VNU-HUS MAT1206E/3508: Introduction to AI

Reasoning with Uncertainty

Hoàng Anh Đức

Bộ môn Tin học, Khoa Toán-Cơ-Tin học
Đại học KHTN, ĐHQG Hà Nội
hoanganhduc@hus.edu.vn

Textbook Basis

These lecture slides were constructed based primarily on the contents of:

- Wolfgang Ertel (2025). *Introduction to Artificial Intelligence*. 3rd. Springer. DOI: 10.1007/978-3-658-43102-0

Contents

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

Additional Materials

Prof. Ertel's Lectures at Ravensburg-Weingarten University in 2011

- https://youtu.be/IW-HIOPqgsk&t=4455 (Computing with Probabilities)
- https://youtu.be/wbbAA8og4D8 (Computing with Probabilities, The Principle of Maximum Entropy)
- https://youtu.be/MWAWjCUuDUs (The Maximum Entropy Method)
- https://youtu.be/sQLzN6zWosY (The Maximum Entropy Method, LEXMED)
- https://youtu.be/xfv8xIk1-x4 (LEXMED, Reasoning with Bayesian Networks)
- https://youtu.be/z-WrA1xbkdY (Reasoning with Bayesian Networks)
- https://youtu.be/gMjuL5vMo04 (Reasoning with Bayesian Networks)

3 Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Introduction

Recall: The Flying Penguin Example

1. Tweety is a penguin
2. Penguins are birds
3. Birds can fly

Formalized in PL1, the *knowledge base* KB is:

$$\textit{penguin}(\textit{tweety})$$
$$\textit{penguin}(x) \Rightarrow \textit{bird}(x)$$
$$\textit{bird}(x) \Rightarrow \textit{fly}(x)$$

- It can be derived (for example, by resolution): $\textit{fly}(\textit{tweety})$.
- If $\textit{penguin}(x) \Rightarrow \neg\textit{fly}(x)$ (= “Penguins cannot fly”) is added to the knowledge base KB, then $\neg\textit{fly}(\textit{tweety})$ can also be derived.

$\Rightarrow$ *The knowledge base is inconsistent* (both $\textit{fly}(\textit{tweety})$ and $\neg\textit{fly}(\textit{tweety})$ can be derived), and *the logic is monotonic* (adding knowledge cannot retract an earlier conclusion).

$\Rightarrow$ *Probabilistic Logic* is useful.

- Formalize the statement “Nearly all birds can fly” (e.g., “99% of all birds can fly”).
- Correctly carry out inferences on it.

Additional Materials
4 Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Introduction

Reasoning with uncertain or incomplete knowledge is important

- In everyday situations and also in many technical applications of AI, *heuristic processes are very important*.
  - Example: Use heuristic techniques when looking for a parking space in city traffic.
- *Heuristics alone are often not enough*, especially when *a quick decision is needed given incomplete knowledge*.

Additional Materials
5 Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Introduction

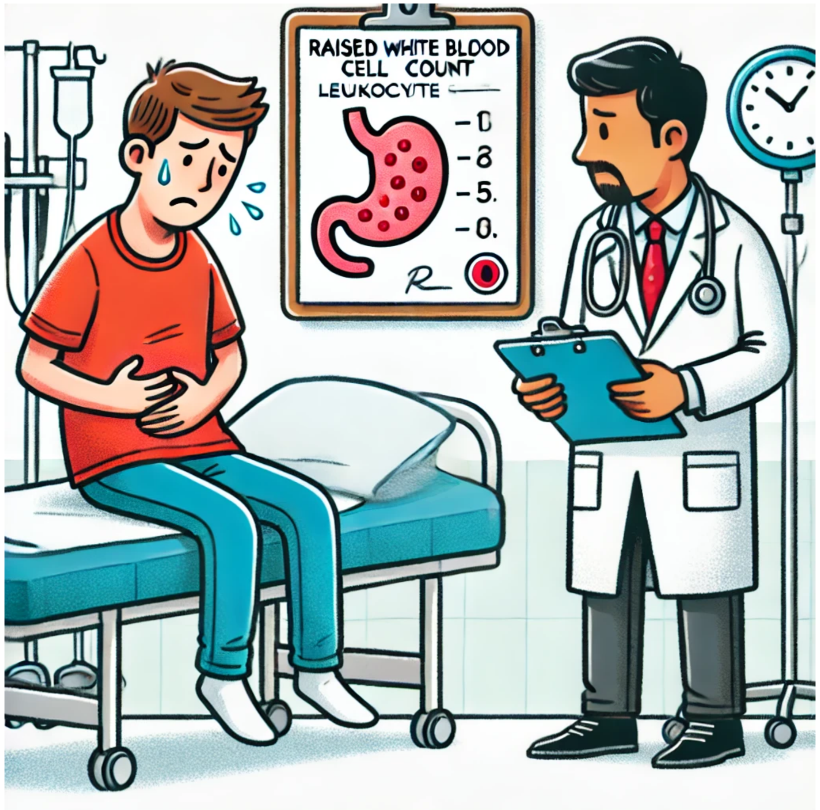
Example 1

If a patient experiences pain in the right lower abdomen and a raised white blood cell (leukocyte) count, this raises the suspicion that it might be appendicitis.

Stomach pain right lower $\wedge$ Leukocytes $> 10000 \rightarrow$ Appendicitis

If *Stomach pain right lower* $\wedge$ *Leukocytes* > 10000 *is true*, we can use Modus Ponens to derive Appendicitis

$\Rightarrow$ This model is *too coarse*.



Additional Materials
6 Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Introduction

Additional Materials
7 Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

When building their medical expert system MYCIN, Shortliffe and Buchanan [Shortliffe 1976] developed a calculus using so-called *certainty factors*, which *allowed the certainty of facts and rules to be represented*.

- Each *rule* $A \to B$ is assigned a *certainty factor* β.
- In MYCIN, β expresses how strongly the rule supports B when A holds; it is *not generally the conditional probability* $P(B \mid A)$.
  - Stomach pain right lower $\wedge$ Leukocytes $> 10000 \to_{0.6}$ Appendicitis
- Formulas for connecting the factors of rules
- Calculus is incorrect
- Inconsistent results could be derived

Introduction

- *Non-monotonic logic* and *default logic* represent *defeasible conclusions* that may be *withdrawn when an exception becomes known*.
  - Here, only *bird*$(x) \Rightarrow$ *fly*(x) is a *default*; the penguin fact and classification rule are *strict*.
  - The *strict exception* *penguin*$(x) \Rightarrow \neg$*fly*(x) *blocks the flying default for penguins*.
- These logics *do not by themselves express numerical degrees of belief*.
- *Dempster–Shafer theory:* assigns a *belief function* *Bel*(A) to a proposition A.
- *Fuzzy logic:* represents *degrees of truth or membership* for *vague concepts*, *rather than probabilities of well-defined propositions*.

Additional Materials

8 Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

References

Introduction

Reasoning with conditional probabilities

- *Conditional probabilities*
  - When probabilities describe an agent's beliefs, $P(B \mid A)$ is the agent's belief in B after learning A.
  - It does not by itself mean that A causes B.
- *Subjective probabilities*
  - A subjective probability represents an agent's degree of belief.
  - It depends on the information available to the agent.
  - Agents with different information may assign different probabilities.
  - All probability assignments must obey the probability axioms.
  - If the available information gives no reason to favor one possibility, we may model the possibilities as equally likely. *This is an assumption. It is not a conclusion from the evidence.*
- *Probabilities describe uncertainty*.
- Choosing an action also requires comparing its possible consequences or costs.
- *Probability theory* is well-founded.
- Methods for uncertain and incomplete knowledge:
  - Maximum entropy method (MaxEnt) and LEXMED.
  - Bayesian networks.

Additional Materials
9 Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Example 2

For a single roll of a fair (unbiased) die (experiment),

- The probability of the event “rolling a six” equals $1/6$.
- The probability of the occurrence “rolling an odd number” is equal to $1/2$.

Definition

- *Sample space Ω:* the finite set of *all possible outcomes* for an experiment.
- *Event:* subset of Ω.
  - If the outcome of an experiment is included in an event E, then event E has occurred.
  - A and B are events $\Rightarrow A \cup B$ is an event.
- *Elementary event:* subset of Ω containing exactly one element.
- *Sure event:* Ω.
- *Impossible event:* $\emptyset$.

Additional Materials

Introduction

10 Computing with Probabilities

Conditional Probability

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

References

Computing with Probabilities

Additional Materials
Introduction
11 Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

We will use *propositional logic notation for set operations*.

Set notation	Propositional logic	Description
$A \cap B$	$A \wedge B$	intersection / and
$A \cup B$	$A \vee B$	union / or
$\overline{A}$	$\neg A$	complement / negation
Ω	t	certain event / true
$\emptyset$	f	impossible event / false

- In the table, A and B denote *events*, i.e. *subsets of* Ω.
- A *random variable* is a function $X : \Omega \to \mathcal{X}$; it *assigns a value* $X(\omega)$ to each outcome ω. Here we consider *finite ranges* $\mathcal{X}$.
- For $S \subseteq \mathcal{X}$, $X \in S$ *denotes the event* $\{\omega \in \Omega \mid X(\omega) \in S\}$.
- For the fair die above, *face_number* has range $\{1, \ldots, 6\}$, and $P(\textit{face\_number} \in \{5, 6\}) = 1/3$.

Computing with Probabilities

Laplace Probability

Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be finite, and assume that its outcomes are *equally likely* (the *Laplace assumption*). Under this assumption, define, for every event $A \subseteq \Omega$,

$$P(A) := \frac{|A|}{|\Omega|} = \frac{\text{number of outcomes in } A}{\text{number of possible outcomes}}.$$

Example 3

For a fair die, the probability of an even number is

$$P(\textit{face\_number} \in \{2, 4, 6\}) = \frac{|\{2, 4, 6\}|}{|\{1, 2, 3, 4, 5, 6\}|} = \frac{3}{6} = \frac{1}{2}.$$

Additional Materials
Introduction
12 Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Computing with Probabilities

Additional Materials

Introduction

Computing with Probabilities

Conditional Probability

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

References

- *Consequence:* every *elementary event* $\{\omega\}$ has probability $1/|\Omega|$.
- The counting formulas $P(A) = |A|/|\Omega|$ and $P(A \mid B) = |A \wedge B|/|B|$ apply to a *finite sample space with equally likely outcomes*; *finiteness alone is not enough*. The general definition $P(A \mid B) = P(A \wedge B)/P(B)$ *needs no such equal-likelihood assumption*, but still requires $P(B) > 0$.
    - A *loaded die* has the same finite sample space, but its outcomes *need not be equally likely*.
- A condition such as *eye_color* $=$ *blue* *denotes an event*.
- For a *Boolean random variable* *JohnCalls*, the event *JohnCalls* $= t$ is abbreviated *JohnCalls*; hence $P($*JohnCalls*$)$ *abbreviates* $P($*JohnCalls* $= t)$.

Computing with Probabilities

Additional Materials

Introduction

Computing with Probabilities

Conditional Probability

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

References

Theorem 1

(1) $P(\Omega) = 1$.

(2) $P(\emptyset) = 0$, *which means that the impossible event has a probability of* 0.

(3) *For pairwise exclusive events* A *and* B*, it is true that* $P(A \vee B) = P(A) + P(B)$.

(4) *For two complementary events* A *and* $\neg A$*, it is true that* $P(A) + P(\neg A) = 1$.

(5) *For arbitrary events* A *and* B*, it is true that* $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$.

(6) *For* $A \subseteq B$*, it is true that* $P(A) \leq P(B)$.

(7) *If* $A_1, A_2, \ldots, A_n$ *are the elementary events, then* $\sum_{i=1}^{n} P(A_i) = 1$ *(normalization condition).*

Exercise 1 ([Ertel 2025], Exercise 7.1, p. 173)

Prove the propositions from Theorem 1.

Computing with Probabilities

Additional Materials

Introduction

15 Computing with Probabilities

Conditional Probability

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

References

For binary variables A, B,

- $P(A \wedge B) = P(A, B)$ stands for the probability of the event $A \wedge B$.
- *Distribution* or *joint probability distribution* $\mathbf{P}(A, B)$ of the variables A and B is the vector

$$(P(A, B), P(A, \neg B), P(\neg A, B), P(\neg A, \neg B))$$

- Distribution in matrix form

$\mathbf{P}(A, B)$	$B = t$	$B = f$
$A = t$	$P(A, B)$	$P(A, \neg B)$
$A = f$	$P(\neg A, B)$	$P(\neg A, \neg B)$

Computing with Probabilities

Additional Materials
Introduction
16 Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

In general,

- d variables $X_1, X_2, \ldots, X_d$ with n values each
- The distribution contains the values $P(X_1 = x_1, \ldots, X_d = x_d)$
- $x_1, \ldots, x_d$ each may have n different values
- The distribution can therefore be represented as a d-*dimensional matrix* with a total of n^d *elements*.
- By the *normalization condition*, one of these n^d values is *redundant*.
- Thus, the distribution is characterized by $n^d - 1$ *free parameters*.

Example 4

On Lansdowne Street in Boston, the speed of 100 vehicles is measured. For each measurement it is also noted whether the driver is a student. The results are

Event	Freq.	Relative freq.
Vehicle observed	100	1
Driver is a student (S)	30	0.3
Velocity (speed) too high (V)	10	0.1
Driver is a student and speeding ($S \wedge V$)	5	0.05

Do students speed more frequently than the average person, or than non-students?

Answer: *conditional probability*:

$$P(V \mid S) = \frac{5}{30} = \frac{1}{6} \approx 0.17,\ P(V) = 0.1,\ P(V \mid \neg S) = \frac{5}{70} = \frac{1}{14} \approx 0.071.$$

Thus $P(V \mid S) > P(V) > P(V \mid \neg S)$. Students speed more than the average driver and more than non-students.

Note: These figures describe the observed sample of 100 vehicles; generalizing requires that the sample be representative.

Additional Materials
Introduction
Computing with Probabilities
17 Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Additional Materials
Introduction
Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Definition

For two events A and B with $P(B) > 0$, the *conditional probability* of A given B is defined by

$$P(A \mid B) := \frac{P(A \wedge B)}{P(B)}.$$

If $P(B) = 0$, elementary conditional probability is *undefined*.

Under the *Laplace assumption*,

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

Additional Materials
Introduction
Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Definition

If $P(B) > 0$, events A and B are called *independent* if

$$P(A \mid B) = P(A).$$

If $P(B) = 0$, we define B to be independent of every event.

Theorem 2

For independent events A and B, $P(A \wedge B) = P(A)P(B)$.

Exercise 2

Prove Theorem 2.

- *Product Rule:*

$$\mathbf{P}(A, B) = \mathbf{P}(A \mid B)\mathbf{P}(B)$$

- *Chain Rule:* (Obtained by repeatedly applying the product rule.)

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \dots, X_{i-1}).$$

Additional Materials
Introduction
Computing with Probabilities
20 Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Example 5

- A roll of two dice.
- If *the two dice are independent*, the probability of rolling two sixes

$$P(D_1 = 6 \wedge D_2 = 6) = P(D_1 = 6) \cdot P(D_2 = 6) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

- If, by some magic power, *die 2 is always the same as die 1*,

$$P(D_1 = 6 \wedge D_2 = 6) = \frac{1}{6}.$$

Example 6

- For three events, the chain rule is

$$P(A_1, A_2, A_3) = P(A_3 \mid A_1, A_2) \cdot P(A_2 \mid A_1) \cdot P(A_1).$$

- We randomly draw 3 cards without replacement from deck with 52 cards.
- Event $A_i = \{$draw an ace in the i-th try$\}$.
- The probability that we have picked three aces

$$P(A_1, A_2, A_3) = P(A_3 \mid A_1, A_2) \cdot P(A_2 \mid A_1) \cdot P(A_1)$$
$$= \frac{2}{50} \cdot \frac{3}{51} \cdot \frac{4}{52} = \frac{1}{5525}.$$

Additional Materials
Introduction
Computing with Probabilities
21 Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Example 7

- Let X_1, X_2 be binary with $0 < P(X_i = t) < 1$ for $i = 1, 2$. Then

$$
\begin{aligned}
&P(X_1 = t, X_2 = t) = P(X_1, X_2) \\
&= P(X_1 = t \mid X_2 = t) \cdot P(X_2 = t) = P(X_1 \mid X_2) \cdot P(X_2), \\
&P(X_1 = t, X_2 = f) = P(X_1, \neg X_2) \\
&= P(X_1 = t \mid X_2 = f) \cdot P(X_2 = f) = P(X_1 \mid \neg X_2) \cdot P(\neg X_2).
\end{aligned}
$$

- Then, the chain rule for distribution $\mathbf{P}(X_1, X_2)$ reads

$$\mathbf{P}(X_1, X_2) = \mathbf{P}(X_2 \mid X_1) \cdot \mathbf{P}(X_1),$$

which means

$$\mathbf{P}(X_1, X_2) = \begin{pmatrix} P(X_1, X_2) \\ P(X_1, \neg X_2) \\ P(\neg X_1, X_2) \\ P(\neg X_1, \neg X_2) \end{pmatrix} = \begin{pmatrix} P(X_2 \mid X_1) \cdot P(X_1) \\ P(\neg X_2 \mid X_1) \cdot P(X_1) \\ P(X_2 \mid \neg X_1) \cdot P(\neg X_1) \\ P(\neg X_2 \mid \neg X_1) \cdot P(\neg X_1) \end{pmatrix}$$

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
22 Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

- Since $A \leftrightarrow (A \wedge B) \vee (A \wedge \neg B)$ is true for binary variables A and B, we also have

$$P(A) = P((A \wedge B) \vee (A \wedge \neg B))$$
$$= P(A \wedge B) + P(A \wedge \neg B).$$

$A \wedge B$ and $A \wedge \neg B$ are pairwise exclusive

- In general,

$$P(X_1 = x_1, \ldots, X_{d-1} = x_{d-1})$$
$$= \sum_{x_d} P(X_1 = x_1, \ldots, X_{d-1} = x_{d-1}, X_d = x_d)$$

The application of this formula is called *marginalization*.

- Marginalization can also be applied to distribution $\mathbf{P}(X_1, \ldots, X_d)$. The resulting distribution $\mathbf{P}(X_1, \ldots, X_{d-1})$ is called the *marginal distribution*.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

23 Conditional Probability

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

References

Additional Materials
Introduction
Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Example 8

Leuko Leukocyte value higher than 10000

App Patient has appendicitis (appendix inflammation)

$\mathbf{P}(App, Leuko)$	App	$\neg App$	Total
$Leuko$	0.23	0.31	0.54
$\neg Leuko$	0.05	0.41	0.46
Total	0.28	0.72	1

For example, it holds:

$$P(Leuko) = P(App, Leuko) + P(\neg App, Leuko) = 0.54$$

$$P(Leuko \mid App) = \frac{P(Leuko, App)}{P(App)} = \frac{0.23}{0.28} \approx 0.82.$$

Theorem 3 (Bayes' Theorem)

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

Exercise 3

Prove Theorem 3.

Example 9 (Appendicitis example)

$$P(\mathit{App} \mid \mathit{Leuko}) = \frac{P(\mathit{Leuko} \mid \mathit{App}) \cdot P(\mathit{App})}{P(\mathit{Leuko})} = \frac{0.82 \cdot 0.28}{0.54} \approx 0.43.$$

- $P(\mathit{Leuko} \mid \mathit{App})$: known appendicitis $\rightarrow$ chance of high leukocytes; often *similar in different places*.
- $P(\mathit{App} \mid \mathit{Leuko})$: known high leukocytes $\rightarrow$ chance of appendicitis; *depends on the local population* (via $P(\mathit{App})$ and $P(\mathit{Leuko})$).
  - Example: high leukocytes are more common where many people get infections (e.g. tropics).
- *Bayes' theorem*: from $P(\mathit{Leuko} \mid \mathit{App})$ plus local $P(\mathit{App})$ and $P(\mathit{Leuko})$ to $P(\mathit{App} \mid \mathit{Leuko})$ *for diagnosis*.

Additional Materials
Introduction
Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Example 10

- Sales representative: "Very reliable burglar alarm, reports any burglar with 99% certainty"
- A: Alarm, B: Burglar. The seller claims $P(A \mid B) = 0.99$
- Thus with high certainty: *If alarm then burglary!*
- *No! Be careful!*
- What does this mean when we hear the alarm go off?
  - Suppose we (the buyer) live in a relatively safe area in which break-ins are rare, with $P(B) = 0.001$ *on a given day*.
  - Assume that the alarm system is triggered not only by burglars, but also by animals, such as birds or cats in the yard, which results in $P(A) = 0.1$ *on a given day*.
  - Thus, $P(B \mid A) = (P(A \mid B) \cdot P(B))/P(A) \approx 0.01 \Rightarrow$ *There will be too many false alarms!*
- Additionally, we have $P(A) = P(A \mid B) \cdot P(B) + P(A \mid \neg B) \cdot P(\neg B) = 0.00099 + P(A \mid \neg B) \cdot 0.999 = 0.1$, which implies $P(A \mid \neg B) \approx 0.1 \Rightarrow$ The alarm will be triggered roughly every tenth day that there is not a break-in

Additional Materials
Introduction
Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

Exercise 4 ([Ertel 2025], Exercise 7.2, p. 173)

The gardening hobbyist Max wants to statistically analyze his yearly harvest of peas. For every pea pod he picks he measures its length x_i in centimeters and its weight y_i in grams. He divides the peas into two classes, the good and the bad (empty pods). The measured data (x_i, y_i) are

good pea:

x	1	2	2	3	3	4	4	5	6
y	2	3	4	4	5	5	6	6	6

bad pea:

x	4	6	6	7
y	2	2	3	3

(a) From the data, compute the probabilities $P(y > 3 \mid \textit{Class} = \textit{good})$ and $P(y \leq 3 \mid \textit{Class} = \textit{good})$. Then use Bayes' formula to determine $P(\textit{Class} = \textit{good} \mid y > 3)$ and $P(\textit{Class} = \textit{good} \mid y \leq 3)$.

(b) Which of the probabilities computed in subproblem (a) contradict the statement "All good peas are heavier than 3 grams"? (More than one may apply; briefly explain why each one you list contradicts the statement.)

The Principle of Maximum Entropy

- A *calculus for reasoning under uncertainty* can be realized using probability theory.
- Often too little knowledge for solving the necessary equations $\Rightarrow$ new ideas are needed.
- Idea from E.T. Jaynes (Physicist): Given missing knowledge, one can maximize the entropy of the desired probability distribution.
  - More precisely,
    - Take the precisely stated prior data or testable information about a probability distribution. *[What you already know.]*
    - Consider the set of all candidate probability distributions that satisfy those constraints. *[What are the possibilities given what you know?]*
    - Choose the distribution from this set that maximizes the (information) entropy. *[What is the least biased choice given what you know?]*
  - Intuition: MaxEnt picks the distribution that agrees with what you know and is otherwise as uniform as possible – it does not introduce any extra (unjustified) structure.
- Application to the LEXMED project.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

The Principle of Maximum Entropy

Let X be a discrete random variable with possible values $x_1, x_2, \ldots, x_n$ and probability distribution $\mathbf{P}(X) = (p_1, p_2, \ldots, p_n)$, where $p_i = P(X = x_i)$.

Definition

The *(information) entropy* H of the distribution $\mathbf{P}(X)$ is defined as $H(\mathbf{P}) = -\sum_{i=1}^{n} p_i \log p_i$ (with the convention $0 \log 0 := 0$).

- Entropy is a *measure of the uncertainty* associated with a random variable.
  - The *higher the entropy*, the *more uncertain or unpredictable the variable* is.
  - If one outcome has probability 1 and all others 0, then the entropy is 0 (no uncertainty).
  - If all outcomes are equally likely, then the entropy is maximized (maximum uncertainty).
- Entropy is measured in *nats* when using the natural logarithm ($\ln$) and in *bits* when using the base-2 logarithm ($\log_2$). (The choice of base for $\log$ depends on the context and application.)

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- Modus Ponens

$$\frac{A, A \Rightarrow B}{B}$$

- Generalization to probability rules

$$\frac{P(A) = \alpha, P(B \mid A) = \beta}{P(B) = ?}$$

Given: probabilities α, β, with $0 < \alpha < 1$; **Find:** $P(B)$.

- *Marginalization*

$$\begin{aligned} P(B) &= P(A, B) + P(\neg A, B) \\ &= P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A) \end{aligned}$$

The values of $P(A)$, $P(\neg A)$, and $P(B \mid A)$ are known. But *$P(B \mid \neg A)$ is unknown*.

- Since $0 \leq P(B \mid \neg A) \leq 1$, we can bound $P(B)$ by

$$P(B \mid A) \cdot P(A) \leq P(B) \leq P(B \mid A) \cdot P(A) + P(\neg A).$$

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
30 An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- Distribution

$$\mathbf{P}(A,B) = (P(A,B), P(A,\neg B), P(\neg A,B), P(\neg A,\neg B))$$

- Abbreviation

$$\begin{aligned} p_1 &= P(A,B) \\ p_2 &= P(A,\neg B) \\ p_3 &= P(\neg A,B) \\ p_4 &= P(\neg A,\neg B) \end{aligned}$$

- These four parameters (unknowns) $p_1, \ldots, p_4$ define the distribution.
- Out of it, any probability for A and B can be calculated.
- Four equations are required to calculate these unknowns.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

The Principle of Maximum Entropy
An Inference Rule for Probabilities

- Normalization condition: $p_1 + p_2 + p_3 + p_4 = 1$.
- From the given values $P(A) = \alpha$ and $P(B \mid A) = \beta$ we calculate

$$P(A, B) = P(B \mid A) \cdot P(A) = \alpha\beta$$
$$P(A) = P(A, B) + P(A, \neg B).$$

- So far, we have the following system of three equations

$$\begin{aligned} p_1 + p_2 + p_3 + p_4 &= 1 \\ p_1 &= \alpha\beta \\ p_1 + p_2 &= \alpha \end{aligned}$$

- Solve it as far as is possible, we get

$$\begin{aligned} p_1 &= \alpha\beta \\ p_2 &= \alpha(1-\beta) \\ p_3 + p_4 &= 1 - \alpha \end{aligned}$$

One equation is missing!

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

- To come to a definite solution despite this missing knowledge, we change our point of view. We use the given equations as constraints for an optimization problem.
- The full joint distribution is

$$\mathbf{p} = (p_1, p_2, p_3, p_4),$$

with $p_1 = \alpha\beta$ and $p_2 = \alpha(1-\beta)$ already fixed, and free variables $p_3, p_4 \geq 0$ under $p_3 + p_4 = 1 - \alpha$.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
33 An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- The *full entropy* is

$$H(\mathbf{p}) = -\sum_{i=1}^{4} p_i \ln p_i = -p_1 \ln p_1 - p_2 \ln p_2 - p_3 \ln p_3 - p_4 \ln p_4.$$

- Define the free part

$$H_{\text{free}}(p_3, p_4) := -p_3 \ln p_3 - p_4 \ln p_4.$$

  Then $H(\mathbf{p}) = H_{\text{free}}(p_3, p_4) + C$, where the constant $C = -p_1 \ln p_1 - p_2 \ln p_2$ does *not* depend on p_3, p_4.
- Therefore maximizing $H(\mathbf{p})$ over p_3, p_4 is the same as maximizing $H_{\text{free}}(p_3, p_4)$ under $p_3 + p_4 = 1 - \alpha$.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

Why should the entropy function be maximized?

- The *entropy* *measures the uncertainty of a distribution* up to a constant factor.
- *Negative entropy* is then a *measure of the amount of information a distribution contains, up to an additive constant*.
- *Maximizing the entropy minimizes the information content of the distribution*.
- Because *we are missing information about the distribution*, it must somehow be added in. We could fix an ad hoc value, for example $p_3 = 0.1$. Yet *it is better to determine the values p_3 and p_4 such that the information added is minimal*.

- For differentiable f and g with $\nabla g \neq 0$ on the constraint $g = 0$ (the partial derivatives of g are not all zero), find a *maximum or minimum* of $f(x_1, \ldots, x_n)$ by introducing a multiplier λ and forming

$$L(x_1, \ldots, x_n, \lambda) = f(x_1, \ldots, x_n) + \lambda g(x_1, \ldots, x_n).$$

- Solve

$$\frac{\partial L}{\partial x_i} = 0 \quad (i = 1, \ldots, n), \qquad \frac{\partial L}{\partial \lambda} = g(x_1, \ldots, x_n) = 0.$$

- Evaluate f at the solutions of the above system and at any relevant boundary points, then compare the values to determine the maximum or minimum. (For example, if $x_i \geq 0$, then evaluate at the boundary points with $x_i = 0$.)

[Strang and Herman 2016], Sec. 4.8

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- **Problem:** Maximizing $H_{\text{free}}(p_3, p_4) = -p_3 \ln p_3 - p_4 \ln p_4$ subject to $p_3 + p_4 - 1 + \alpha = 0$.
- Lagrange function:

$$\begin{aligned} L &= H_{\text{free}}(p_3, p_4) + \lambda(p_3 + p_4 - 1 + \alpha) \\ &= -p_3 \ln p_3 - p_4 \ln p_4 + \lambda(p_3 + p_4 - 1 + \alpha) \end{aligned}$$

- Taking the partial derivatives with respect to p_3 and p_4

$$\frac{\partial L}{\partial p_3} = -\ln p_3 - 1 + \lambda = 0, \qquad \frac{\partial L}{\partial p_4} = -\ln p_4 - 1 + \lambda = 0.$$

- These two equations along with the constraint give us a system of three equations and three unknowns p_3, p_4, λ. Solving it, we have $p_3 = p_4 = (1-\alpha)/2$. Evaluating H_{free} gives $(1-\alpha)(\ln 2 - \ln(1-\alpha))$ at this solution and $-(1-\alpha)\ln(1-\alpha)$ at the boundary points $(0, 1-\alpha)$ and $(1-\alpha, 0)$. The first is larger by $(1-\alpha)\ln 2$, so this is the *maximum*.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

The Principle of Maximum Entropy
An Inference Rule for Probabilities

$$P(B) = P(A,B) + P(\neg A, B) = p_1 + p_3 = \alpha\beta + \frac{1-\alpha}{2}$$
$$= \alpha(\beta - \frac{1}{2}) + \frac{1}{2} = P(A)(P(B \mid A) - \frac{1}{2}) + \frac{1}{2}.$$

- When $P(A) = 1$, $P(B) = P(B \mid A) = \beta$; in particular, $\beta = 1$ gives $P(B) = 1$ and $\beta = 0$ gives $P(B) = 0$.
- $P(B \mid A) = \beta$ requires $P(A) > 0$. If $P(A) = 0$, $P(B \mid A)$ is undefined; the graph's common left endpoint is only the limit $P(A) \to 0^+$, where the MaxEnt value tends to $1/2$.

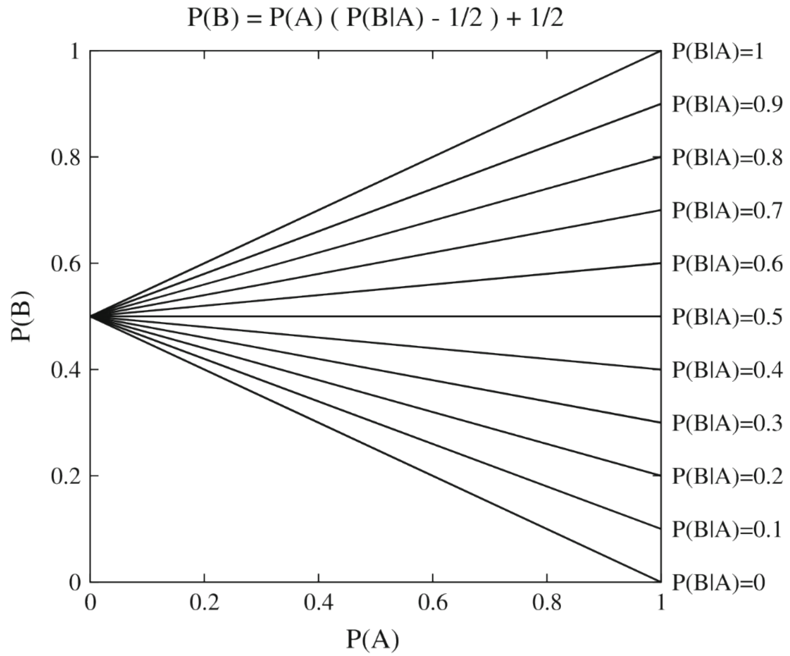


Figure: Curves for $0 < P(A) \leq 1$; the common left endpoint shows the limit as $P(A) \to 0^+$.

- When $P(B \mid A) = 1/2$, always $P(B) = 1/2$.

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

The Principle of Maximum Entropy

An Inference Rule for Probabilities

A *set of probabilistic equations* is called *consistent* if *there is at least one solution*, that is, one distribution which satisfies all equations.

Theorem 4

Let there be a consistent set of linear probabilistic equations. Then there exists a unique maximum for the entropy function with the given equations as constraints. The MaxEnt distribution thereby defined has minimum information content under the constraints.

- It follows from this theorem that *there is no distribution which satisfies the constraints and has higher entropy than the MaxEnt distribution*.
- Under the same stated constraints, choosing a lower-entropy distribution requires additional information or preferences not expressed by those constraints.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- p_3 and p_4 always occur symmetrically.
- Therefore, $p_3 = p_4$ (indifference).

Definition

If an arbitrary exchange of two or more variables in the Lagrange equations results in equivalent equations, these variables are called *indifferent*.

Theorem 5

If a set of variables $\{p_{i_1}, p_{i_2}, \ldots, p_{i_k}\}$ *is indifferent, then the maximum of the entropy under the given constraints is at the point where* $p_{i_1} = p_{i_2} = \cdots = p_{i_k}$.

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- No knowledge given $\Rightarrow$ All varibles are indifferent. (*Indifference Principle*.)
- No constraints beside the normalization condition $p_1 + p_2 + \cdots + p_n = 1$.
- We can set $p_1 = \cdots = p_n = \dfrac{1}{n}$.
- Given a complete lack of knowledge, all worlds are equally probable. That is, the distribution is uniform.

Example 11 (Special case: two variables A and B)

- $P(A, B) = P(A, \neg B) = P(\neg A, B) = P(\neg A, \neg B) = 1/4$.
- $P(A) = P(B) = 1/2$ and $P(B \mid A) = 1/2$.

As soon as the value of a condition deviates from the one derived from the uniform distribution, the probabilities of the worlds shift.

Example 12 (Special case: two variables A and B)

- Assume that only $P(B \mid A) = \beta$ is known; hence $P(A) = p_1 + p_2 > 0$.
- Then $P(A, B) = \beta P(A)$, so $p_1 = \beta(p_1 + p_2)$.
- We derived two constraints:

$$\beta p_2 + (\beta - 1)p_1 = 0$$
$$p_1 + p_2 + p_3 + p_4 - 1 = 0$$

- The Lagrange equations can no longer be solved symbolically so easily.
- Solving them numerically yields $p_3 = p_4$ (indifference of p_3 and p_4 in the constraints).

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

Exercise 5 ([Ertel 2025], Exercise 7.3, p. 174)

You are supposed to predict the afternoon weather using a few simple weather values from the morning of this day. The classical probability calculation for this requires a complete model, which is given in the following table.

Sky	*Bar*	*Prec*	$P(Sky, Bar, Prec)$
Clear	Rising	Dry	0.40
Clear	Rising	Raining	0.07
Clear	Falling	Dry	0.08
Clear	Falling	Raining	0.10
Cloudy	Rising	Dry	0.09
Cloudy	Rising	Raining	0.11
Cloudy	Falling	Dry	0.03

Sky: The sky is clear or cloudy in the morning
Bar: Barometer rising or falling in the morning
Prec: Raining or dry in the afternoon

(a) How many events are in the distribution for these three variables?

(b) Compute $P(\textit{Prec} = \textit{dry} \mid \textit{Sky} = \textit{clear}, \textit{Bar} = \textit{rising})$.

(c) Compute $P(\textit{Prec} = \textit{rain} \mid \textit{Sky} = \textit{cloudy})$.

(d) What would you do if the last row were missing from the table?

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
43 Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

Exercise 6 ([Ertel 2025], Exercise 7.4*, p. 174)

In a television quiz show, the contestant must choose between three closed doors. Behind one door the prize awaits: a car. Behind both of the other doors are goats. The contestant chooses a door, e.g. number one. The host, who knows where the car is, opens another door, e.g. number three, and a goat appears. The contestant is now given the opportunity to choose between the two remaining doors (one and two). What is the better choice from his point of view? To stay with the door he originally chose or to switch to the other closed door?

Exercise 7 ([Ertel 2025], Exercise 7.5, p. 174)

Using the Lagrange multiplier method, show that, without explicit constraints, the uniform distribution $p_1 = p_2 = \cdots = p_n = 1/n$ represents maximum entropy. Do not forget the implicitly ever-present constraint $p_1 + p_2 + \cdots + p_n = 1$. How can we show this same result using indifference?

We will now show that, *for modeling reasoning, conditional probability is better than what is known in logic as material implication*.

Classical logic

Case	A	B	$A \Rightarrow B$
1	t	t	t
2	t	f	f
3	f	t	t
4	f	f	t

Probabilities (matching cases)

Case	$P(A)$	$P(B)$	$P(B \mid A)$
1	1	1	1
2	1	0	0
3	0	1	Undefined
4	0	0	Undefined

Matching case numbers connect each truth assignment on the left to the probabilities it determines on the right.

Question: What value does $P(B \mid A)$ have, if only $P(A) = \alpha$ and $P(B) = \gamma$ are given?

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- $p_1 = P(A, B)$, $p_2 = P(A, \neg B)$, $p_3 = P(\neg A, B)$, $p_4 = P(\neg A, \neg B)$.
- Constraints:

$$p_1 + p_2 = \alpha$$
$$p_1 + p_3 = \gamma$$
$$p_1 + p_2 + p_3 + p_4 = 1$$

- Again we maximize entropy under the given constraints and obtain:

$$p_1 = \alpha\gamma, \quad p_2 = \alpha(1-\gamma), \quad p_3 = \gamma(1-\alpha), \quad p_4 = (1-\alpha)(1-\gamma)$$

- From $p_1 = \alpha\gamma$, we have $P(A, B) = P(A) \cdot P(B)$, which means A and B are independent.

Exercise 8 ([Ertel 2025], Exercise 7.8*, p. 175)

Prove that the above-mentioned results for p_1, p_2, p_3, p_4 are correct.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- From definition, $P(B \mid A) = \dfrac{P(A, B)}{P(A)}$.
- In the MaxEnt distribution above, A and B are independent. Hence, if $P(A) = \alpha > 0$, then $P(B \mid A) = P(B) = \gamma$.
- If $P(A) = 0$, then $P(B \mid A)$ is undefined.

Classical logic

Condition on A	$A \Rightarrow B$
$A = t$	B
$A = f$	t

MaxEnt probability: $P(B) = \gamma$

Condition on $P(A)$	$P(B \mid A)$
$P(A) = \alpha > 0$	γ
$P(A) = 0$	Undefined

Both tables concern the premise A, but a condition on its *truth value* ($A = t$ or $A = f$) and a condition on its *probability* ($P(A) > 0$ or $P(A) = 0$) are different kinds of conditions.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

47 Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

The Principle of Maximum Entropy

MaxEnt-Systems

- Often, MaxEnt optimization has no symbolic solution.
- Therefore: *numerical entropy maximization*.
- *SPIRIT-Shell* (Symmetrical Probabilistic Intensional Reasoning in Inference Networks in Transition) was developed at Fernuniversität Hagen; the university still lists the Java software and a free trial, but its linked download page is currently unavailable.
- *PIT* (Probability Induction Tool) was developed at the Technical University of Munich. Its former public web service is unavailable, so the syntax and outputs below are *historical examples*.
- *PIT* used Sequential Quadratic Programming (SQP) to find an extremum of the entropy function numerically.
- As input, PIT expected a file containing the constraints:

```
var A{t,f}, B{t,f};
P([A=t]) = 0.6;
P([B=t] | [A=t]) = 0.3;
QP([B=t]);
QP([B=t] | [A=t]);
```

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

- Example query: `QP([B=t])`
- Former web front end: `www.pit-systems.de` (inactive)
- Reported result

Nr.	Truth value	Probability	Query
1	UNSPECIFIED	3.800e-01	`QP([B=t]);`
2	UNSPECIFIED	3.000e-01	`QP([A=t]-|>[B=t]);`

- $P(B) = 0.38$ and $P(B \mid A) = 0.3$.
- In the PIT output, `QP([A=t]-|>[B=t])` is an *alternative form* of `QP([B=t] | [A=t])`; *both denote* $P(B \mid A)$ [Ertel 2025], p. 147, footnote 8.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

Exercise 9 ([Ertel 2025], Exercise 7.6, p. 174)

Use the PIT system (`http://www.pit-systems.de`) or SPIRIT (`http://www.xspirit.de`) to calculate the MaxEnt solution for $P(B)$ under the constraint $P(A) = \alpha$ and $P(B \mid A) = \beta$. Which disadvantage of PIT, compared with calculation by hand, do you notice here?

Exercise 10 ([Ertel 2025], Exercise 7.7, p. 175)

Given the constraints $P(A) = \alpha$ and $P(A \vee B) = \beta$, manually calculate $P(B)$ using the MaxEnt method. Use PIT to check your solution.

Current availability

The wording above is reproduced from Ertel. The former public PIT service is unavailable. Fernuniversität Hagen still lists a free SPIRIT-Shell Java trial, but its linked download page is currently unavailable. If necessary, use another numerical MaxEnt solver.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

- Recall the flying penguin example from the introduction:
  - Knowledge base: “Tweety is a penguin”, “penguins are birds”, and “birds can fly”
  - One can logically derive: $\mathit{fly}(\mathit{tweety})$.
  - Adding “penguins cannot fly” made the knowledge base *inconsistent* ($\neg\mathit{fly}(\mathit{tweety})$ can be derived as well).
  - The *logic* is *monotonic*: adding knowledge cannot retract the earlier conclusion.
- In the probabilistic formulation below, we will ask the generic query $P(\mathit{flies} \mid \mathit{penguin})$ rather than name Tweety explicitly. We encode the three general statements about penguins, birds, and flying as probabilistic constraints.
- Two things will follow: the *probabilistic* model *stays consistent*, and its answer changes once the exception “penguins cannot fly” is added. The probabilistic inference is *non-monotonic*.

$$
\begin{array}{rl}
P(\textit{bird} \mid \textit{penguin}) = 1 & \text{“penguins are birds”} \\
P(\textit{flies} \mid \textit{bird}) \in [0.95, 1] & \text{“(almost all) birds can fly”} \\
P(\textit{flies} \mid \textit{penguin}) = 0 & \text{“penguins cannot fly”}
\end{array}
$$

- Example input in the historical PIT syntax:

```
var penguin{yes,no}, bird{yes,no}, flies{yes,no}

P([bird=yes] | [penguin=yes]) = 1;
P([flies=yes] | [bird=yes]) IN [0.95,1];
P([flies=yes] | [penguin=yes]) = 0;

QP([flies=yes]| [penguin=yes]);
```

- Reported answer

Nr.	Truth value	Probability	Query
1	UNSPECIFIED	0.000e+00	`QP([penguin=yes]-|>[flies=yes]);`

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

52 The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

The Principle of Maximum Entropy

The Tweety Example: Non-Monotonic Inference

- The example demonstrates non-monotonic inference only when we compare the answers obtained *before and after the exception* $P(\textit{flies} \mid \textit{penguin}) = 0$ *is entered (in PIT syntax* `P([flies=yes] | [penguin=yes]) = 0;`*)*.
- Before the exception:
  - Only $P(\textit{bird} \mid \textit{penguin}) = 1$ and $P(\textit{flies} \mid \textit{bird}) \in [0.95, 1]$ are given.
  - Nothing links penguins to flying except through birds, so MaxEnt gives $P(\textit{flies} \mid \textit{penguin}) = P(\textit{flies} \mid \textit{bird})$. Without the interval, indifference between flying and not flying would give probability $1/2$.
  - The constraint $[0.95, 1]$ excludes $1/2$, and increasing the probability above 0.95 makes the outcomes less balanced and lowers entropy.
  - Hence MaxEnt selects the boundary value $P(\textit{flies} \mid \textit{penguin}) = 0.95$.
- Entering the exception and recalculating gives $P(\textit{flies} \mid \textit{penguin}) = 0$, the answer reported on the previous slide.
- Adding knowledge lowers the answer from 0.95 to 0, so *an earlier conclusion is withdrawn*. The specific rule about penguins overrides the general rule about birds; this is *non-monotonic inference*.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- Consistency requires only that the rule *allow non-flying birds*; the point constraint $P(\textit{flies} \mid \textit{bird}) = 0.95$ would also work. Setting this probability to 1 would make the three constraints *inconsistent*, as in the logical version [Ertel 2025], p. 147. The *interval* represents *uncertainty about the exact probability*.
- In the Tweety example,

$$P(\textit{flies} \mid \textit{bird}) \in [0.95, 1]$$

  expresses the strong statement "almost all birds fly": *at least* 95% *of birds fly*.
- A weaker interpretation of "normally birds fly" is simply that a bird is *more likely to fly than not to fly*:

$$P(\textit{flies} \mid \textit{bird}) > 0.5.$$

  The value 0.5 is excluded because, at exactly 0.5, flying and not flying are equally likely, so neither outcome is the normal one.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
The Tweety Example
LEXMED
Reasoning with Bayesian Networks
Summary
References

- The *strict condition* $P(\textit{flies} \mid \textit{bird}) > 0.5$ expresses the intended meaning: a bird is *more likely to fly than not to fly*. For a MaxEnt computation, however, the *open lower boundary* may leave no distribution that *attains the largest possible entropy*; feasible distributions may only approach the best value as the probability approaches 0.5.
- To avoid this problem, choose a *fixed margin* $0 < \varepsilon < 0.5$ and use the *closed constraint*

$$p := P(\textit{flies} \mid \textit{bird}) \in [0.5 + \varepsilon, 1],$$
$$p - P(\neg\textit{flies} \mid \textit{bird}) \geq 2\varepsilon.$$

  The second line follows because $P(\neg\textit{flies} \mid \textit{bird}) = 1 - p$. Thus ε is the *required margin* by which $P(\textit{flies} \mid \textit{bird})$ exceeds 0.5, while 2ε is the *minimum difference* $P(\textit{flies} \mid \textit{bird}) - P(\neg\textit{flies} \mid \textit{bird})$. For $\varepsilon = 0.01$, $P(\textit{flies} \mid \textit{bird}) \geq 0.51$ and $P(\neg\textit{flies} \mid \textit{bird}) \leq 0.49$, so this conditional-probability difference is *at least* 0.02.
- The next section applies MaxEnt to the medical expert system *LEXMED*.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

An Inference Rule for Probabilities

Maximum Entropy Without Explicit Constraints

Conditional Probability Versus Material Implication

MaxEnt-Systems

The Tweety Example

LEXMED

Reasoning with Bayesian Networks

Summary

References

LEXMED

- Manfred Schramm, Walter Rampf, Wolfgang Ertel
- Ravensburg-Weingarten University of Applied Sciences + Weingarten 14-Nothelfer Hospital + Technical University Munich
- LEXMED = Medical expert system capable of learning.

The project was funded by the german state of Baden-Wuerttemberg, the AOK Baden-Württemberg, the Ravensburg-Weingarten University of Applied Sciences and by the hospital 14 Nothelfer in Weingarten.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

56 LEXMED

Appendicitis Diagnosis with Formal Methods

Hybrid Probabilistic Knowledge Base

Application of LEXMED

Function of LEXMED

Risk Management Using the Cost Matrix

Performance

Application Areas and Experiences

Reasoning with Bayesian Networks

Summary

References

LEXMED

Personal Details	unknown	values	
Gender	⊙	○ male ○ female	?
Age-group	⊙	○ 0-5 ○ 6-10 ○ 11-15 ○ 16-20 ○ 21-25 ○ 26-35 ○ 36-45 ○ 46-55 ○ 56-65 ○ 65-	?
Results of examination	**not done**	**values**	
1st quadrant	⊙	○ yes ○ no	?
2nd quadrant	⊙	○ yes ○ no	?
3rd quadrant	⊙	○ yes ○ no	?
4th quadrant	⊙	○ yes ○ no	?
guarding	⊙	○ local ○ global ○ none	?
rebound tenderness	⊙	○ yes ○ no	?
pain on tapping	⊙	○ yes ○ no	?
rectal pain	⊙	○ yes ○ no	?
bowel sounds	⊙	○ weak ○ normal ○ increased ○ none	?
abnormal ultrasound	⊙	○ yes ○ no	?
abnormal urine sediment	⊙	○ yes ○ no	?
temperature range (rectal)	⊙	○ -37.3 ○ 37.4-37.6 ○ 37.7-38.0 ○ 38.1-38.4 ○ 38.5-38.9 ○ 39.0-	?
leucocyte count	⊙	○ 0-6k ○ 6k-8k ○ 8k-10k ○ 10k-12k ○ 12k-15k ○ 15k-20k ○ 20k-	?

Abfragen

Figure: LEXMED Query form.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
57 LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

LEXMED

	Result of the PIT diagnosis			
Diagnosis	App. inflamed	App. perforated	Negative	Other
Probability	0.70	0.17	0.06	0.07

Figure: LEXMED Answer.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
58 LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

- The most common serious cause of acute abdominal pain is appendicitis.
- Even today, *diagnosis can be difficult in many cases*.
    - Up to about 20% of removed appendices showed no pathological evidence of appendicitis
    - There are also cases, where an inflamed appendix is not recognized
- In 1972, *de Dombal (Great Britain)* developed *an expert system for the diagnosis of acute stomach pain*.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
59 Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Appendicitis Diagnosis with Formal Methods

Hybrid Probabilistic Knowledge Base

Application of LEXMED

Function of LEXMED

Risk Management Using the Cost Matrix

Performance

Application Areas and Experiences

Reasoning with Bayesian Networks

Summary

References

Nearly all of the formal diagnostic processes used in medicine to date have been based on *scores*.

- *For each value of a symptom* (for example *fever* or *lower right stomach pain*) the doctor notes a certain number of *points*.
- If *the sum of the points is over a certain value (threshold)*, a certain *decision is recommended* (for example operation).

With n symptoms $S_1, S_2, \ldots, S_n$, a score can formally be defined as

$$\text{Diagnose} = \begin{cases} \text{Appendicitis} & \text{if } w_1 S_1 + \cdots + w_n S_n > \Theta, \\ \text{negative} & \text{else.} \end{cases}$$

- Scores are too weak for the modelling of complex relations.
- Score systems cannot consider "contexts".
  - E.g. they cannot distinguish between the leukocyte values of elderly and medium age people.
- They demand high requirements on databases (representative).

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Query to the expert system: What is the probability of an inflamed appendix if the patient is a 23-year-old man with pain in the right lower abdomen and a leukocyte count of $13{,}000$?

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
Age	*0–5, 6–10, 11–15, 16–20, 21–25, 26–35, 36–45, 46–55, 56–65, 65–*	10	*Age10*
Pain 1st Quad.	*Yes, no*	2	*P1Q2*
Pain 2nd Quad.	*Yes, no*	2	*P2Q2*
Pain 3rd Quad.	*Yes, no*	2	*P3Q2*
Pain 4th Quad.	*Yes, no*	2	*P4Q2*
Guarding	*Local, global, none*	3	*Gua3*
Rebound tenderness	*Yes, no*	2	*Reb2*
Pain on tapping	*Yes, no*	2	*Tapp2*
Rectal pain	*Yes, no*	2	*RecP2*
Bowel sounds	*Weak, normal, increased, none*	4	*BowS4*
Abnormal ultrasound	*Yes, no*	2	*Sono2*
Abnormal urine sedim.	*Yes, no*	2	*Urin2*
Temperature (rectal)	*–37.3, 37.4–37.6, 37.7–38.0, 38.1–38.4, 38.5–38.9, 39.0–*	6	*TRec6*
Leukocytes	*0–6k, 6k–8k, 8k–10k, 10k–12k, 12k–15k, 15k–20k, 20k–*	7	*Leuko7*
Diagnosis	*Inflamed, perforated, negative, other*	4	*Diag4*

Figure: Symptoms used for the query in LEXMED and their values. The number of values for the each symptom is given in the column marked #.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Appendicitis Diagnosis with Formal Methods

Hybrid Probabilistic Knowledge Base

Application of LEXMED

Function of LEXMED

Risk Management Using the Cost Matrix

Performance

Application Areas and Experiences

Reasoning with Bayesian Networks

Summary

References

Query to the expert system:

$$P(\textit{Diag4} = \textit{inflamed} \vee \textit{Diag4} = \textit{perforated} \mid$$
$$\textit{Sex2} = \textit{male} \wedge \textit{Age10} \in 21\text{–}25 \wedge \textit{P3Q2} = \textit{yes}$$
$$\wedge\ \textit{Leuko7} \in 12\text{–}15k)$$

Knowledge Base:

Database

15000 patients from
Baden-Württemberg 1995

Expert knowledge

Dr. Rampf

Dr. Hontschik

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

- LEXMED calculates the probabilities of various diagnoses using the probability distribution of all relevant variables (see the previous table).
- The full joint table has one entry for each possible assignment:

$$2^{10} \cdot 10 \cdot 3 \cdot 4 \cdot 6 \cdot 7 \cdot 4 = 20\,643\,840.$$

- Normalization leaves $20\,643\,839$ free parameters.
- Any rule set with less than $20\,643\,839$ probability values may not specify the full joint distribution completely.
- A *complete distribution* is required.
- A human expert can not deliver $20\,643\,839$ values!
- Use *MaxEnt method*. The generalization of about 500 rules to a complete probability model is done in LEXMED by maximizing the entropy with the 500 rules as constraints.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Appendicitis Diagnosis with Formal Methods

Hybrid Probabilistic Knowledge Base

Application of LEXMED

Function of LEXMED

Risk Management Using the Cost Matrix

Performance

Application Areas and Experiences

Reasoning with Bayesian Networks

Summary

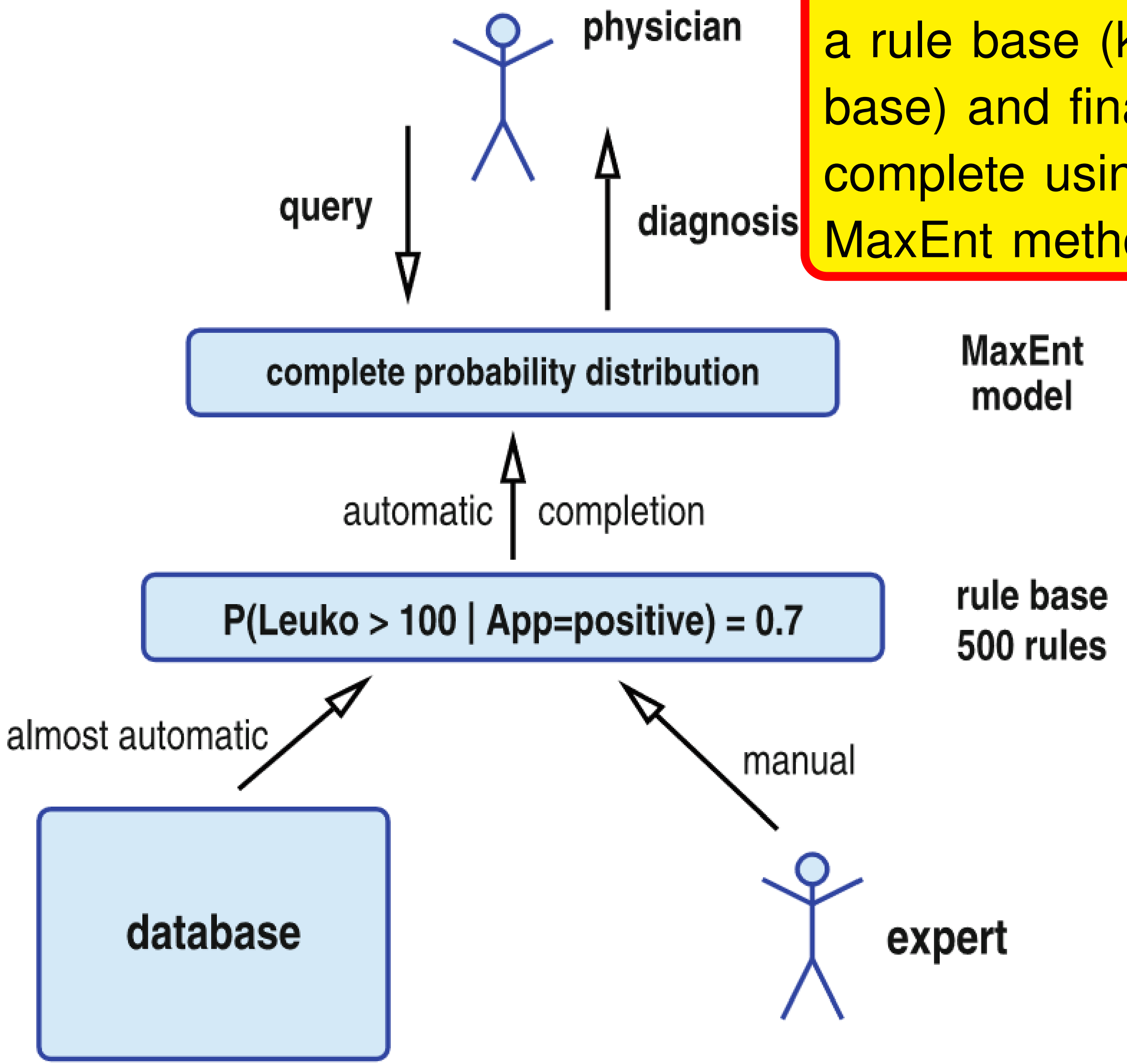
References

LEXMED

Hybrid Probabilistic Knowledge Base

Probabilistic rules are generated from data and expert knowledge, which are integrated in a rule base (knowledge base) and finally made complete using the MaxEnt method.

Knowledge Processing:



Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

System Architecture:

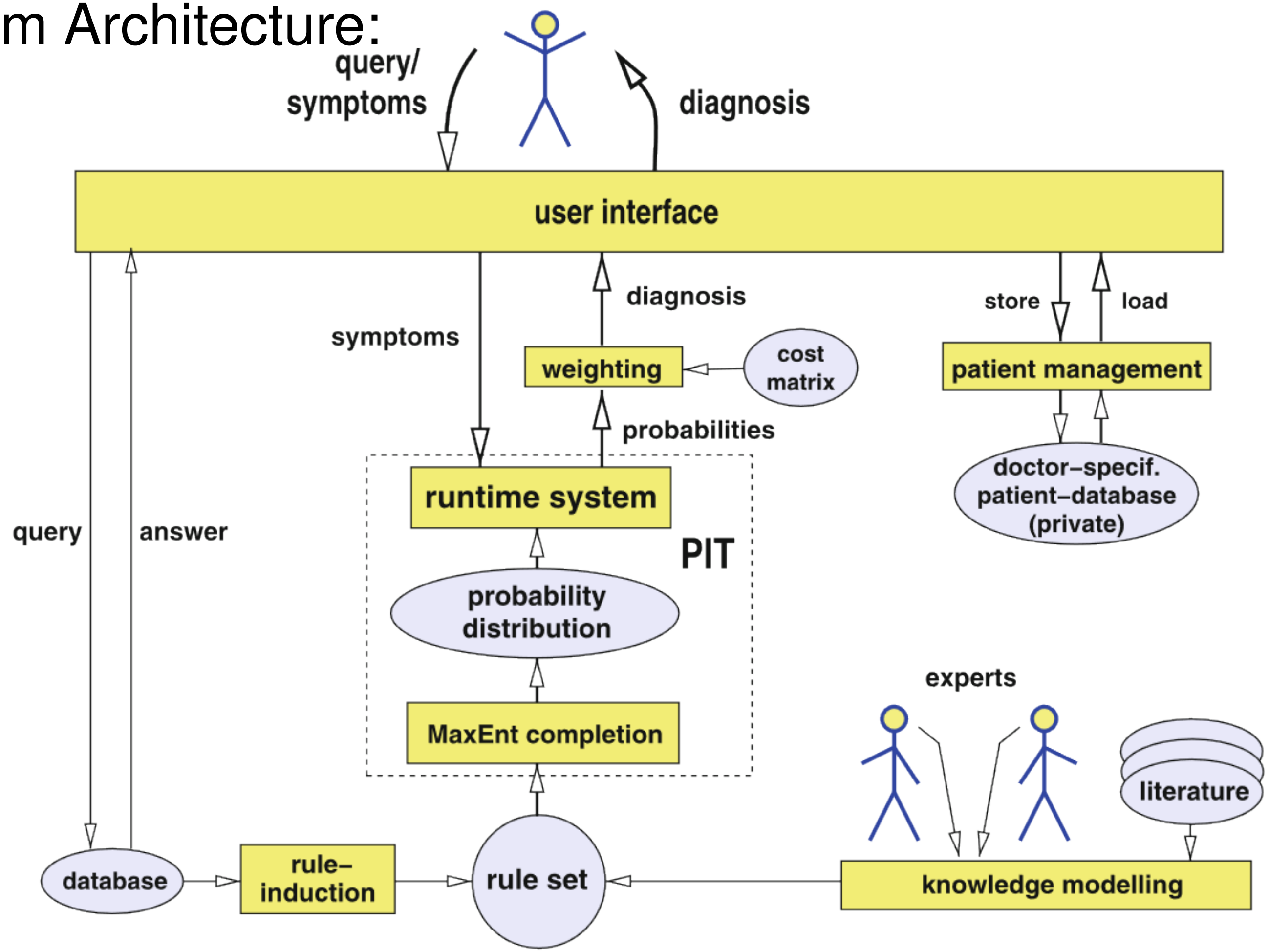


Figure: Rules are generated from the database as well as from expert knowledge. From these, MaxEnt creates a complete probability distribution. For a user query, the probability of every possible diagnosis is calculated. Using the cost matrix (will be defined later) a decision is then suggested.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Appendicitis Diagnosis with Formal Methods

Hybrid Probabilistic Knowledge Base

66 Application of LEXMED

Function of LEXMED

Risk Management Using the Cost Matrix

Performance

Application Areas and Experiences

Reasoning with Bayesian Networks

Summary

References

The usage of LEXMED is simple and self-explanatory.

- The doctor visits the LEXMED home page at `www.lexmed.de`.
- Doctor inputs the results of his examination into the input form.
    - If certain examination results are missing as input (for example the sonogram results), then the doctor chooses the entry *not examined*.
    - Entering more symptom values gives LEXMED more information for its assessment. However, a particular new result may strengthen the current leading diagnosis, weaken it, or make no difference.
- LEXMED outputs the probabilities for the four different diagnoses as well as a suggestion for a treatment.
- Each registered user has access to a private patient database, in which input data can be archived.
- Thus data and diagnoses from earlier patients can be easily compared with those of a new patient.

Knowledge is formalized using probabilistic propositions, e.g.,

$$P(\textit{Leuko7} > 20\,000 \mid \textit{Diag4} = \textit{inflamed}) = 0.09.$$

Note: Instead of single numerical values, we might also use intervals (i.e. $[0.06, 0.12]$).

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Appendicitis Diagnosis with Formal Methods

Hybrid Probabilistic Knowledge Base

Application of LEXMED

68 Function of LEXMED

Risk Management Using the Cost Matrix

Performance

Application Areas and Experiences

Reasoning with Bayesian Networks

Summary

References

Learning of Rules by Statistical Induction

- *Raw data in LEXMED:* 54 different (anonymized) values for $14,646$ patients (whose appendixes were surgically removed).
- After a statistical analysis, 14 *symptoms* among 54 attributes are selected and used *for the query in LEXMED*.
- Two steps to *create the rules* from this databases:
  (1) Determining the dependency structure of the symptoms.
  (2) Filling this structure with the respective probability rules.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

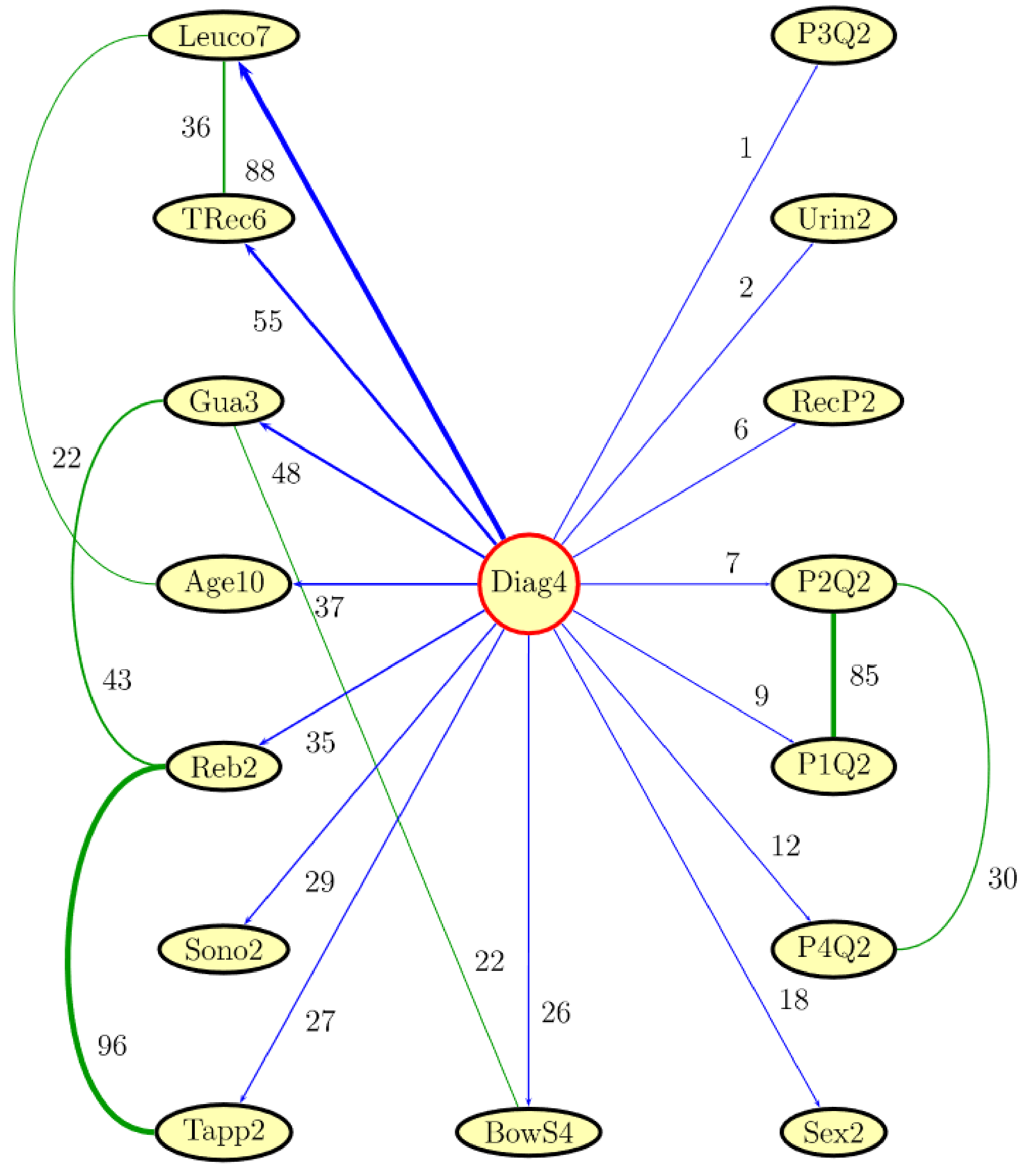
Dependency graph computed from the database (see the next slide)

- Node: variable (symptom + diagnosis)
- Edge: directed
- Edge's thickness: measures the correlation of the variables
  - Two independent variables: correlation = 0.
  - The pair correlation for each of the 14 symptoms with *Diag4* was computed. (Blue edges).
  - All triple correlations between the diagnosis and two symptoms were calculated. Of these, only the strongest values have been drawn as additional edges between the two participating symptoms. (Green edges.)

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

LEXMED

Function of LEXMED



Estimating the *Rule Probabilities*

- Structure of the dependency graph = structure of the learned rules.
- The rules have different complexities.
  - Rules which only describe the *distribution of the possible diagnoses* (a priori rules), e.g., $P(\textit{Diag4} = \textit{inflamed}) = 0.40$.
  - Rules which describe the *dependency between the diagnosis and a symptom* (rules with simple conditions), e.g., $P(\textit{Sono2} = \textit{yes} \mid \textit{Diag4} = \textit{inflamed}) = 0.43$.
  - Rules which describe the *dependency between the diagnosis and two symptoms*, e.g., $P(\textit{P4Q2} = \textit{yes} \mid \textit{Diag4} = \textit{inflamed} \wedge \textit{P2Q2} = \textit{yes}) = 0.61$.
- The numerical values for these rules are estimated by *counting their frequency in the database*.
  - For example, $P(\textit{Sono2} = \textit{yes} \mid \textit{Diag4} = \textit{inflamed}) = 0.43$ because we count in the database and calculate

$$\frac{|\textit{Diag4} = \textit{inflamed} \wedge \textit{Sono2} = \textit{yes}|}{|\textit{Diag4} = \textit{inflamed}|} = 0.43.$$

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

LEXMED

Function of LEXMED

Some of the LEXMED rules with probability intervals written in the historical PIT syntax. “$*$” stands for “$\wedge$” here.

```
P([Leuco7=0-6k]    | [Diag4=negativ] * [Age10=16-20]) = [0.132,0.156];
P([Leuco7=6-8k]    | [Diag4=negativ] * [Age10=16-20]) = [0.257,0.281];
P([Leuco7=8-10k]   | [Diag4=negativ] * [Age10=16-20]) = [0.250,0.274];
P([Leuco7=10-12k]  | [Diag4=negativ] * [Age10=16-20]) = [0.159,0.183];
P([Leuco7=12-15k]  | [Diag4=negativ] * [Age10=16-20]) = [0.087,0.112];
P([Leuco7=15-20k]  | [Diag4=negativ] * [Age10=16-20]) = [0.032,0.056];
P([Leuco7=20k-]    | [Diag4=negativ] * [Age10=16-20]) = [0.000,0.023];
P([Leuco7=0-6k]    | [Diag4=negativ] * [Age10=21-25]) = [0.132,0.172];
P([Leuco7=6-8k]    | [Diag4=negativ] * [Age10=21-25]) = [0.227,0.266];
P([Leuco7=8-10k]   | [Diag4=negativ] * [Age10=21-25]) = [0.211,0.250];
P([Leuco7=10-12k]  | [Diag4=negativ] * [Age10=21-25]) = [0.166,0.205];
P([Leuco7=12-15k]  | [Diag4=negativ] * [Age10=21-25]) = [0.081,0.120];
P([Leuco7=15-20k]  | [Diag4=negativ] * [Age10=21-25]) = [0.041,0.081];
P([Leuco7=20k-]    | [Diag4=negativ] * [Age10=21-25]) = [0.004,0.043];
```

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Expert Rules

- Rules for non-specific abdominal pain (NSAP) receive their values from propositions of medical experts.
- To model the uncertainty of expert knowledge, the use of probability intervals has proven effective.
- Once the expert rules have been created, the rule base is finished.

In LEXMED, the complete probability model was calculated by PIT using the maximum-entropy method.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Appendicitis Diagnosis with Formal Methods

Hybrid Probabilistic Knowledge Base

Application of LEXMED

Function of LEXMED

Risk Management Using the Cost Matrix

Performance

Application Areas and Experiences

Reasoning with Bayesian Networks

Summary

References

- Using its efficiently stored probability model, LEXMED calculates the probabilities for the four possible diagnoses within a few seconds. For example, we assume the following output:

	Results of the PIT diagnosis			
Diagnosis	Appendix inflamed	Appendix perforated	Negative	Other
Probability	0.24	0.16	0.57	0.03

- A decision must be made based on these four probability values.
- *How to derive an optimal decision from these probabilities?*
- Ask LEXMED to calculate a recommended decision.

Question

How can the computed probabilities now be translated optimally into decisions?

Naive algorithm

- assign a decision to each diagnosis
- ultimately select the decision that corresponds to the highest probability

Example 13 (Naive Algorithm)

- 0.4 for the diagnosis *appendicitis* (inflamed or perforated), 0.55 for the diagnosis *negative*, and 0.05 for the diagnosis *other*.
- Decide "no operation" (which may be too risky).

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Cost-oriented method

- Comparing the *costs of the possible errors that can occur for each decision*. (= costs for wrong decisions)
  - *The error is quantified* in the form of "(hypothetical) additional cost of the current decision compared to the optimum".
  - The given values contain the *costs to the hospital, to the insurance company, the patient* (for example risk of post-operative complications), *and to other parties* (for example absence from work), *taking into account long term consequences*.
  - *Optimal decisions* have (additional) *costs* 0.
- *The entries are finally averaged for each decision*, that is, summed while taking into account their frequencies.
- Finally, the *decision with the smallest average cost of error* is suggested.

LEXMED

Risk Management Using the Cost Matrix

Therapy	Probability of various diagnoses				
	inflamed	*perforated*	*negative*	*other*	
	0.25	0.15	0.55	0.05	
Operation	0	500	5800	6000	3565
Emergency operation	500	0	6300	6500	3915
Ambulant observ.	12000	150000	0	16500	26325
Other	3000	5000	1300	0	2215
Stationary observ.	3500	7000	400	600	2175

Figure: The cost matrix of LEXMED together with a patient's computed diagnosis probabilities.

- Computed probabilities for the four possible diagnoses: $(0.25, 0.15, 0.55, 0.05)$.
- The last column of the table contains the result of the calculations of the average expected costs of the errors.
  - E.g., the cost of the errors corresponds to Operation:

$$0.25 \cdot 0 + 0.15 \cdot 500 + 0.55 \cdot 5800 + 0.05 \cdot 6000 = 3565.$$

Cost matrix in the binary case

- Diagnosis: *Appendicitis* and *NSAP* (a.k.a *non-specific abdominal pain*).

$$P(\textit{Appendicitis}) = p_1$$
$$P(\textit{NSAP}) = p_2$$

- Therapies: *operation*, *ambulant observ.* (= send patient home).
- Cost matrix:

	Appendicitis	NSAP
operation	0	k_2
ambulant observ.	k_1	0

$$\begin{pmatrix} 0 & k_2 \\ k_1 & 0 \end{pmatrix}$$

  - *Correct decision:* cost 0.
  - *False positive:* cost k_2 = expected costs which occur when a patient without an inflamed appendix is operated on
  - *False negative:* cost k_1 = expected costs which occur when deciding to send the patient home in the case of appendicitis

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Cost matrix in the binary case

- Average additional cost for the two possible treatments:

$$\begin{pmatrix} 0 & k_2 \\ k_1 & 0 \end{pmatrix} \cdot \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} k_2 p_2 \\ k_1 p_1 \end{pmatrix}$$

- Multiplying the vector $(k_2 p_2, k_1 p_1)^T$ by any *positive* scalar, say $1/k_1 > 0$, does not affect the final decision (as we only care about which one is smaller).
  $\Rightarrow$ Only the relationship $k = k_2/k_1$ is relevant.
  $\Rightarrow$ Same result with the cost matrix $\begin{pmatrix} 0 & k \\ 1 & 0 \end{pmatrix}$.
- *Risk management*
  - By changing k we can fit the "working point" of the diagnosis system.
  - $k \to \infty$: extremely risky setting, no patient will ever be operated on $\Rightarrow$ the system gives no False positive but many False negatives.
  - $k = 0$: all patients are operated.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Exercise 11 ([Ertel 2025], Exercise 7.9*, p. 175)

A probabilistic algorithm calculates the likelihood p that an inbound email is spam. To classify the emails in classes *delete* and *read*, a cost matrix is then applied to the result.

(a) Give a cost matrix (2×2 matrix) for the spam filter. Assume here that it costs the user 10 cents to delete an email, while the loss of an email costs 10 dollars. (Note: 1 dollar = 100 cents.)

(b) Show that, for the case of a 2×2 matrix, the application of the cost matrix is equivalent to the application of a threshold on the spam probability and determine the threshold.

LEXMED
Performance

In a prospective study, LEXMED was evaluated on a *test set of 185 cases of suspected appendicitis*. To compare with similar studies, LEXMED's output was restricted to two diagnoses: *appendicitis* and *NSAP*. [Ertel 2025], p. 158
For each k $(0 \leq k < \infty)$, the *sensitivity* and *specificity* were measured on this test set:

$$\begin{aligned}
\textit{Sensitivity} &= P(\textit{classified positive} \mid \textit{positive}) \\
&= \frac{|\textit{classified positive} \text{ and } \textit{positive}|}{|\textit{positive}|} \\
\textit{Specificity} &= P(\textit{classified negative} \mid \textit{negative}) \\
&= \frac{|\textit{classified negative} \text{ and } \textit{negative}|}{|\textit{negative}|}
\end{aligned}$$

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
- Appendicitis Diagnosis with Formal Methods
- Hybrid Probabilistic Knowledge Base
- Application of LEXMED
- Function of LEXMED
- Risk Management Using the Cost Matrix
- Performance
- Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

The previous slide defined *sensitivity* and *specificity* for each value of k. Applying these measures to the *185 prospective cases* described in [Ertel 2025], p. 158 gives the following result.

- At $k = 0.6$, LEXMED achieved a *sensitivity of 88%* and a *specificity of 87%*. [Ertel 2025], p. 159
- Thus, at this operating point, LEXMED correctly classified 88% of the appendicitis cases as positive and 87% of the NSAP cases as negative.
- This is a very good result from a *small study*. The next slide discusses how LEXMED was intended to be used in practice. [Ertel 2025], p. 160

- Use in the diagnosis
- Quality assurance: comparing the diagnosis quality of hospitals with expert systems.
- Since 1999 in use in the 14-Nothelfer hospital in Weingarten
- In the reported small study, LEXMED's diagnostic performance was comparable to that of an experienced surgeon; it was intended as a second opinion, not a replacement for clinical judgment. [Ertel 2025], p. 160
- Commercial marketing very difficult
- Wrong time?
- Wish of patients for personal care!
- When will computer-aided diagnosis become an established medical tool?

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Appendicitis Diagnosis with Formal Methods
Hybrid Probabilistic Knowledge Base
Application of LEXMED
Function of LEXMED
Risk Management Using the Cost Matrix
Performance
84 Application Areas and Experiences
Reasoning with Bayesian Networks
Summary
References

Reasoning with Bayesian Networks

- d variables $X_1, \ldots, X_d$, each with n values, have n^d possible joint assignments.
- A full joint-probability table therefore has n^d *entries*, of which $n^d - 1$ *are free parameters* after normalization.
- *Bayesian networks* exploit conditional-independence structure to represent many such distributions with far fewer parameters.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- Simplest case: all variables are *mutually independent*[1]

$$\mathbf{P}(X_1, X_2, \ldots, X_d) = \mathbf{P}(X_1) \cdot \mathbf{P}(X_2) \cdot \cdots \cdot \mathbf{P}(X_d)$$

- For independent events A, B with $P(B) > 0$, conditional probabilities become trivial:

$$P(A \mid B) = P(A).$$

- The situation becomes more interesting when *only a portion of the variables are independent or independent under certain conditions*. For reasoning in AI, the dependencies between variables happen to be important and must be utilized.

[1] Pairwise independence requires only that each pair X_i, X_j be independent. Mutual independence requires the joint distribution of every subcollection to factorize, including the full joint distribution shown here. Ertel [Ertel 2025], p. 161 writes "pairwise independent" but then states the full factorization; for $d \geq 3$ these two conditions are not equivalent. We therefore replace Ertel's wording by "mutually independent".

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

Semantics of Bayesian Networks

Summary

References

Example 14 (Alarm-Example, [Pearl 1988]; [Russell and Norvig 2010])

- Bob: single, has an alarm system in his house.
- John and Mary: neighbors of Bob in the houses next door to the left and right, respectively.
- Bob asks John and Mary to call him at his office if they hear the alarm.
- *Knowledge Base:*
  - Variables: $J =$ "John calls", $M =$ "Mary calls", *Al* $=$ "Alarm siren sounds", *Bur* $=$ "Burglary", *Ear* $=$ "Earthquake"
  - Calling behaviors of John and Mary

$$P(J \mid Al) = 0.90 \qquad P(M \mid Al) = 0.70$$
$$P(J \mid \neg Al) = 0.05 \qquad P(M \mid \neg Al) = 0.01$$

  - The alarm is triggered by a burglary, but can also be triggered by a (weak) earthquake, which can lead to a false alarm.

$$P(Al \mid Bur, Ear) = 0.95 \qquad P(Al \mid \neg Bur, Ear) = 0.29$$
$$P(Al \mid Bur, \neg Ear) = 0.94 \qquad P(Al \mid \neg Bur, \neg Ear) = 0.001$$

  - A priori probabilities: $P(Bur) = 0.001$, $P(Ear) = 0.002$. (*Bur* and *Ear* are independent.)
- *Requests:* $P(Bur \mid J \vee M)$, $P(J \mid Bur)$, $P(M \mid Bur)$

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
87 Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- A *Bayesian network* is a directed acyclic graph (DAG) in which
  - each node represents a random variable,
  - each edge $X_i \rightarrow X_j$ indicates that X_i is a *parent* of X_j and, when a causal interpretation is intended, is usually read as a *direct influence* of X_i on X_j, and
  - each node is associated with a conditional probability table (CPT) that specifies its conditional distribution given its parents.
- The structure of the graph encodes *conditional independence* assumptions that can be exploited to simplify the representation of the joint probability distribution.
- The joint probability distribution over all variables $X_1, \dots, X_d$ can be expressed as

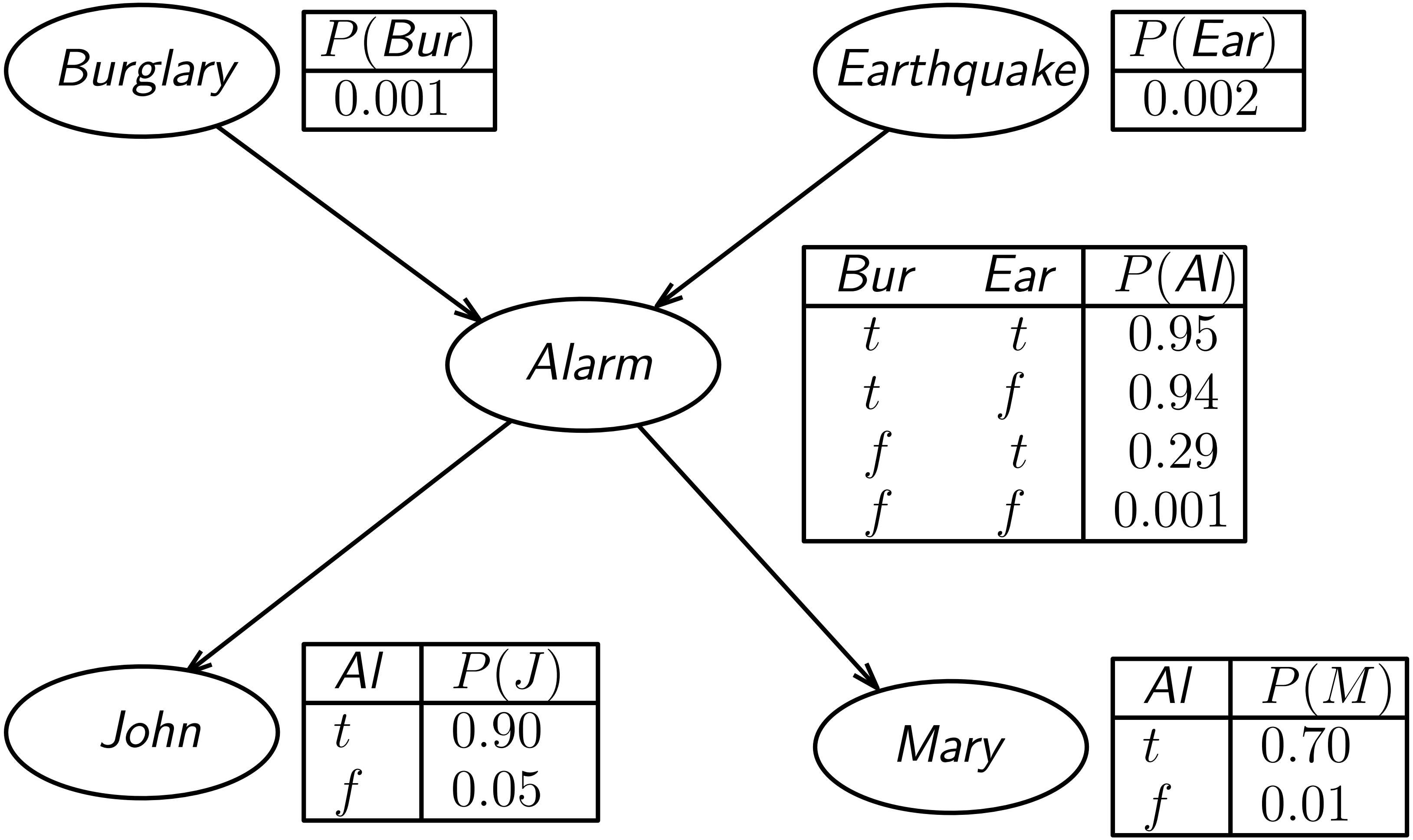
$$\mathbf{P}(X_1, X_2, \dots, X_d) = \prod_{i=1}^{d} \mathbf{P}(X_i \mid \text{Parents}(X_i)),$$

where Parents(X_i) denotes the set of parent nodes of X_i

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Reasoning with Bayesian Networks

Graphical Representation of Knowledge as a Bayesian Network



$P(Bur)$
0.001

$P(Ear)$
0.002

Bur	Ear	$P(Al)$
t	t	0.95
t	f	0.94
f	t	0.29
f	f	0.001

Al	$P(J)$
t	0.90
f	0.05

Al	$P(M)$
t	0.70
f	0.01

Figure: Bayesian network for the alarm example with the associated CPTs (conditional probability tables). The CPT of a node lists all the conditional probabilities of the node's variable conditioned on all the nodes connected by incoming edges.

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Definition

Two variables A and B are called *conditionally independent* given C if, for every a, b, c with $P(C = c) > 0$,

$$P(A = a, B = b \mid C = c) = P(A = a \mid C = c)P(B = b \mid C = c).$$

Remark

- independent $\not\Rightarrow$ conditionally independent.
- conditionally independent $\not\Rightarrow$ independent.

- *A and B are independent events* means knowing that A happened would not tell you anything about whether B happened (or vice versa).
- *A and B are conditionally independent events, given C* means that if you already knew that C happened, then knowing that A happened would not tell you further information about whether B happened.

Example 15 (independent $\not\Rightarrow$ cond. independent)

There are two fair, separate coins: each lands heads with probability 0.5. Toss both coins once.

Variable	Value
E(*Same result*)	t, f
F(*First coin*)	H(*Head*), T(*Tail*)
S(*Second coin*)	H(*Head*), T(*Tail*)

Head **Tail**

- $\mathbf{P}(F, S) = \mathbf{P}(F) \cdot \mathbf{P}(S)$
  - The two results are *independent*: knowing one result alone gives no information about the other.
- $\mathbf{P}(F, S \mid E) \neq \mathbf{P}(F \mid E) \cdot \mathbf{P}(S \mid E)$
  - Suppose that, *by some magic power*, we know the value of E without knowing either individual coin result. If $E = t$, only HH and TT remain possible; if $E = f$, only HT and TH remain possible. *In either case, the result of the first coin tells you exactly the result of the second coin and vice versa.*

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
91 Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Example 16 (cond. independent $\not\Rightarrow$ independent)

There are two biased coins. Each toss of *coin1* results in heads with probability 0.99, while each toss of *coin2* results in heads with probability 0.01. Select one of the two coins uniformly at random, keep that coin, and toss it twice. Given the selected coin, the two tosses are independent.

Variable	**Value**
C(*Coin*)	c_1(*coin1*), c_2(*coin2*)
F(*First toss*)	H(*Head*), T(*Tail*)
S(*Second toss*)	H(*Head*), T(*Tail*)

Head **Tail**

- $\mathbf{P}(F, S \mid C) = \mathbf{P}(F \mid C) \cdot \mathbf{P}(S \mid C)$
  - Given the selected coin, knowing the first toss does not help predict the second toss.
- $\mathbf{P}(F, S) \neq \mathbf{P}(F) \cdot \mathbf{P}(S)$
  - If the selected coin is unknown, the first toss provides information about it. For example, a head on the first toss is strong evidence that *coin1* was selected, so the second toss is very likely to result in heads.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Exercise 12

Recall the preceding experiment. One coin is selected uniformly at random, kept, and tossed twice. Given the selected coin, the two tosses are independent.

For each toss of the selected coin:

	Selected	Heads/toss
coin1	$1/2$	0.99
coin2	$1/2$	0.01

Let

- C_1: "*coin1* was selected";
- C_2: "*coin2* was selected";
- H_1: "the first toss is heads";
- H_2: "the second toss is heads".

(a) Compute $P(H_1)$.

(b) Compute $P(C_1 \mid H_1)$ and $P(C_2 \mid H_1)$.

(c) Using conditional independence given the selected coin, compute $P(H_2 \mid H_1)$ by considering separately the events C_1 and C_2.

(d) Compute $P(H_2)$ and compare it with $P(H_2 \mid H_1)$. What does this show about the independence of the two tosses when the selected coin is unknown?

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Example 17 (Alarm-Example (cont.))

- John and Mary independently react to an alarm.
  $\mathbf{P}(J, M \mid \mathit{Al}) = \mathbf{P}(J \mid \mathit{Al}) \cdot \mathbf{P}(M \mid \mathit{Al})$.
- Thus, *given an alarm, two variables J and M are independent*.
- (Without any condition,) J and M are not independent, that is, $\mathbf{P}(J, M) \neq \mathbf{P}(J) \cdot \mathbf{P}(M)$. **[Why?]**
  - **Hint:** It is enough to find one value assignment for which independence fails; use $J = t$ and $M = t$.
  - Calculate $P(\mathit{Al})$ by marginalization, using independence of *Bur* and *Ear*.
    (Result: $P(\mathit{Al}) \approx 0.00252$.)
  - Then calculate $P(J)$ and $P(M)$ from the computed $P(\mathit{Al})$.
    (Results: $P(J) \approx 0.05214$, $P(M) \approx 0.01174$.)
  - Similarly, calculate $P(J, M)$ using conditional independence of J and M given *Al*.
    (Result: $P(J, M) \approx 0.002084$.)
  - Compare $P(J, M)$ and $P(J) \cdot P(M)$.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

Semantics of Bayesian Networks

Summary

References

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Example 18 (Alarm-Example (cont.))

- John react to an alarm, but does not react to a burglary. (This could be, for example, because of a high wall that blocks his view on Bob's property, but he can still hear the alarm.)

$$\mathbf{P}(J, \mathit{Bur} \mid \mathit{Al}) = \mathbf{P}(J \mid \mathit{Al}) \cdot \mathbf{P}(\mathit{Bur} \mid \mathit{Al}).$$

- Given an alarm, the variables J and *Ear*, M and *Bur*, as well as M and *Ear* are also independent.

$$\mathbf{P}(J, \mathit{Ear} \mid \mathit{Al}) = \mathbf{P}(J \mid \mathit{Al}) \cdot \mathbf{P}(\mathit{Ear} \mid \mathit{Al})$$
$$\mathbf{P}(M, \mathit{Bur} \mid \mathit{Al}) = \mathbf{P}(M \mid \mathit{Al}) \cdot \mathbf{P}(\mathit{Bur} \mid \mathit{Al})$$
$$\mathbf{P}(M, \mathit{Ear} \mid \mathit{Al}) = \mathbf{P}(M \mid \mathit{Al}) \cdot \mathbf{P}(\mathit{Ear} \mid \mathit{Al})$$

Theorem 6

The following equations are pairwise equivalent, which means that each individual equation describes the conditional independence for the variables A and B given C.

$$\mathbf{P}(A, B \mid C) = \mathbf{P}(A \mid C) \cdot \mathbf{P}(B \mid C) \quad (1)$$

$$\mathbf{P}(A \mid B, C) = \mathbf{P}(A \mid C) \quad (2)$$

$$\mathbf{P}(B \mid A, C) = \mathbf{P}(B \mid C) \quad (3)$$

Proof.

We prove Eq. (1) $\Leftrightarrow$ Eq. (2). Similarly for Eq. (1) and Eq. (3).

(a) Chain rule: $\mathbf{P}(A, B, C) = \mathbf{P}(A \mid B, C)\mathbf{P}(B \mid C)\mathbf{P}(C)$.

(b) Definition: $\mathbf{P}(A, B, C) = \mathbf{P}(A, B \mid C)\mathbf{P}(C)$.

(c) Eq. (1) $\Rightarrow$ Eq. (2): From Eq. (1) and (b), $\mathbf{P}(A, B, C) = \mathbf{P}(A \mid C)\mathbf{P}(B \mid C)\mathbf{P}(C)$. Comparing with (a).

(d) Eq. (2) $\Rightarrow$ Eq. (1): From Eq. (2) and (a), $\mathbf{P}(A, B, C) = \mathbf{P}(A \mid C)\mathbf{P}(B \mid C)\mathbf{P}(C)$. Comparing with (b).

$\square$

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
96 Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Reasoning with Bayesian Networks

Practical Application

All probabilities in the alarm CPTs lie strictly between 0 and 1, so every conditioning event displayed below has positive probability.

$$P(J \mid \mathit{Bur}) = \frac{P(J, \mathit{Bur})}{P(\mathit{Bur})} = \frac{P(J, \mathit{Bur}, \mathit{Al}) + P(J, \mathit{Bur}, \neg \mathit{Al})}{P(\mathit{Bur})}$$

$$\begin{aligned} \mathbf{P}(J, \mathit{Bur}, \mathit{Al}) &= \mathbf{P}(J \mid \mathit{Bur}, \mathit{Al})\mathbf{P}(\mathit{Al} \mid \mathit{Bur})\mathbf{P}(\mathit{Bur}) \\ &= \mathbf{P}(J \mid \mathit{Al})\mathbf{P}(\mathit{Al} \mid \mathit{Bur})\mathbf{P}(\mathit{Bur}) \end{aligned}$$

Chain rule

J and *Bur* are independent given *Al*

$$\begin{aligned} P(J \mid \mathit{Bur}) &= \frac{P(J \mid \mathit{Al})P(\mathit{Al} \mid \mathit{Bur})P(\mathit{Bur})}{P(\mathit{Bur})} \\ &+ \frac{P(J \mid \neg \mathit{Al})P(\neg \mathit{Al} \mid \mathit{Bur})P(\mathit{Bur})}{P(\mathit{Bur})} \\ &= P(J \mid \mathit{Al})P(\mathit{Al} \mid \mathit{Bur}) + P(J \mid \neg \mathit{Al})P(\neg \mathit{Al} \mid \mathit{Bur}) \end{aligned}$$

Reasoning with Uncertainty
Hoàng Anh Đức

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Since *Bur* and *Ear* are independent,

$$
\begin{aligned}
P(\mathit{Al} \mid \mathit{Bur}) &= P(\mathit{Al} \mid \mathit{Bur}, \mathit{Ear}) P(\mathit{Ear}) \\
&\quad + P(\mathit{Al} \mid \mathit{Bur}, \neg \mathit{Ear}) P(\neg \mathit{Ear}) \\
&= 0.95 \cdot 0.002 + 0.94 \cdot 0.998 = 0.94002.
\end{aligned}
$$

Thus, $P(\mathit{Al} \mid \mathit{Bur}) \approx 0.94$ and $P(\neg \mathit{Al} \mid \mathit{Bur}) = 0.05998 \approx 0.06$.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

- Independent Variables
- Graphical Representation of Knowledge as a Bayesian Network
- Conditional Independence
- Practical Application
- Software for Bayesian Networks
- Development of Bayesian Networks
- Semantics of Bayesian Networks

Summary

References

Therefore,

$$P(J \mid \mathit{Bur}) = P(J \mid \mathit{Al})P(\mathit{Al} \mid \mathit{Bur}) + P(J \mid \neg\mathit{Al})P(\neg\mathit{Al} \mid \mathit{Bur})$$
$$= 0.9 \cdot 0.94002 + 0.05 \cdot 0.05998 = 0.849017 \approx 0.849.$$

Analogously, $P(M \mid \mathit{Bur}) = 0.6586138 \approx 0.659.$

Similar to $P(J \mid \mathit{Bur})$, we can calculate

$$\begin{aligned} P(J, M \mid \mathit{Bur}) &= P(J \mid \mathit{Al})P(M \mid \mathit{Al})P(\mathit{Al} \mid \mathit{Bur}) \\ &+ P(J \mid \neg\mathit{Al})P(M \mid \neg\mathit{Al})P(\neg\mathit{Al} \mid \mathit{Bur}) \\ &= 0.9 \cdot 0.7 \cdot 0.94002 + 0.05 \cdot 0.01 \cdot 0.05998 \\ &= 0.59224259 \approx 0.5922. \end{aligned}$$

John calls for about 85% of all break-ins and Mary for about 66% of all break-ins. Both of them call for about 59% of all break-ins.

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

$$
\begin{aligned}
P(J \vee M \mid \textit{Bur}) &= 1 - P(\neg J, \neg M \mid \textit{Bur}) \\
P(\neg J, \neg M \mid \textit{Bur}) &= P(\neg J \mid \textit{Al}) P(\neg M \mid \textit{Al}) P(\textit{Al} \mid \textit{Bur}) \\
&\quad + P(\neg J \mid \neg \textit{Al}) P(\neg M \mid \neg \textit{Al}) P(\neg \textit{Al} \mid \textit{Bur}) \\
&= 0.1 \cdot 0.3 \cdot 0.94002 + 0.95 \cdot 0.99 \cdot 0.05998 \\
&= 0.08461179, \\
P(J \vee M \mid \textit{Bur}) &= 1 - 0.08461179 = 0.91538821 \approx 0.915.
\end{aligned}
$$

$$
\begin{aligned}
P(J \vee M) &= P(J) + P(M) - P(J, M) \approx 0.06179, \\
P(\textit{Bur} \mid J \vee M) &= \frac{P(J \vee M \mid \textit{Bur}) P(\textit{Bur})}{P(J \vee M)} \approx 0.0148.
\end{aligned}
$$

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

- Independent Variables
- Graphical Representation of Knowledge as a Bayesian Network
- Conditional Independence
- Practical Application
- Software for Bayesian Networks
- Development of Bayesian Networks
- Semantics of Bayesian Networks

Summary

References

Reasoning with Bayesian Networks

Practical Application

$$P(\textit{Bur} \mid J) = \frac{P(J \mid \textit{Bur})P(\textit{Bur})}{P(J)} \approx 0.0163,$$

$$P(\textit{Bur} \mid M) = \frac{P(M \mid \textit{Bur})P(\textit{Bur})}{P(M)} \approx 0.0561,$$

$$P(\textit{Bur} \mid J, M) = \frac{P(J, M \mid \textit{Bur})P(\textit{Bur})}{P(J, M)} \approx 0.2842.$$

- Only about 1.6% of the cases in which John calls involve a burglary; for Mary, the corresponding probability is about 5.6%. Although Mary calls less often when the alarm sounds ($P(M \mid \textit{Al}) = 0.70 < 0.90 = P(J \mid \textit{Al})$), she is only one fifth as likely as John to call when it is silent ($P(M \mid \neg \textit{Al}) = 0.01 = \frac{1}{5}P(J \mid \neg \textit{Al})$), so, in the full network, her call is stronger evidence of a burglary.
- If both John and Mary call, the probability of a burglary is about 28.4%, the highest of the three cases.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

Semantics of Bayesian Networks

Summary

References

Conditioning

$$P(A \mid B) = \sum_c P(A \mid B, C = c) P(C = c \mid B).$$

If furthermore A and B are conditionally independent given C, this formula simplifies to

$$P(A \mid B) = \sum_c P(A \mid C = c) P(C = c \mid B).$$

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Historical PIT input for the alarm example.

```
var Alarm{t,f}, Burglary{t,f}, Earthquake{t,f}, John{t,f}, Mary{t,f};

P([Earthquake=t]) = 0.002;
P([Burglary=t]) = 0.001;
P([Alarm=t] | [Burglary=t] AND [Earthquake=t]) = 0.95;
P([Alarm=t] | [Burglary=t] AND [Earthquake=f]) = 0.94;
P([Alarm=t] | [Burglary=f] AND [Earthquake=t]) = 0.29;
P([Alarm=t] | [Burglary=f] AND [Earthquake=f]) = 0.001;
P([John=t] | [Alarm=t]) = 0.90;
P([John=t] | [Alarm=f]) = 0.05;
P([Mary=t] | [Alarm=t]) = 0.70;
P([Mary=t] | [Alarm=f]) = 0.01;

QP([Burglary=t] | [John=t] AND [Mary=t]);
```

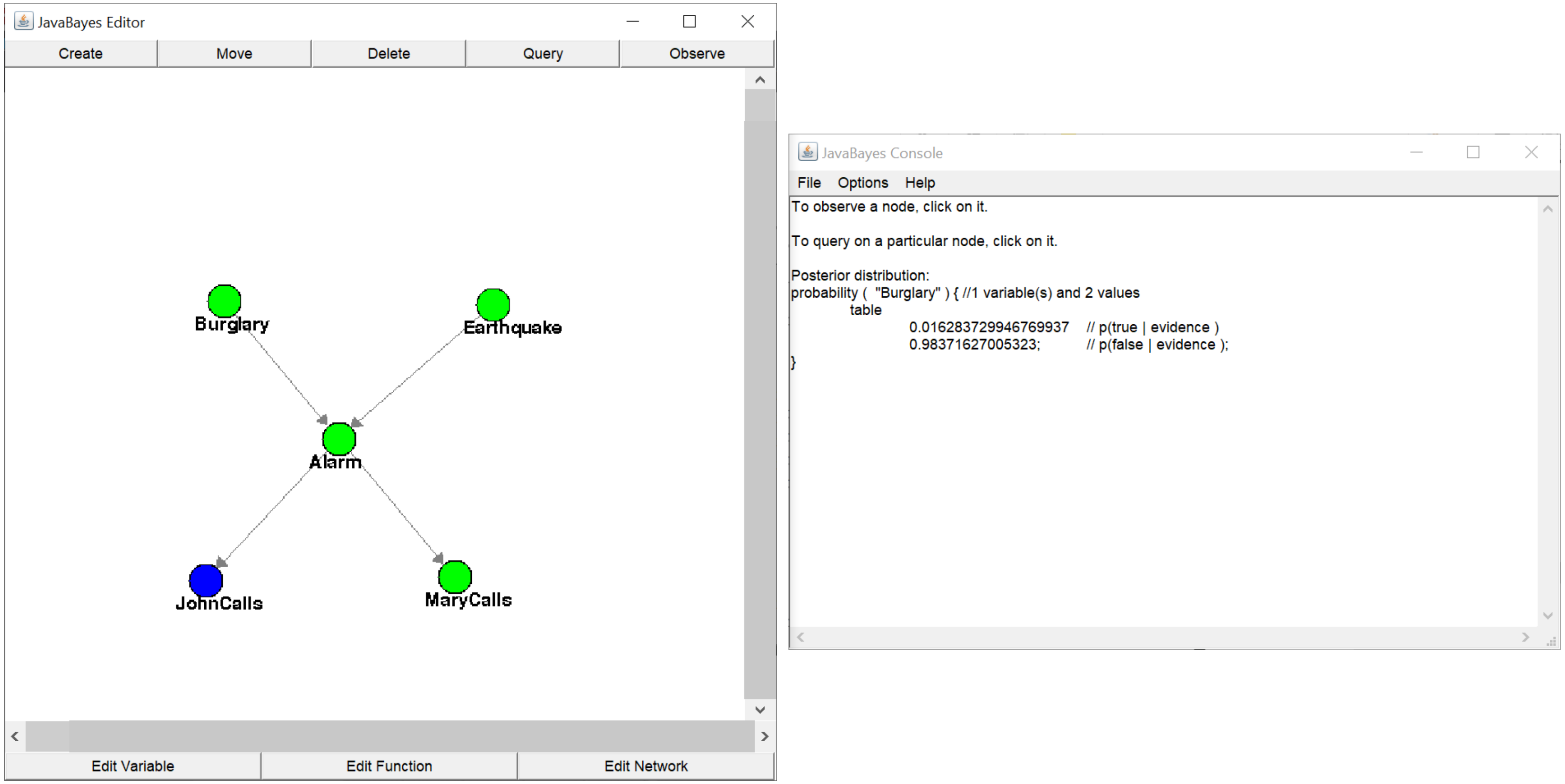
Reported response:

`P([Burglary=t] | [John=t] AND [Mary=t]) = 0.2841.`

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- PIT was a MaxEnt-based probabilistic reasoning tool rather than a classical Bayesian network tool.
- It accepted general probabilistic rules and queries. For consistent inputs, it completed the probability model using MaxEnt and answered the query; for inconsistent inputs, it reported the inconsistency.
- On input of CPTs or equivalent rules, its MaxEnt model gave the same conditional independences and answers as the corresponding Bayesian network.
- The former public PIT service is no longer available at the address cited by Ertel.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References



- A classic system is JavaBayes. Two windows: graphical editor + console
- With the graphical network editor, nodes and edges can be manipulated and the values in the CPTs edited.
- The values of variables can be assigned with “Observe” and the values of other variables called up with “Query”. The answers to queries then appear in the console

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- More powerful is the professional tool Hugin.
- Continuous variables possible.
- Can also learn Bayesian networks, that is, generate the network fully automatically from statistical data.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

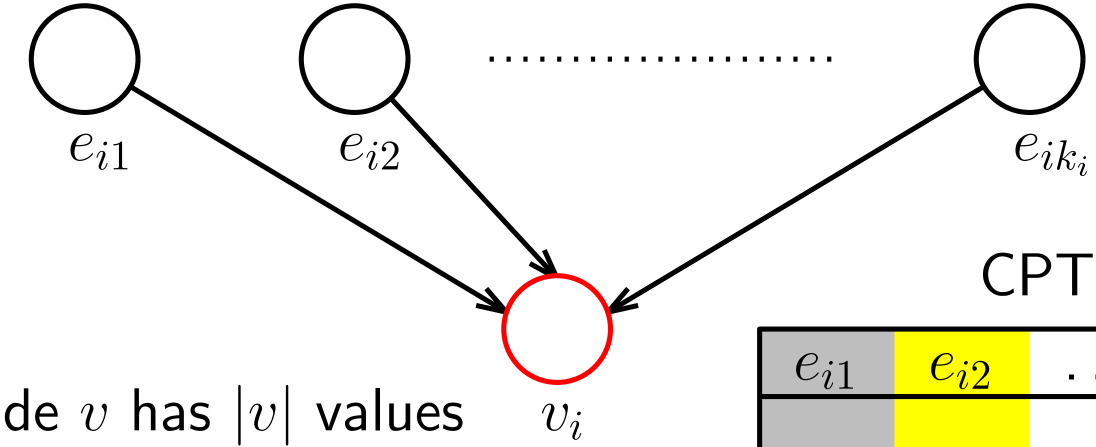
- For variables $v_1, \ldots, v_n$ with domain sizes $|v_1|, \ldots, |v_n|$, the full joint table has

$$\prod_{i=1}^{n} |v_i|$$

entries.

- Normalization leaves $\prod_{i=1}^{n} |v_i| - 1$ free parameters.
- Alarm example: $2^5 = 32$ table entries and $2^5 - 1 = 31$ free parameters.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References



A node v has $|v|$ values

Size of CPT at v_i

$$(|v_i| - 1) \prod_{j=1}^{k_i} |e_{ij}|$$

Total size of all CPTs

$$\sum_{i=1}^{n} (|v_i| - 1) \prod_{j=1}^{k_i} |e_{ij}|$$

CPT of v_i

e_{i1}	e_{i2}	...	e_{ik_i}	v_i
		...		

$$\mathbf{P}(v_i \mid e_{i1}, e_{i2}, \ldots, e_{ik_i})$$

(Require all combinations of values)

Alarm example: $2 + 2 + 4 + 1 + 1 = 10$ free CPT parameters.

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Suppose that there are n variables and that every variable has b possible values.

- Let $k_i = |\,\text{Parents}(X_i)|$ be the number of parents of X_i.
- There are b^{k_i} possible settings of those parents. For each setting, the b probabilities for X_i sum to one, leaving $b-1$ free values. Thus the CPT of X_i contains

  $$(b-1)b^{k_i}$$

  free parameters.
- Hence the network contains exactly

  $$N_{\text{BN}} = \sum_{i=1}^{n} (b-1)b^{k_i}$$

  free CPT parameters.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- If every node has *at most k parents*, then $k_i \leq k$ for every i, and therefore

$$N_{\mathrm{BN}} = \sum_{i=1}^{n} (b-1) b^{k_i} \leq n(b-1)b^k.$$

- The full joint distribution contains $b^n - 1$ free parameters. A complete ordered DAG has exactly the same count because

$$(b-1)\sum_{i=0}^{n-1} b^i = b^n - 1.$$

- For fixed b and k, the representation uses $O(n)$ parameters. Exact inference can nevertheless require exponential time and space in the worst case.

Why "at most" rather than "exactly"?

Every finite DAG has at least one root node with no parents. Thus, for $k > 0$, not every node can have exactly k parents.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

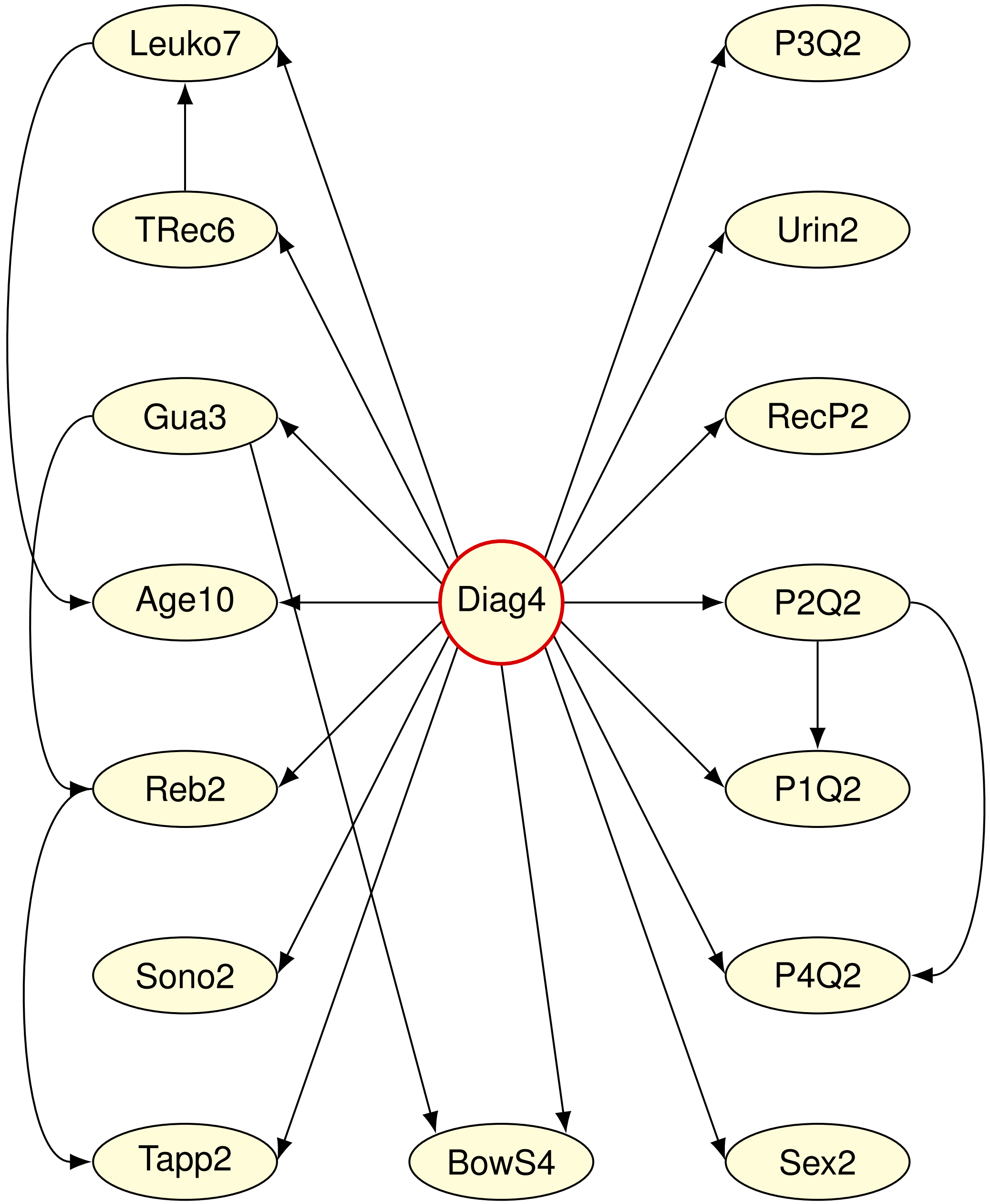
Software for Bayesian Networks

Development of Bayesian Networks

Semantics of Bayesian Networks

Summary

References



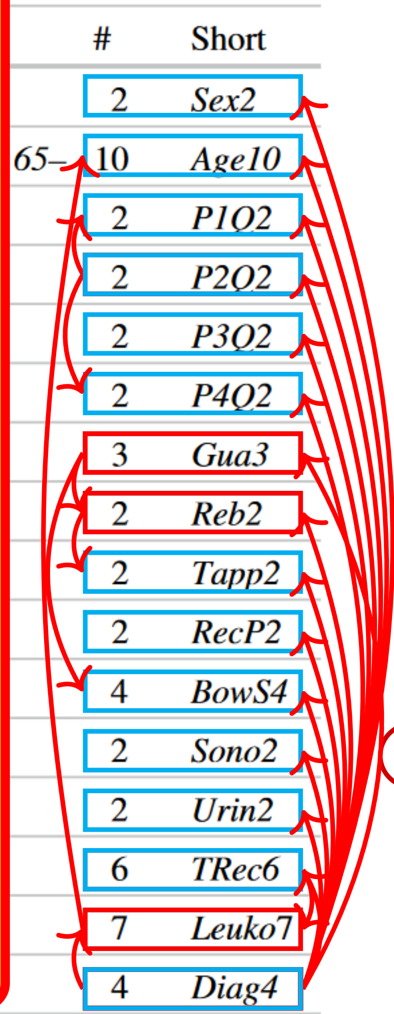
Bayesian network for the LEXMED application

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
111 Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- Full joint distribution: $20\,643\,839$ free parameters.
- Bayesian network: 523 free CPT parameters. **[Why?]**

$$\sum_{i=1}^{n}(|v_i|-1)\prod_{j=1}^{k_i}|e_{ij}| = 6\cdot 6\cdot 4+5\cdot 4+2\cdot 4+ 9\cdot 7\cdot 4+1\cdot 3\cdot 4+1\cdot 4+ 1\cdot 2\cdot 4+3\cdot 3\cdot 4+1\cdot 4+ 1\cdot 4\cdot 2+1\cdot 4\cdot 2+1\cdot 4+ 1\cdot 4+1\cdot 4+1\cdot 4+3 = 523.$$

Ertel's calculated result is 521, but the four-valued root *Diag4* contributes $4-1=3$, not 1, so the last value in the above equation is 3.



Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Construction of a Bayesian network usually proceeds in *two stages*:

(1) *Design of the network structure:* This stage is usually *performed manually*.

(2) *Entering the probabilities in the CPTs:* Manually entering the values in the case of many variables is very tedious. If (as for example with LEXMED) a database is available, this step can be *automated through estimating the CPT entries by counting frequencies*.

Construction of a Bayesian network

- **Step 1:** *Choose variables and an ordering.* Choose the variables and an ordering $X_1, \ldots, X_n$. Any ordering is allowed; *causes before effects often yields a simpler network*.
- **Step 2:** *Select a minimal parent set.* For each X_i, choose an inclusion-minimal set $\text{Parents}(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$. It must satisfy $\mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) = \mathbf{P}(X_i \mid \text{Parents}(X_i))$ for every assignment for which the conditional probabilities are defined. *Inclusion-minimal* means that *no selected parent can be removed while preserving this equality*.
- **Step 3:** *Add the edges.* For every selected parent $X_j \in \text{Parents}(X_i)$, *add the directed edge* $X_j \to X_i$.
- **Step 4:** *Specify the CPTs.* For every node, specify $\mathbf{P}(X_i \mid \text{Parents}(X_i))$: *list the probability of each value of* X_i *for every combination of its parents' values*.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

Semantics of Bayesian Networks

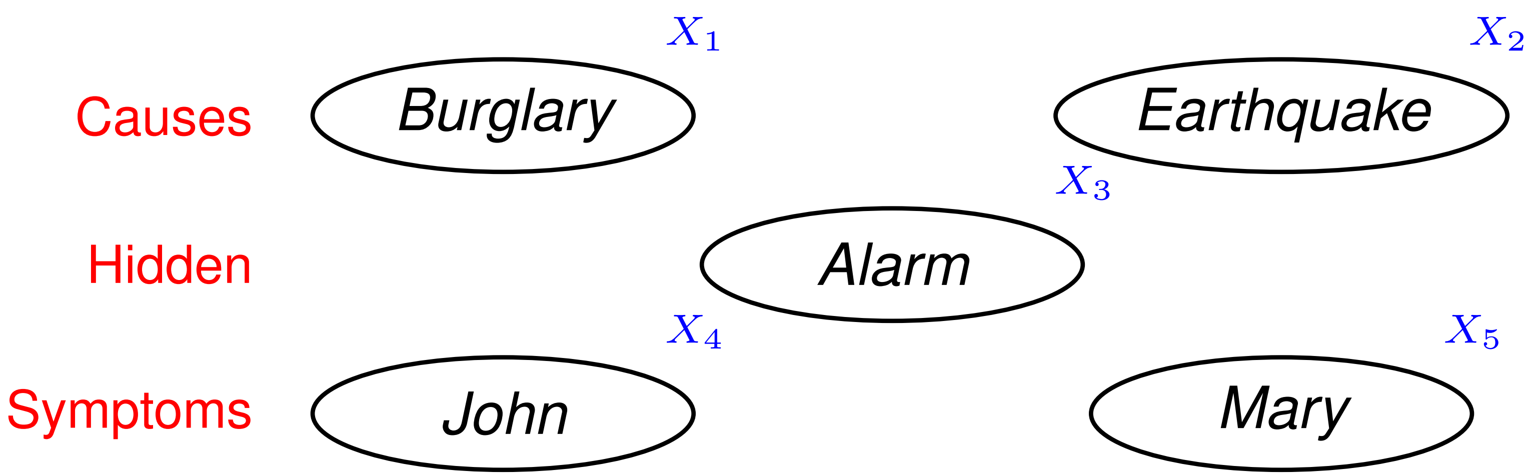
Summary

References

Construction of the network in the alarm example.

- **Causes:** *burglary* and *earthquake*
- **Symptoms:** *John* and *Mary*
- *Alarm:* hidden variable
  - Because John and Mary do not directly react to a burglar or earthquake, rather only to the alarm, it is appropriate to add this as an additional variable which is not observable by Bob.
- Considering causality: *going from cause to effect*

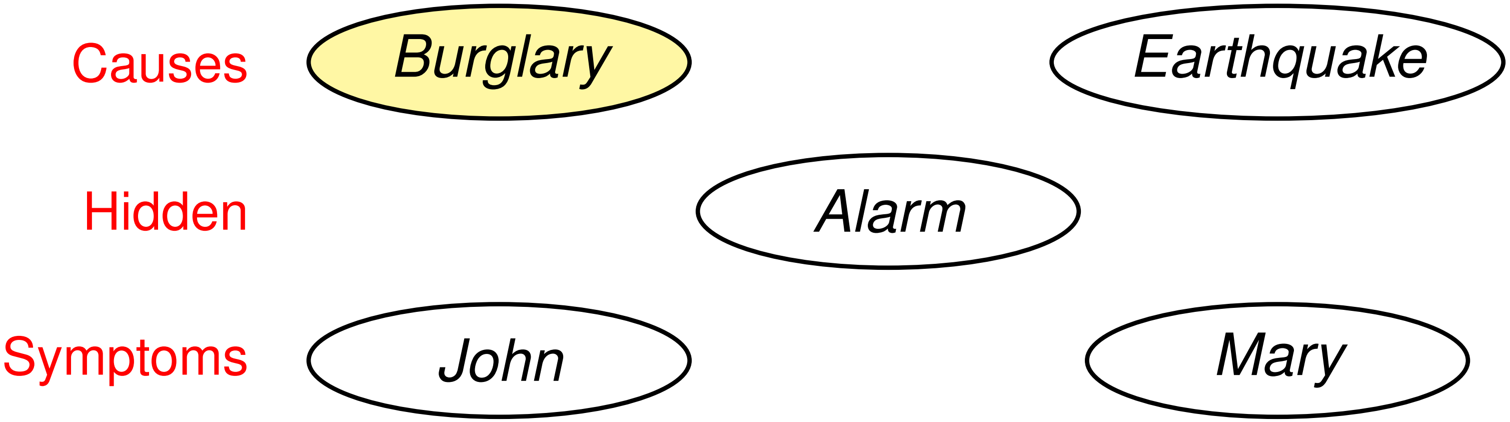
Step 1: Choose variables and an ordering. We choose $(X_1, X_2, X_3, X_4, X_5) = (\textit{Bur}, \textit{Ear}, \textit{Al}, J, M)$.



Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Step 2: Select a minimal parent set. Current node: $X_1 = \textit{Bur}$.
There are no earlier variables, so $\mathrm{Parents}(\textit{Bur}) = \varnothing$.

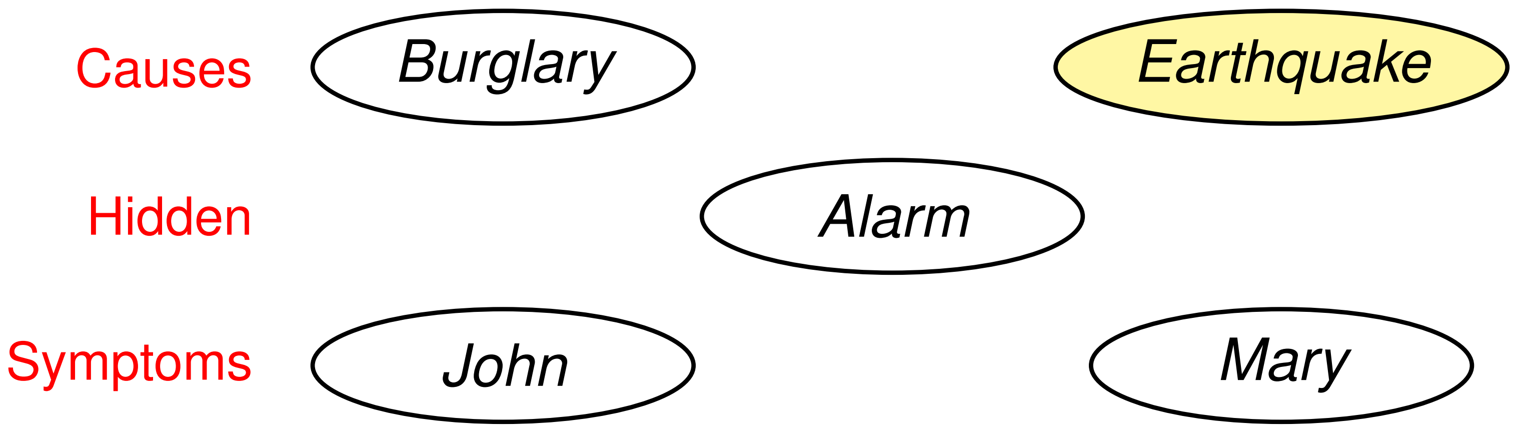
Step 3: Add the edges. Add no edge.



Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Step 2: Select a minimal parent set. Current node: $X_2 = \mathit{Ear}$. The only earlier variable is *Bur*. From the assumption that burglary and earthquake are independent [Ertel 2025], p. 162, *Bur* can be omitted. Hence $\mathrm{Parents}(\mathit{Ear}) = \varnothing$.

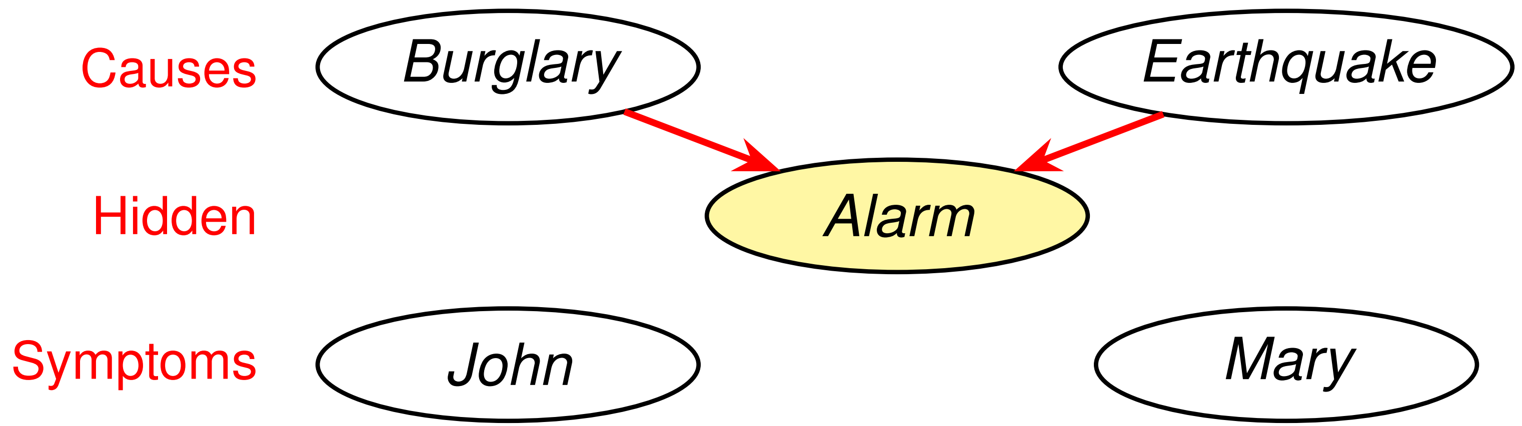
Step 3: Add the edges. Add no edge.



Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Step 2: Select a minimal parent set. Current node: $X_3 =$ *Al*. The earlier variables are *Bur* and *Ear*. The given values $P(\mathit{Al} \mid \mathit{Bur}, \mathit{Ear}) = 0.95$ and $P(\mathit{Al} \mid \mathit{Bur}, \neg\mathit{Ear}) = 0.94$ differ, so *Ear* cannot be omitted; $P(\mathit{Al} \mid \mathit{Bur}, \mathit{Ear}) = 0.95$ and $P(\mathit{Al} \mid \neg\mathit{Bur}, \mathit{Ear}) = 0.29$ differ, so *Bur* cannot be omitted [Ertel 2025], p. 162. Hence $\mathrm{Parents}(\mathit{Al}) = \{\mathit{Bur}, \mathit{Ear}\}$.

Step 3: Add the edges. Add $\mathit{Bur} \to \mathit{Al}$ and $\mathit{Ear} \to \mathit{Al}$.



Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

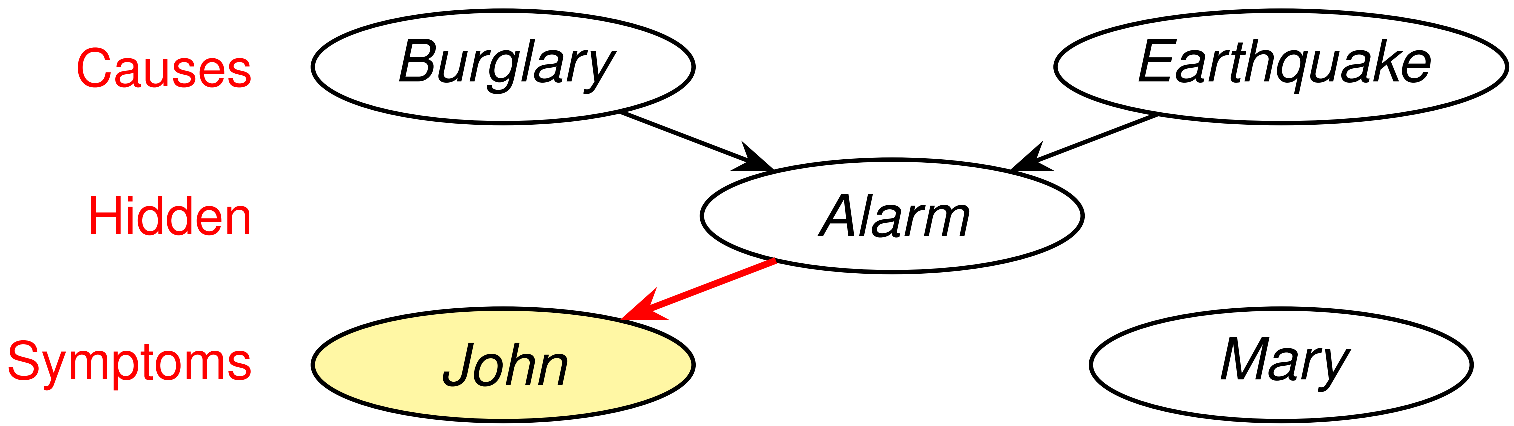
Semantics of Bayesian Networks

Summary

References

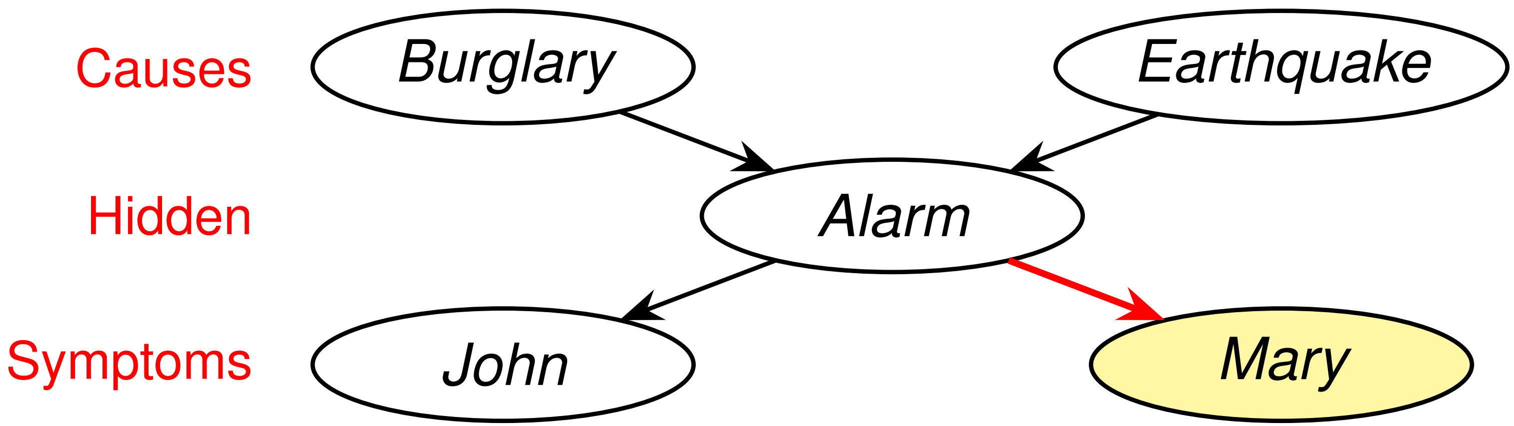
Step 2: Select a minimal parent set. Current node: $X_4 = J$. The earlier variables are *Bur*, *Ear*, and *Al*. From the assumption that John reacts only to the alarm, not directly to either cause [Ertel 2025], p. 169, *Bur* and *Ear* can be omitted. The given values $P(J \mid \textit{Al}) = 0.90$ and $P(J \mid \neg \textit{Al}) = 0.05$ differ, so *Al* cannot be omitted [Ertel 2025], p. 162. Hence $\mathrm{Parents}(J) = \{\textit{Al}\}$.

Step 3: Add the edges. Add $\textit{Al} \to J$.



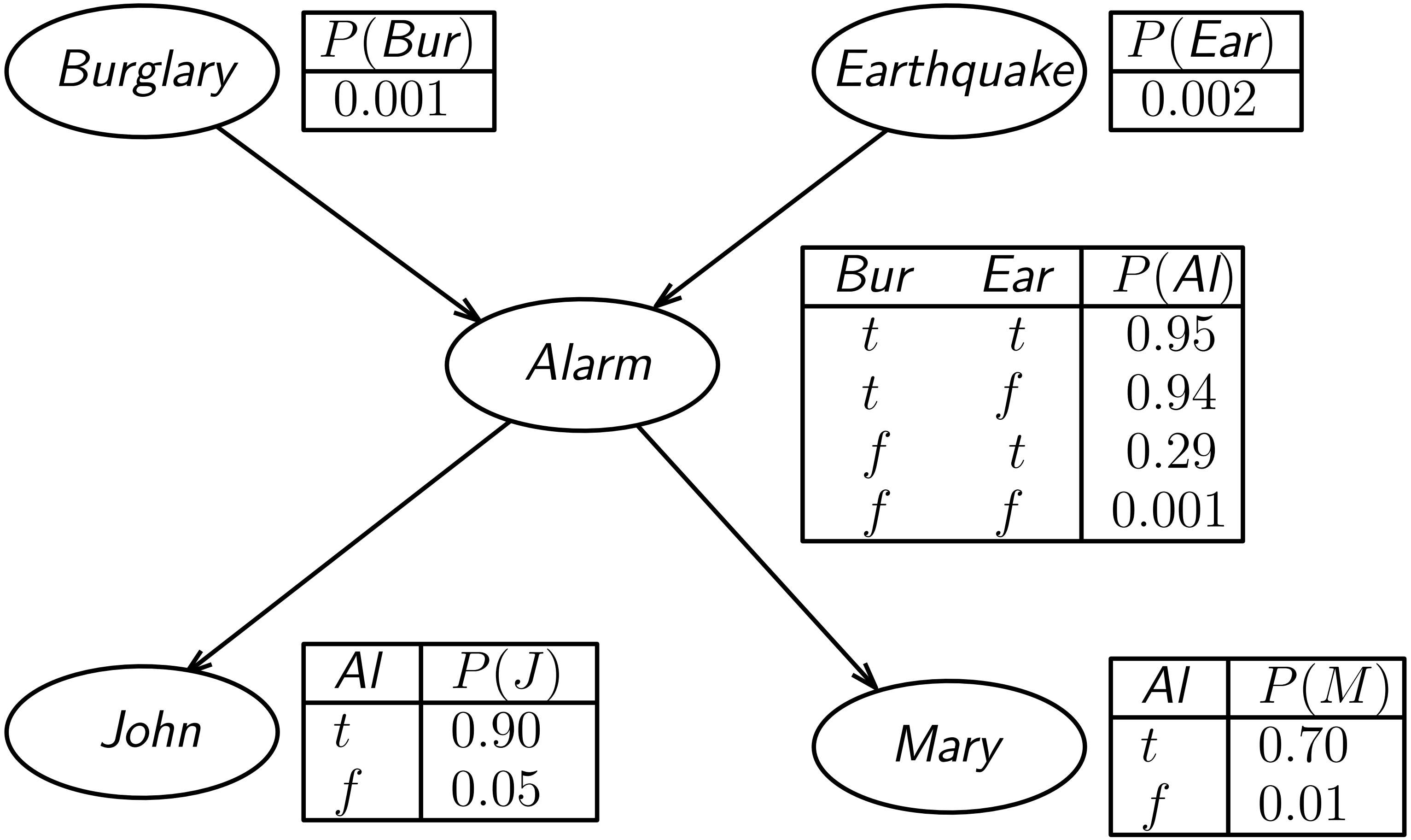
Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
116 Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Step 2: Select a minimal parent set. Current node: $X_5 = M$. Burglary and earthquake affect Mary's call only through the alarm, and given the alarm state, John's call gives no further information about Mary's call [Russell and Norvig 2021], p. 416. Thus $\mathbf{P}(M \mid \textit{Bur}, \textit{Ear}, \textit{Al}, J) = \mathbf{P}(M \mid \textit{Al})$, so *Bur*, *Ear*, and J can be omitted. Since $P(M \mid \textit{Al}) = 0.70 \neq 0.01 = P(M \mid \neg \textit{Al})$, *Al* cannot be omitted. Hence $\mathrm{Parents}(M) = \{\textit{Al}\}$. **Step 3: Add the edges.** Add $\textit{Al} \rightarrow M$.



Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Step 4: Specify the CPTs. Use the probabilities introduced earlier in the alarm example [Ertel 2025], p. 162:



$P(Bur)$
0.001

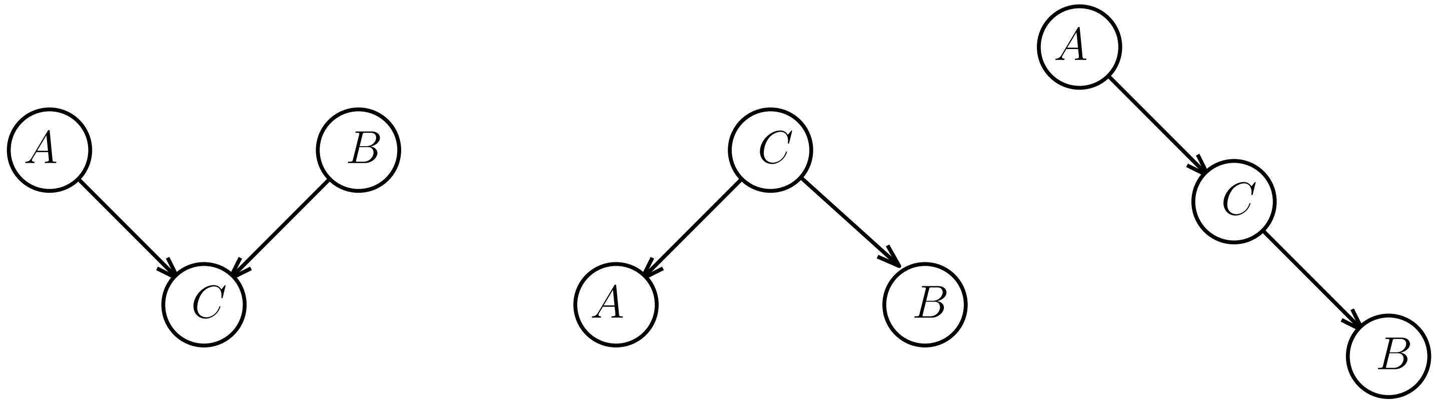
$P(Ear)$
0.002

Bur	Ear	$P(Al)$
t	t	0.95
t	f	0.94
f	t	0.29
f	f	0.001

Al	$P(J)$
t	0.90
f	0.05

Al	$P(M)$
t	0.70
f	0.01

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References



A and B are independent

A and B are independent given C

Figure: Independence patterns used when selecting parents: A and B are independent in the collider structure (left), and conditionally independent given C in the common-cause and chain structures (middle and right).

Alarm-network instances are *Bur* $\rightarrow$ *Al* $\leftarrow$ *Ear*, $J \leftarrow$ *Al* $\rightarrow M$, and *Bur* $\rightarrow$ *Al* $\rightarrow J$. Whether learning about one variable tells us anything about another depends on the whole network and on which other variables have already been observed.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
117 Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- A sparse Bayesian network can be *far more compact than the full joint distribution*.
- If each of n Boolean variables has at most k parents, then each CPT has at most 2^k free parameters and all CPTs together have at most $n2^k$.
- The full joint table has 2^n entries, of which $2^n - 1$ are free parameters.
- Example: for $n = 30$ and $k = 5$,

$$n2^k = 960, \qquad 2^n - 1 = 1\,073\,741\,823.$$

- A complete ordered DAG has exactly $2^n - 1$ free CPT parameters because $\sum_{i=0}^{n-1} 2^i = 2^n - 1$.
- Thus, for fixed k, representation size grows linearly in n; exact inference can nevertheless require exponential time and space in the worst case.
- **Practical tradeoff:** small dependencies may be omitted when the added accuracy does not justify larger CPTs; for example, one might decline to add *Ear* $\to J$ and *Ear* $\to M$.

Reasoning with Uncertainty
Hoàng Anh Đức

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

- The alarm example used *one causes-before-effects ordering*; the construction algorithm *permits any variable ordering*.
- A *causes-before-effects ordering* often *gives fewer parents and makes the required probability assessments easier to understand*.
- A *poor ordering* may require *more edges and CPT parameters*, even though the resulting network can represent the same joint distribution.

Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

Semantics of Bayesian Networks

Summary

References

Having constructed the *graph and its CPTs*, we now state the *conditional-independence semantics encoded by the graph*.

Requirements

- A Bayesian network is a *directed acyclic graph (DAG)*.
- The *variables are numbered in a topological order*, so *every parent of X_i has an index smaller than i*.

For this topological ordering, the *parent-selection condition in Step 2* gives, for each X_i whenever the conditional probabilities are defined,

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

Thus, given its *parents*, X_i is *conditionally independent of the set of its earlier non-parent variables*.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

Semantics of Bayesian Networks

Summary

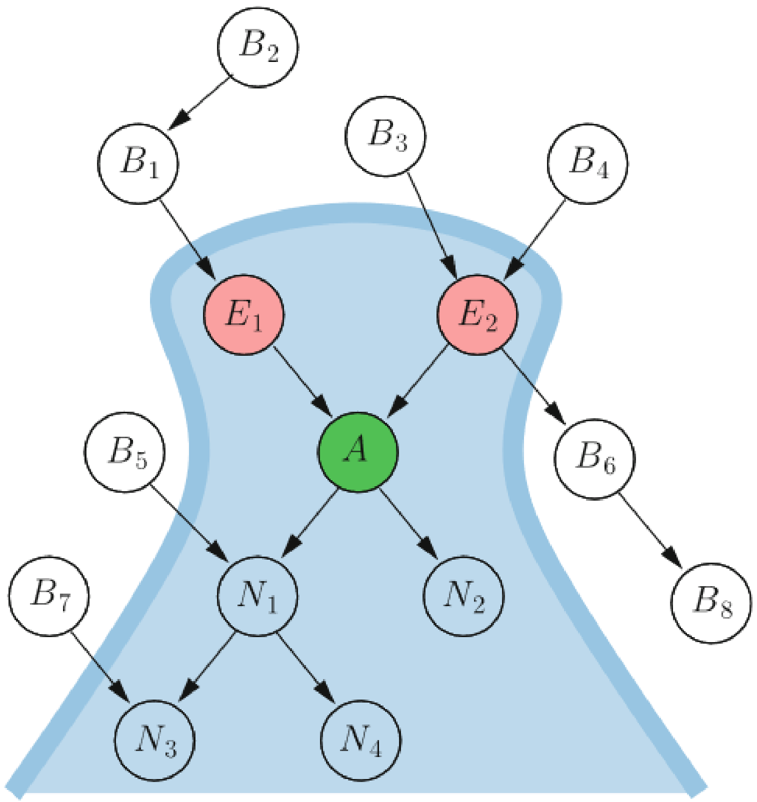
References

More generally [Ertel 2025], p. 170:

Theorem 7

Given its parents, a node in a Bayesian network is conditionally independent of the set of all other nodes that are neither its descendants nor its parents.

Example of conditional independence in a Bayesian network. Given its *parents* E_1 and E_2, A is *conditionally independent of the set of non-descendant nodes* $B_1, \ldots, B_8$.



Reasoning with Uncertainty

Hoàng Anh Đức

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Independent Variables

Graphical Representation of Knowledge as a Bayesian Network

Conditional Independence

Practical Application

Software for Bayesian Networks

Development of Bayesian Networks

121 Semantics of Bayesian Networks

Summary

References

- The pointwise chain rule and the parent-selection condition yield the *Bayesian-network factorization*; the CPTs supply a row for every parent assignment:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid \text{Parents}(X_i)).$$

- Applying this factorization to the completed alarm network gives

$$\begin{aligned}&\mathbf{P}(\mathit{Bur}, \mathit{Ear}, \mathit{Al}, J, M)\\ &\quad = \mathbf{P}(\mathit{Bur})\mathbf{P}(\mathit{Ear})\mathbf{P}(\mathit{Al} \mid \mathit{Bur}, \mathit{Ear})\mathbf{P}(J \mid \mathit{Al})\mathbf{P}(M \mid \mathit{Al}).\end{aligned}$$

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
122 Semantics of Bayesian Networks
Summary
References

Basics of Bayesian Networks

- A *Bayesian network* is defined by:
  - A set of variables and a set of directed edges between these variables.
  - Each variable has finitely many possible values.
  - The variables together with the edges form a directed acyclic graph (DAG), that is, a directed graph with no directed cycle.
  - For every variable A the CPT (that is, the table of conditional probabilities $\mathbf{P}(A \mid \textit{Parents}(A))$) is given.
- Two variables A and B are called *conditionally independent* given C if, for every a, b, c with $P(C = c) > 0$, $P(A = a, B = b \mid C = c) = P(A = a \mid C = c)P(B = b \mid C = c)$.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Probability Rules and Their Domains

- *Bayes' theorem.* If $P(A) > 0$ and $P(B) > 0$, then

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

- *Marginalization.* Always,

$$P(B) = P(A, B) + P(\neg A, B).$$

If $0 < P(A) < 1$, then also

$$P(B) = P(B \mid A)P(A) + P(B \mid \neg A)P(\neg A).$$

- *Conditioning.* If $P(B) > 0$ and write c for the event $C = c$ and let $S_B := \{c : P(B, c) > 0\}$. Then

$$P(A \mid B) = \sum_{c \in S_B} P(A \mid B, c)P(c \mid B).$$

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Basics of Bayesian Networks (cont.)

- In a topological order, the parent-selection condition is

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

whenever these conditionals are defined.

- A causes-before-effects ordering often yields fewer parents and simpler CPTs.
- *Bayesian-network factorization:*

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid \mathrm{Parents}(X_i)).$$

The CPTs supply a row for every parent assignment.

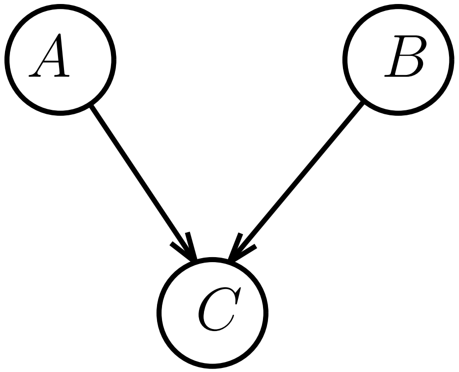
Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Exercise 13 ([Ertel 2025], Exercise 7.10, p. 175)

Given a Bayesian network with the three binary variables A, B, C and $P(A) = 0.2$, $P(B) = 0.9$, as well as the CPT shown below:

(a) Compute $P(A \mid B)$.
(b) Compute $P(C \mid A)$.

A	B	$P(C)$
t	f	0.1
t	t	0.2
f	t	0.9
f	f	0.4



Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
126 Semantics of Bayesian Networks
Summary
References

Exercise 14 ([Ertel 2025], Exercise 7.11, p. 175)

For the alarm example (Example 14), calculate the following conditional probabilities:

(a) Calculate the a priori probabilities $P(\textit{Al})$, $P(J)$, $P(M)$.

(b) Calculate $P(M \mid \textit{Bur})$ using the product rule, marginalization, the chain rule, and conditional independence.

(c) Use Bayes' formula to calculate $P(\textit{Bur} \mid M)$.

(d) Compute $P(\textit{Al} \mid J, M)$ and $P(\textit{Bur} \mid J, M)$.

(e) Show that the variables J and M are not independent.

(f) Check all of your results with JavaBayes and with PIT. (The PIT service is *no longer active*. Check with another Bayesian-network tool or by direct calculation.)

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
127 Semantics of Bayesian Networks
Summary
References

Exercise 15 ([Ertel 2025], Exercise 7.11 (cont.), p. 175)

(g) Design a Bayesian network for the alarm example, but with the altered variable ordering M, *Al*, *Ear*, *Bur*, J. According to the semantics of Bayesian networks, only the necessary edges must be drawn in. (**Hint:** the variable order given here does NOT represent causality. Thus it will be difficult to intuitively determine conditional independences.)

(h) In the original Bayesian network of the alarm example, the earthquake node is removed. Which CPTs does this change? (Why these in particular?)

(i) Calculate the CPT of the alarm node in the new network.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
Semantics of Bayesian Networks
Summary
References

Exercise 16 ([Ertel 2025], Exercise 7.12, p. 176)

A diagnostic system is to be made for a dynamo-powered bicycle light using a Bayesian network. The variables in the following table are given.

Abbr.	Meaning	*Values*
Li	Light is on	t/f
Str	Street condition	*dry*, *wet*, *snow*
Flw	Dynamo flywheel worn out	t/f
R	Dynamo sliding	t/f
V	Dynamo shows voltage	t/f
B	Light bulb o.k.	t/f
K	Cable o.k.	t/f

The following variables are pairwise independent: *Str*, *Flw*, B, K. Furthermore: (R, B), (R, K), (V, B), (V, K) are independent and the following equation holds:

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Independent Variables
Graphical Representation of Knowledge as a Bayesian Network
Conditional Independence
Practical Application
Software for Bayesian Networks
Development of Bayesian Networks
129 Semantics of Bayesian Networks
Summary
References

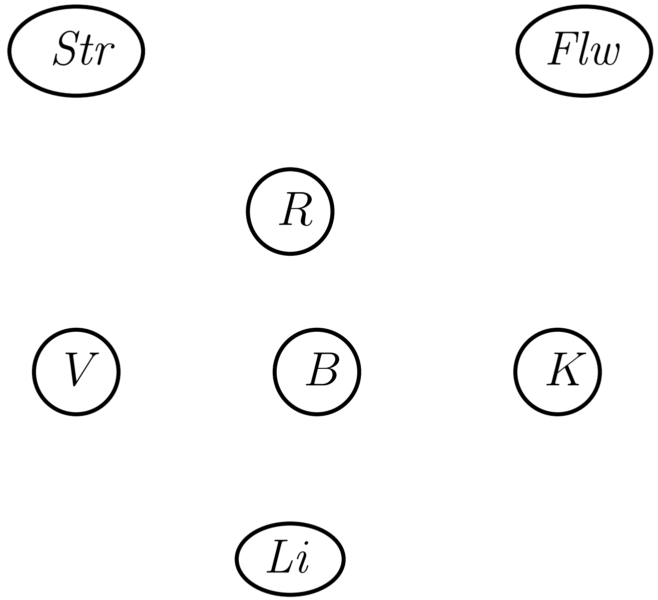
$$\mathbf{P}(\mathit{Li} \mid V, R) = \mathbf{P}(\mathit{Li} \mid V)$$
$$\mathbf{P}(V \mid R, \mathit{Str}) = \mathbf{P}(V \mid R)$$
$$\mathbf{P}(V \mid R, \mathit{Flw}) = \mathbf{P}(V \mid R)$$

(a) Draw all of the edges into the graph (taking causality into account).

(b) Enter all missing CPTs into the graph (table of conditional probabilities). Freely insert plausible values for the probabilities.

(c) Show that the network does not contain an edge $(\mathit{Str}, \mathit{Li})$.

(d) Compute $P(V \mid \mathit{Str} = \mathit{snow})$.



V	B	K	$P(Li)$
t	t	t	0.99
t	t	f	0.01
t	f	t	0.01
t	f	f	0.001
f	t	t	0.3
f	t	f	0.005
f	f	t	0.005
f	f	f	0

Summary

- *Probabilistic logic* for reasoning under uncertain knowledge.
- Method of *maximum entropy* models *non-monotonic reasoning*.
- Bayesian networks as *special case* of MaxEnt.
- Bayesian networks rely on *independence assumptions*.
- In a Bayesian network, all CPTs must be *filled completely*.
- With MaxEnt, *arbitrary knowledge* can be formulated.
  - E.g.: “I am pretty sure that A is true.”: $P(A) \in [0.6, 1]$.
  - The freedom that the developer has when modeling with MaxEnt can be a *disadvantage* (especially for a beginner) because, in contrast to the Bayesian approach, it is not necessarily clear what knowledge should be modeled.

Summary

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
132 Summary
References

- Combining MaxEnt and Bayesian networks
  - Begin by building a network according to the Bayesian methodology: enter the *required edges*, then *fill the CPTs with values*.
  - If certain values for the CPTs are unavailable, then they can be replaced with *intervals* or by other probabilistic logic formulas.
  - Such a network no longer has the *special semantics* of a Bayesian network. It must be processed and completed by a MaxEnt system.
- Arbitrary rule sets may be *inconsistent*: $P(A) = 0.7$ and $P(A) = 0.8$.
- The historical PIT system could *detect* inconsistent rule sets.
- In some cases reasoning is possible anyway.

Summary

- Medical expert system LEXMED
  - can be modeled and implemented using MaxEnt and Bayesian networks
- In a prospective study of 185 cases, LEXMED achieved a *sensitivity of 88%* and a *specificity of 87%* at $k = 0.6$. This is a very good result from a *small study*. [Ertel 2025], pp. 158–160
- Scores are equivalent to the special case *Naive-Bayes*, that is, to the assumption that all symptoms are *conditionally independent* given the diagnosis.
- In the LEXMED example we showed that it is possible to build an expert system for reasoning under uncertainty that is capable of *discovering (learning) knowledge* from the data in a database

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

133 Summary

References

Summary

- Nowadays, *Bayesian inference* is very important and well-developed
- We have completely left out the handling of *continuous variables*.
- For the case of *normally distributed random variables* there are procedures and systems.
- For arbitrary distributions, however, the *computational complexity* is a big problem.
- In addition to the directed networks that are heavily based on *causality*, there are also *undirected networks*.

Additional Materials
Introduction
Computing with Probabilities
The Principle of Maximum Entropy
LEXMED
Reasoning with Bayesian Networks
Summary
References

References

- Ertel, Wolfgang (2025). *Introduction to Artificial Intelligence*. 3rd. Springer. DOI: 10.1007/978-3-658-43102-0.
- Russell, Stuart J. and Peter Norvig (2021). *Artificial Intelligence: A Modern Approach*. 4th. Pearson.
- Strang, Gilbert and Edwin Herman (Mar. 30, 2016). *Calculus Volume 3: Lagrange Multipliers*. OpenStax, Rice University. URL: https://openstax.org/books/calculus-volume-3/pages/4-8-lagrange-multipliers (visited on 08/14/2026).
- Russell, Stuart J. and Peter Norvig (2010). *Artificial Intelligence: A Modern Approach*. 3rd. Pearson.
- Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier. DOI: 10.1016/C2009-0-27609-4.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

135 References

References (cont.)

Shortliffe, Edward (1976). *Computer-based medical consultations: MYCIN*. Elsevier.

Additional Materials

Introduction

Computing with Probabilities

The Principle of Maximum Entropy

LEXMED

Reasoning with Bayesian Networks

Summary

References