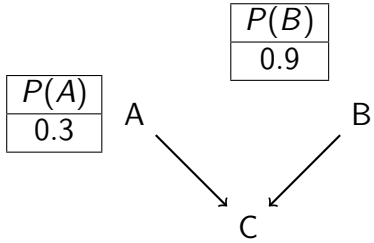
Compute $P(B|C)$ with A hidden:

$$
\begin{aligned}
P(B|C) &= \alpha P(B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a, B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a) P(B{=}1) P(C{=}1|A{=}a, B{=}1) \\
&= \alpha (P(A{=}0) P(B{=}1) P(C{=}1|A{=}0, B{=}1) \\
&\quad + P(A{=}1) P(B{=}1) P(C{=}1|A{=}1, B{=}1)) \\
&= \alpha (0.7 \cdot 0.9 \cdot 0.8 + 0.3 \cdot 0.9 \cdot 0.2) \\
&= \alpha \cdot 0.558 \\
P(\neg B|C) &= \alpha P(B{=}0, C{=}1) = \ldots \\
&= \alpha (0.7 \cdot 0.1 \cdot 0.9 + 0.3 \cdot 0.1 \cdot 0.0) = \alpha \cdot 0.063
\end{aligned}
$$

(Here $\alpha = \frac{1}{P(C)}$ is a corrective term ...)

Hence $P(B|C) = 0.899$, $P(\neg B|C) = 0.101$

Inference by Enumeration: Illustration



A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

Compute $P(B|C)$ with A hidden:

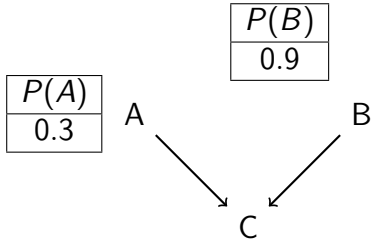
$$
\begin{aligned}
P(B|C) &= \alpha P(B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a, B{=}1, C{=}1) \\
&= \alpha \sum_{a \in \{0,1\}} P(A{=}a) P(B{=}1) P(C{=}1|A{=}a, B{=}1) \\
&= \begin{array}{l} \alpha (P(A{=}0) P(B{=}1) P(C{=}1|A{=}0, B{=}1) \\ \quad + P(A{=}1) P(B{=}1) P(C{=}1|A{=}1, B{=}1)) \end{array} \\
&= \alpha (0.7 \cdot 0.9 \cdot 0.8 + 0.3 \cdot 0.9 \cdot 0.2) \\
&= \alpha \cdot 0.558
\end{aligned}
$$

$$
\begin{aligned}
P(\neg B|C) &= \alpha P(B{=}0, C{=}1) = \ldots \\
&= \alpha (0.7 \cdot 0.1 \cdot 0.9 + 0.3 \cdot 0.1 \cdot 0.0) = \alpha \cdot 0.063
\end{aligned}
$$

(Here $\alpha = \frac{1}{P(C)}$ is a corrective term ...)

Hence $P(B|C) = 0.899$, $P(\neg B|C) = 0.101$

Inference by Enumeration: Illustration



A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

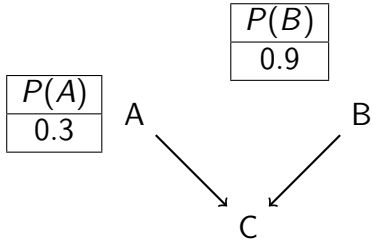
Compute $P(B|C)$ with A hidden:

$$
\begin{aligned}
P(B|C) &= \alpha P(B{=}1, C{=}1) \\
&= \alpha \sum_{a\in\{0,1\}} P(A{=}a, B{=}1, C{=}1) \\
&= \alpha \sum_{a\in\{0,1\}} P(A{=}a)P(B{=}1)P(C{=}1|A{=}a, B{=}1) \\
&= \begin{array}{l}\alpha(P(A{=}0)P(B{=}1)P(C{=}1|A{=}0, B{=}1) \\ \quad +P(A{=}1)P(B{=}1)P(C{=}1|A{=}1, B{=}1))\end{array} \\
&= \alpha(0.7{\cdot}0.9{\cdot}0.8{+}0.3{\cdot}0.9{\cdot}0.2) \\
&= \alpha{\cdot}0.558 \\
P(\neg B|C) &= \alpha P(B{=}0, C{=}1){=}\ldots \\
&= \alpha(0.7{\cdot}0.1{\cdot}0.9{+}0.3{\cdot}0.1{\cdot}0.0){=}\alpha{\cdot}0.063
\end{aligned}
$$

(Here $\alpha = \frac{1}{P(C)}$ is a corrective term ...)

Hence $P(B|C) = 0.899$, $P(\neg B|C) = 0.101$

Inference by Enumeration: Illustration



A	B	$P(C\|A,B)$
0	0	0.9
0	1	0.8
1	0	0.0
1	1	0.2

Compute $P(B|C)$ with A hidden:

$$
\begin{aligned}
P(B|C) &= \alpha P(B{=}1, C{=}1) \\
&= \alpha \sum_{a\in\{0,1\}} P(A{=}a, B{=}1, C{=}1) \\
&= \alpha \sum_{a\in\{0,1\}} P(A{=}a)P(B{=}1)P(C{=}1|A{=}a, B{=}1) \\
&= \begin{array}{l}\alpha(P(A{=}0)P(B{=}1)P(C{=}1|A{=}0, B{=}1) \\ \quad +P(A{=}1)P(B{=}1)P(C{=}1|A{=}1, B{=}1))\end{array} \\
&= \alpha(0.7{\cdot}0.9{\cdot}0.8{+}0.3{\cdot}0.9{\cdot}0.2) \\
&= \alpha{\cdot}0.558 \\
P(\neg B|C) &= \alpha P(B{=}0, C{=}1){=}\ldots \\
&= \alpha(0.7{\cdot}0.1{\cdot}0.9{+}0.3{\cdot}0.1{\cdot}0.0){=}\alpha{\cdot}0.063
\end{aligned}
$$

(Here $\alpha = \frac{1}{P(C)}$ is a corrective term ...)

Hence $P(B|C) = 0.899$, $P(\neg B|C) = 0.101$

Inference by Enumeration: Procedure

$X = (x_1, \ldots, x_n)$ is variables topologically sorted (parents first)
$v : X \to \{0, 1\}$ is partial mapping expressing fixed values

function probab(X,v)
if $X = ()$ **then return** 1.0;
let $(y_1, \ldots, y_k) = X$;
if $v(y_1)$ is defined
then return $P(y_1|\text{Parents}(y_1))\cdot\text{probab}((y_2, \ldots, y_k),v)$
else return $P(y_1|\text{Parents}(y_1))\cdot\text{probab}((y_2, \ldots, y_k),v \cup \{(y_1, 1)\})$
$+P(\neg y_1|\text{Parents}(y_1))\cdot\text{probab}((y_2, \ldots, y_k),v \cup \{(y_1, 0)\})$

This has a runtime that is exponential in the number of hidden variables
Much computation repeated unnecessarily
More clever methods avoid redundant computations

Inference by Sampling

Generate one sample:

1. Randomly assign values to root variables according to their CPT
2. Similarly assign values to other variables, once parents have values
3. Stop when all variables have values

The samples obtained are unbiased!

Inference:

1. Generate lots of samples
   - If $\phi \wedge \psi$ holds in a sampled valuation, $c_\phi = c_\phi + 1$
   - If ψ holds in a sampled valuation, $c_\psi = c_\psi + 1$
2. Compute result: $P(\phi|\psi) = \frac{P(\phi \wedge \psi)}{P(\psi)} = \frac{c_\phi}{c_\psi}$

Issue: With high-dimensional problems and many variables in ϕ and ψ, needs astronomic numbers of samples to get the counts c_ϕ and c_ψ high enough.

Curse of Dimensionality

Example

How likely is a Finnish plumber who likes *Star Trek* and playing golf to like tea? (how many relevant data points about this available?)

As the number of variables (dimensions) increases, the number of data points would have to increase exponentially to fill the space.

Consequence:

- Most value combinations don't appear in the data
- Needs huge amounts of data to determine $P(A|\phi)$ when ϕ has many variables
- Same issue with lots of statistical methods more generally

Curse of Dimensionality

Example

How likely is a Finnish plumber who likes *Star Trek* and playing golf to like tea? (how many relevant data points about this available?)

As the number of variables (dimensions) increases, the number of data points would have to increase exponentially to fill the space.

Consequence:

- Most value combinations don't appear in the data
- Needs huge amounts of data to determine $P(A|\phi)$ when ϕ has many variables
- Same issue with lots of statistical methods more generally

Inference by Model-Counting

- BN inference is #P-complete
- Weighted model-counting (WMC) is #P-complete
- Several effective methods for solving WMC

This suggests reduction from BN inference to WMC

Weighted Model-Counting

Definition

The model-counting problem is computing $|\{v : X \to \{0,1\} | v \models \phi\}|$ for given a propositional formula ϕ over X.

Definition

The weighted model-counting problem is computing $\sum\{W(v) | v \models \phi\}|$ where

- $W(v) = \prod_{l \in L}(v(l) \cdot w(l))$ and
- $L = X \cup \{\neg x | x \in X\}$,

for propositional formulas ϕ over X and weights $w : L \to \mathbb{R}$.

Inference by Weighted Model-Counting

BN with variables a, b, c, d, Parents$(c) = \{a, b\}$, Parents$(d) = \{c\}$

Atoms for each variable/value:

$$\begin{array}{cccc} \lambda_a & \lambda_{\neg a} & \lambda_b & \lambda_{\neg b} \\ \lambda_c & \lambda_{\neg c} & \lambda_d & \lambda_{\neg d} \end{array}$$

Atoms for every CPT entry:

$$\begin{array}{cccc} \theta_a & \theta_{\neg a} & \theta_b & \theta_{\neg b} \\ \theta_{c|ab} & \theta_{c|a\neg b} & \theta_{c|\neg ab} & \theta_{c|\neg a\neg b} \\ \theta_{\neg c|ab} & \theta_{\neg c|a\neg b} & \theta_{\neg c|\neg ab} & \theta_{\neg c|\neg a\neg b} \\ \theta_{d|c} & \theta_{d|\neg c} & \theta_{\neg d|c} & \theta_{\neg d|\neg c} \end{array}$$

Formulas: exactly1$(\lambda_x, \lambda_{\neg x})$ for every x, and...

$\lambda_a \leftrightarrow \theta_a$
$\lambda_c \wedge \lambda_a \wedge \lambda_b \leftrightarrow \theta_{c|ab}$
$\lambda_{\neg c} \wedge \lambda_a \wedge \lambda_b \leftrightarrow \theta_{\neg c|ab}$
$\lambda_d \wedge \lambda_c \leftrightarrow \theta_{d|c}$

$\lambda_{\neg a} \leftrightarrow \theta_{\neg a}$
$\lambda_c \wedge \lambda_a \wedge \lambda_{\neg b} \leftrightarrow \theta_{c|a\neg b}$
$\lambda_{\neg c} \wedge \lambda_a \wedge \lambda_{\neg b} \leftrightarrow \theta_{\neg c|a\neg b}$
$\lambda_d \wedge \lambda_{\neg c} \leftrightarrow \theta_{d|\neg c}$

$\lambda_b \leftrightarrow \theta_b$
$\lambda_c \wedge \lambda_{\neg a} \wedge \lambda_b \leftrightarrow \theta_{c|\neg ab}$
$\lambda_{\neg c} \wedge \lambda_{\neg a} \wedge \lambda_b \leftrightarrow \theta_{\neg c|\neg ab}$
$\lambda_{\neg d} \wedge \lambda_c \leftrightarrow \theta_{\neg d|c}$

$\lambda_{\neg b} \leftrightarrow \theta_{\neg b}$
$\lambda_c \wedge \lambda_{\neg a} \wedge \lambda_{\neg b} \leftrightarrow \theta_{c|\neg a\neg b}$
$\lambda_{\neg c} \wedge \lambda_{\neg a} \wedge \lambda_{\neg b} \leftrightarrow \theta_{\neg c|\neg a\neg b}$
$\lambda_{\neg d} \wedge \lambda_{\neg c} \leftrightarrow \theta_{\neg d|\neg c}$

All $\lambda_?$ have weight 1.0, and others come from CPTs, e.g. $w(\theta_{\neg a|b\neg c}) = P(\neg a|b, \neg c)$

Inference by Weighted Model-Counting

Call the previous formulas Σ and the given weights w.

Properties:

- Weighted model count W for $\Sigma \cup \{x_1, \ldots, x_n\}$ with weights w satisfies $P(x_1, \ldots, x_n) = W$
- $P(X|E)$ obtained by WMCing $P(X, E)$ and $P(E)$ and dividing

Nothing clever in the formulas/weights
All cleverness is in the model-counter algorithms!

Inference by Weighted Model-Counting

Call the previous formulas Σ and the given weights w.

Properties:

- Weighted model count W for $\Sigma \cup \{x_1, \ldots, x_n\}$ with weights w satisfies $P(x_1, \ldots, x_n) = W$
- $P(X|E)$ obtained by WMCing $P(X, E)$ and $P(E)$ and dividing

Nothing clever in the formulas/weights
All cleverness is in the model-counter algorithms!

Most Probable Explanation (MPE)

Definition

The Most Probable Explanation problem is finding the valuation with the highest probability among valuations that assign given values $E_1 = e_1, \ldots, E_n = e_n$ to evidence variables.

This problem can be reduced to the Weighted MAX-SAT problem, by using the formulas used in BN-inference-by-WMC

Weighted MAX-SAT Problem

Definition

The weighted MAX-SAT problem is finding a valuation v that maximizes $\sum_{x \in X} w(x)$ for given a propositional formula ϕ over X and a weight function $w : X \to \mathbb{R}$.

Usually the weights are on *clauses*, but this special case suffices to us

Example

$$\begin{aligned} \Phi &= \{a \vee b \vee c\} \\ w(a) &= 1 \\ w(b) &= 2 \\ w(c) &= 3 \end{aligned}$$

The valuation that maximizes the weight is $v(a) = v(b) = v(c) = 1$, with weight 6.

MPE by Weighted MAX-SAT

1. For a given BN, take the same formulas we used with BN inference
2. $w(\lambda_?) = 0$ for all $\lambda_?$, and $w(\theta_{X|Y}) = \log P(X|Y)$
3. Maximizing $\sum_{i=x}^{n} \log P(x_i|\text{Parents}(x_i))$ maximizes $\prod_{i=x}^{n} P(x_i|\text{Parents}(x_i))$

Note: Probabilities 0 need special handling because $\lim_{x \to 0} \log x = -\infty$

Conclusion

- Reasoning tasks and applications
  - Compute $P(X)$ or conditional probability $P(X|Y)$
  - Diagnosis
  - State estimation (determine state, based on observations)
  - Many more
- Bayesian networks one of graphical models of uncertain data
  - another type of graphical model: Markov networks
  - lots of other formalisms (Markov logic networks, ...)
- Probabilistic reasoning much different from logic, more complex
- However, some cases solvable by methods for logical reasoning