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

Contents

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)

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

$\Rightarrow$ *The knowledge base is inconsistent*. (Because the logic is monotonic; i.e., new knowledge can not void old knowledge.)
$\Rightarrow$ *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.

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*.

Let's just sit back and think about what to do!

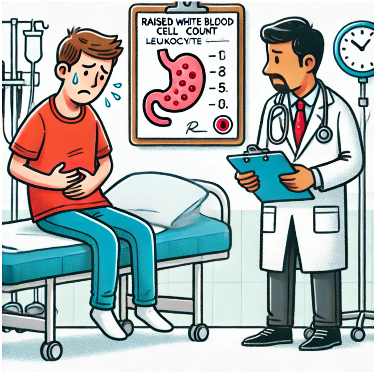
Example 1

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 $\land$ Leukocytes $> 10000 \rightarrow$ Appendicitis

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

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



Introduction

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* $A \rightarrow B$ is assigned a certain *factor* β.
- $A \rightarrow_{\beta} B$ means *the conditional probability* $P(B \mid A) = \beta$.
  - Stomach pain right lower $\wedge$ Leukocytes $> 10000 \rightarrow_{0.6}$ Appendicitis
- Formulas for connecting the factors of rules
- Calculus is incorrect
- Inconsistent results could be derived

Introduction

- There were also (unsuccessful) attempts to solve this problem by using *non-monotonic logic* and *default logic*.
  - Default logic: a special type of non-monotonic logic.
  - In the Flying Penguin Example, the first three given rules are default. The newly added rule *penguin*$(x) \Rightarrow \neg$*fly*(x) "override" the default rule *bird*$(x) \Rightarrow$ *fly*(x) in the case of penguin, and for other birds the default rule still applies.
- *Dempster-Schäfer theory:* assigns a belief function *Bel*(A) to a logical proposition A, whose value gives the degree of evidence for the truth of A.
- *Fuzzy logic:* demonstrates considerable weaknesses when reasoning under uncertainty in more complex applications.
  - The meaning (semantics) of a proposition in fuzzy logic is not clearly defined.

- *Conditional probabilities instead of implication* (as it is known in logic)
  - Significantly better in modeling everyday causal reasoning.
- *Subjective probabilities*
  - For example, if you are in the middle of the street and do not know whether you should turn left or right. (That is, the probabilities of turning left and turning right are unknown.)
  - From mathematical viewpoint, if you don't know the probabilities, you do nothing.
  - From AI viewpoint, you need to make a decision. So (even if you don't know anything) you "assume" that turning left and right have the same probability 0.5 and make a decision based on this "assumption".
  - *The "assumption" you made may not be true but it is subjective to you*.
- *Probability theory* is well-founded.
- Reasoning with uncertain and incomplete knowledge.
  - Maximum entropy method (MaxEnt) and the medical expert system LEXMED.
  - Bayesian networks.

Computing with Probabilities

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:* $\emptyset$.

Computing with Probabilities

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

- A, B, etc.: *random variables*.
- We consider only *discrete random variables with finite value range*.
- Example:
  - The variable *face_number* for a dice roll is discrete with the values $1, 2, 3, 4, 5, 6$.
  - The probability of rolling a five or a six is equal to $1/3$.

$$P(\textit{face\_number} \in \{5, 6\})$$
$$= P(\textit{face\_number} = 5 \vee \textit{face\_number} = 6) = 1/3.$$

Computing with Probabilities

Definition

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be finite. There is no preferred elementary event, which means that we assume a symmetry related to the frequency of how often each elementary event appears. The *probability* $P(A)$ of the event A is then

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of favorable cases for } A}{\text{Number of possible cases}}.$$

Example 3

Throwing a die, the probability for an even number is

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

Computing with Probabilities

- Any elementary event has the probability $1/|\Omega|$ (*Laplace assumption*).
- Applicable only at *finite event sets*.
- To *describe events* we use *variables with the appropriate number of values*.
  - Example: Variable *eye_color* can take on the values *green*, *blue*, *brown*.
  - *eye_color* $=$ *blue* then describes an event because we are dealing with a proposition with the truth values t or f.
- *Binary (boolean) variables* (i.e., variables that can take on the values t and f) are propositions themselves.
  - Write $P(\mathit{JohnCalls})$ instead of $P(\mathit{JohnCalls} = t)$.

Computing with Probabilities

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. 171)

Prove the propositions from Theorem 1.

Computing with Probabilities

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

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 $n^d - 1$ unique values.

Example 4

On Landsdowne 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*.

The probability for *speeding under the condition that the driver is a student*:

$$P(V \mid S) = \frac{|\text{Driver is a student and speeding}|}{|\text{Driver is a student}|} = \frac{5}{30} = \frac{1}{6} \approx 0.17.$$

The probability for *speeding in general*: $P(V) = 0.1$.

Thus, students speed more frequently than the average person.

Definition

For two events A and B, the probability $P(A \mid B)$ for A under the condition B (*conditional probability*) is defined by

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

$P(A \mid B) =$ probability of A regarding event B only, i.e.

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

Indeed, this can be proved as follows.

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

Computing with Probabilities
Conditional Probability

Definition

If, for two events A and B, $P(A \mid B) = P(A)$, then these events are called *independent*. In other words, A and B are independent if *the probability of the event A is not influenced by the event B*.

Theorem 2

For independent events A and B, it follows from the definition that $P(A \wedge B) = P(A) \cdot P(B)$.

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, *dice 2 is always the same as dice 1*, $P(D_1 = 6 \wedge D_2 = 6) = \frac{1}{6}$.

Exercise 2

Prove Theorem 2.

- *Product Rule:* For two events A and B,
  $P(A \wedge B) = P(A \mid B) \cdot P(B)$.
- *Chain Rule:* For random variables $X_1, \ldots, X_n$,

$$
\begin{aligned}
\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1}) \cdot \mathbf{P}(X_1, \ldots, X_{n-1}) \\
&= \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1}) \cdot \mathbf{P}(X_{n-1} \mid X_1, \ldots, X_{n-2}) \\
&\cdot \mathbf{P}(X_1, \ldots, X_{n-2}) \\
&= \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1}) \cdot \mathbf{P}(X_{n-1} \mid X_1, \ldots, X_{n-2}) \\
&\cdot \mathbf{P}(X_1, \ldots, X_{n-2}) \cdot \ldots \cdot \mathbf{P}(X_n \mid X_1) \cdot \mathbf{P}(X_1) \\
&= \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}).
\end{aligned}
$$

(Because the chain rule holds for all values of the (random) variables $X_1, \ldots, X_n$, it has been formulated for the distribution using the symbol $\mathbf{P}$.)

Reasoning with Uncertainty
Hoàng Anh Đức

Example 6

- For $n = 3$ events, the chain rule for events A_1, A_2, A_3 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

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

Example 7

- For $n = 2$ random variables X_1 and X_2 which can both take values t and f, for example, we have

$$P(X_1 = t, X_2 = t) = P(X_1, X_2)$$
$$= P(X_1 = t \mid X_2 = t) \cdot P(X_2 = t) = P(X_1 \mid X_2) \cdot P(X_2),$$
$$P(X_1 = t, X_2 = f) = P(X_1, \neg X_2)$$
$$= P(X_1 = t \mid X_2 = f) \cdot P(X_2 = f) = P(X_1 \mid \neg X_2) \cdot P(\neg X_2).$$

- 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

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

Computing with Probabilities
Conditional Probability

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

$$
\begin{aligned}
P(A) &= P((A \wedge B) \vee (A \wedge \neg B)) \\
&= P(A \wedge B) + P(A \wedge \neg B).
\end{aligned}
$$

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

- In general,

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

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*.

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:

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

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

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \text{ as well as } P(B \mid A) = \frac{P(A \wedge B)}{P(A)}.$$

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)

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

- Assuming that *appendicitis affects the biology of all humans the same*, regardless of ethnicity.
- $P(\mathit{Leuko} \mid \mathit{App})$ is a universal value that is *valid worldwide*.
- $P(\mathit{App} \mid \mathit{Leuko})$, on the other hand, is not universal, because this value is *influenced by the a priori probabilities $P(\mathit{App})$ and $P(\mathit{Leuko})$*. Each of these can *vary according to on's life circumstances*.
  - For example, $P(\mathit{Leuko})$ is dependent on whether a population has a high or low rate of exposure to infectious diseases. In the tropics, this value can differ significantly from that of cold regions.
- *Bayes' theorem*, however, makes it easy for us to *take the universally valid value $P(\mathit{Leuko} \mid \mathit{App})$*, and *compute $P(\mathit{App} \mid \mathit{Leuko})$ which is useful for diagnosis*.

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!*

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 $P(B) = 0.001$.
  - 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 $P(A) = 0.1$.
  - 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!*

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
26 Conditional Probability
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
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
124

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 $P(B) = 0.001$.
  - 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 $P(A) = 0.1$.
  - 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 $P(A) = P(A \mid B) \cdot P(B) + P(A \mid \neg B) \cdot P(B) = 0.00099 + P(A \mid \neg B) \cdot 0.999 = 0.1$, 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

Exercise 4 ([Ertel 2025], Exercise 7.2, p. 171)

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 $P(y > 3 \mid \mathit{Class} = \mathit{good})$ and $P(y \leq 3 \mid \mathit{Class} = \mathit{good})$. Then use Bayes' formula to determine $P(\mathit{Class} = \mathit{good} \mid y > 3)$ and $P(\mathit{Class} = \mathit{good} \mid y \leq 3)$.

(b) Which of the probabilities computed in subproblem (a) contradicts the statement "All good peas are heavier than 3 grams"?

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.

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

$$H(\mathbf{P}) = -\sum_{i=1}^{n} p_i \log p_i$$

- 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.)

- Modus Ponens

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

- Generalization to probability rules

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

Given: two probability values α, β, **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*.

- We cannot make an exact statement about $P(B)$ with classical probability theory, but at the most we can *estimate* $P(B) \geq P(B \mid A) \cdot P(A)$.

The Principle of Maximum Entropy
An Inference Rule for Probabilities

- Distribution

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

- Abbreviation

$$p_1 = P(A, B)$$
$$p_2 = P(A, \neg B)$$
$$p_3 = P(\neg A, B)$$
$$p_4 = P(\neg A, \neg B)$$

- 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.

- 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!

The Principle of Maximum Entropy
An Inference Rule for Probabilities

- To come to a definite solution despite this missing knowledge, we change our point of view. We use the given equation as a constraint for the solution of an optimization problem.
- **Find:** Distribution $\mathbf{p} = (p_3, p_4)$ which maximizes the entropy

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i = -p_3 \ln p_3 - p_4 \ln p_4$$

under the constraint $p_3 + p_4 = 1 - \alpha$.
- 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*.
  - *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*.

The Principle of Maximum Entropy
An Inference Rule for Probabilities

- **Problem:** Maximizing $H(\mathbf{p}) = -p_3 \ln p_3 - p_4 \ln p_4$ w.r.t the constraint $p_3 + p_4 - 1 + \alpha = 0$.
- Method of Lagrange multipliers.
- Lagrange function:

$$\begin{aligned} L &= H(\mathbf{p}) + \lambda(p_3 + p_4 - 1 + \alpha) \\ &= -p_3 \ln p_3 - p_4 \ln p_4 + \lambda(p_3 + p_4 - 1 + \alpha) \end{aligned}$$

- Taking the partial derivatives with respect to p_3 and p_4

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

- 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 $p_3 = p_4 = (1 - \alpha)/2$.

The Principle of Maximum Entropy
An Inference Rule for Probabilities

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

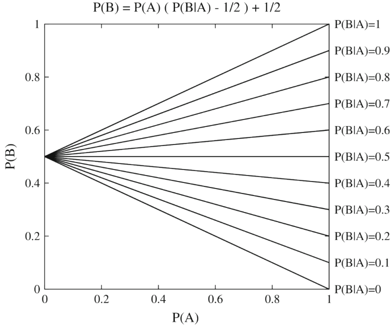


Figure: Curve array for $P(B)$ as a function of $P(A)$ for different values of $P(B \mid A)$.

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 $P(A) = 1$ and $P(B \mid A) = 1$, $P(B) = 1$ (Modus Ponens holds)

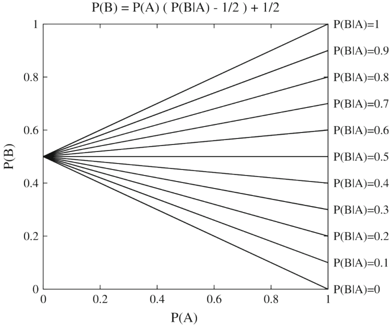


Figure: Curve array for $P(B)$ as a function of $P(A)$ for different values of $P(B \mid A)$.

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

- When $P(A) = 1$ and $P(B \mid A) = 1$, $P(B) = 1$ (Modus Ponens holds)
- When $P(A) = 1$ and $P(B \mid A) = 0$, $P(B) = 0$ (Modus Ponens holds)

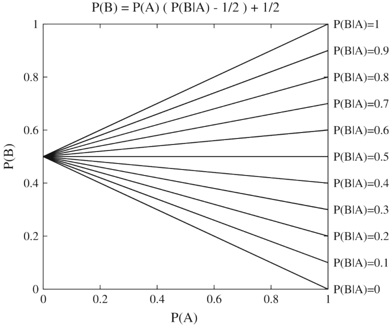


Figure: Curve array for $P(B)$ as a function of $P(A)$ for different values of $P(B \mid A)$.

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

- When $P(A) = 1$ and $P(B \mid A) = 1$, $P(B) = 1$ (Modus Ponens holds)
- When $P(A) = 1$ and $P(B \mid A) = 0$, $P(B) = 0$ (Modus Ponens holds)
- When $P(A) = 0$, Modus Ponens cannot be applied, but our formula results in the value $1/2$ for $P(B)$ irrespective of $P(B \mid A)$. When A is false, we know nothing about B, which reflects our intuition exactly.

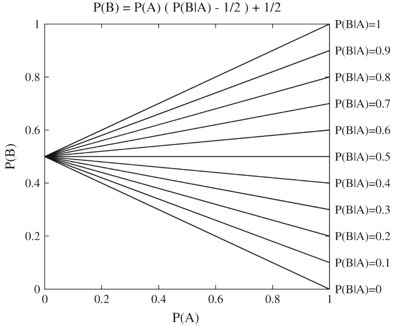


Figure: Curve array for $P(B)$ as a function of $P(A)$ for different values of $P(B \mid A)$.

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

- When $P(A) = 1$ and $P(B \mid A) = 1$, $P(B) = 1$ (Modus Ponens holds)
- When $P(A) = 1$ and $P(B \mid A) = 0$, $P(B) = 0$ (Modus Ponens holds)
- When $P(A) = 0$, Modus Ponens cannot be applied, but our formula results in the value $1/2$ for $P(B)$ irrespective of $P(B \mid A)$. When A is false, we know nothing about B, which reflects our intuition exactly.

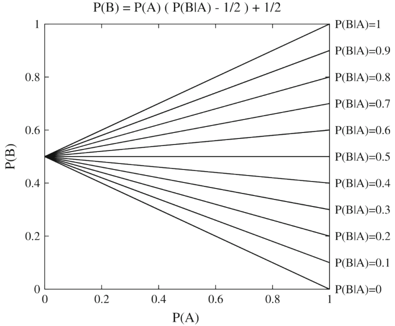


Figure: Curve array for $P(B)$ as a function of $P(A)$ for different values of $P(B \mid A)$.

- We cannot make any prediction about B when $P(B \mid A) = 1/2$.

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*.
- A calculus, which leads to distributions with a higher entropy is adding informations ad hoc, which again is not justified.

- 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 $\{p_{i_1}, p_{i_2}, \dots, p_{i_k}\}$ is indifferent, then the maximum of the entropy under the given constraints is at the point where $p_{i_1} = p_{i_2} = \cdots = p_{i_k}$.

- 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.
- Thus, $P(A, B) = P(B \mid A)P(A) = \beta P(A)$. Therefore, $p_1 = \beta(p_1 + p_2)$.
- We derived two constraints:

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

- No symbolic solutions!
- Solving the Lagrange equations numerically.

Exercise 5 ([Ertel 2025], Exercise 7.3, p. 172)

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

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

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

Exercise 6 ([Ertel 2025], Exercise 7.4, p. 172)

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. 172)

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*.

A	B	$A \Rightarrow B$	$P(A)$	$P(B)$	$P(B \mid A)$
t	t	t	1	1	1
t	f	f	1	0	0
f	t	t	0	1	Undefined
f	f	t	0	0	Undefined

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

- $p_1 = P(A, B)$, $p_2 = P(A, \neg B)$, $p_3 = P(\neg A, B)$, $p_4 = P(\neg A, \neg B)$.
- 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. 173)

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

- From definition, $P(B \mid A) = \dfrac{P(A, B)}{P(A)}$.
- For $P(A) \neq 0$, we have $P(B \mid A) = P(B)$.
- For $P(A) = 0$, $P(B \mid A)$ stays undefined.

A	B	$A \Rightarrow B$	$P(A)$	$P(B)$	$P(B \mid A)$
f	t	t	α	γ	γ
f	f	t	0	γ	Undefined

The Principle of Maximum Entropy

MaxEnt-Systems

- Often, MaxEnt optimization has no symbolic solution.
- Therefore: *numerical entropy maximization*.
- SPIRIT (Symmetrical Probabilistic Intensional Reasoning in Inference Networks in Transition, `www.xspirit.de`): Fernuniversität Hagen.
- PIT (Probability Induction Tool, `http://www.maxent.de`): Munich Technical University.
- PIT uses Sequential Quadratic Programming (SQP) to find an extremum of the entropy function numerically.
- As input, PIT expects a file with 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]);
```

- Request `QP([B=t])`
- Web front end on `www.pit-systems.de` (Inactive)
- 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 $P(B \mid A) = 0.3$.

The Principle of Maximum Entropy

MaxEnt-Systems

- Request `QP([B=t])`
- Web front end on `www.pit-systems.de` (Inactive)
- 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 $P(B \mid A) = 0.3$.

Exercise 9 ([Ertel 2025], Exercise 7.6, p. 172)

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. 172)

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.

$$P(\textit{bird} \mid \textit{penguin}) = 1 \qquad \text{"penguins are birds"}$$
$$P(\textit{flies} \mid \textit{bird}) \in [0.95, 1] \qquad \text{"(almost all) birds can fly"}$$
$$P(\textit{flies} \mid \textit{penguin}) = 0 \qquad \text{"penguins cannot fly"}$$

- PIT input file:

```
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]);
```

- Answer

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

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
Conditional Probability
The Principle of Maximum Entropy
An Inference Rule for Probabilities
Maximum Entropy Without Explicit Constraints
Conditional Probability Versus Material Implication
MaxEnt-Systems
48 The Tweety example
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
124

The Principle of Maximum Entropy
The Tweety example

- Probability intervals are often very helpful
- second rule in the sense of "normally birds fly": $P(\textit{flies}| \mid \textit{bird}) \in (0.5, 1]$
- MaxEnt enables non monotonic inference
- MaxEnt is also successful on challenging benchmarks for non monotonic inference
- Application of MaxEnt within the medical expert system LEXMED

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.

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.

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.

- The most common serious cause of acute abdominal pain is appendicitis.
- Even today, *diagnosis can be difficult in many cases*.
  - Approx. 20% of the surgically removed appendixes without clinical abnormalities
  - 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*.

LEXMED

Appendicitis Diagnosis with Formal Methods

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 $S_1, S_2, \dots, S_n$, a score can formally be defined as

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

- 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).

LEXMED

Hybrid Probabilistic Knowledge Base

Query to the expert system: What is the probability for an inflamed appendix, if the patient is a 23 year old man with pain in the in downright stomach and a leukocyte value of 13000?

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 #.

LEXMED
Hybrid Probabilistic Knowledge Base

Query to the expert system:

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

Knowledge Base:

Database	Expert knowledge
15000 patients from Baden-Württemberg 1995	Dr. Rampf Dr. Hontschik

LEXMED

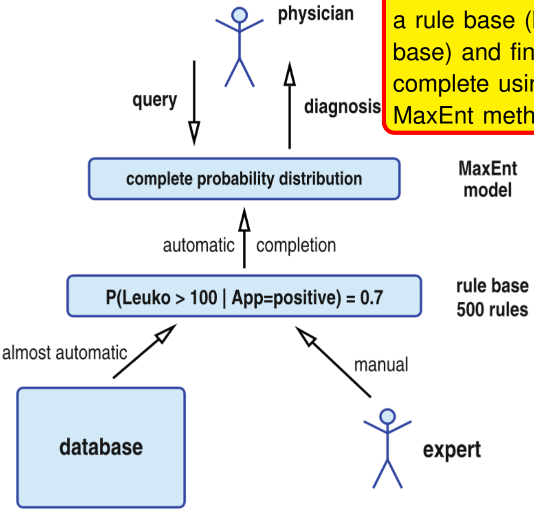
Hybrid Probabilistic Knowledge Base

- LEXMED calculates the probabilities of various diagnoses using the probability distribution of all relevant variables (see the previous table).
- The size of the distribution (that is, the size of the event space):

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

- Normalization condition $\Rightarrow$ $20\,643\,839$ independent values.
- Any rule set with less than $20\,643\,839$ probability values may not describe the event space 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.

Knowledge Processing:



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.

System Architecture:

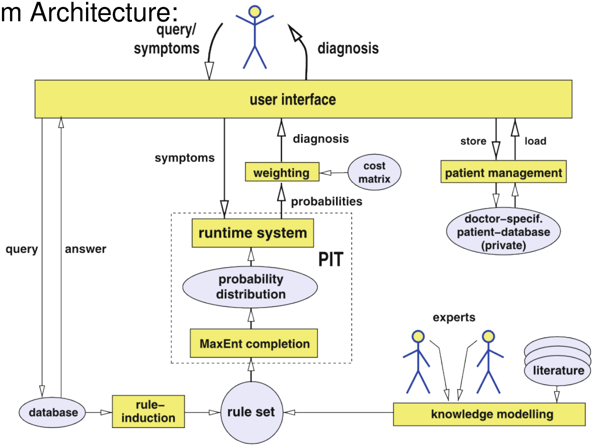


Figure: Rules are generated from the database as well as from expert knowledge. From these, MaxEnt creates a complete probability distribution. For a user query, the probability of every possible diagnosis is calculated. Using the cost matrix (will be defined later) a decision is then suggested.

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*.
  - Naturally the certainty of the diagnosis is higher when more symptom values are input.
- 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.,

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

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

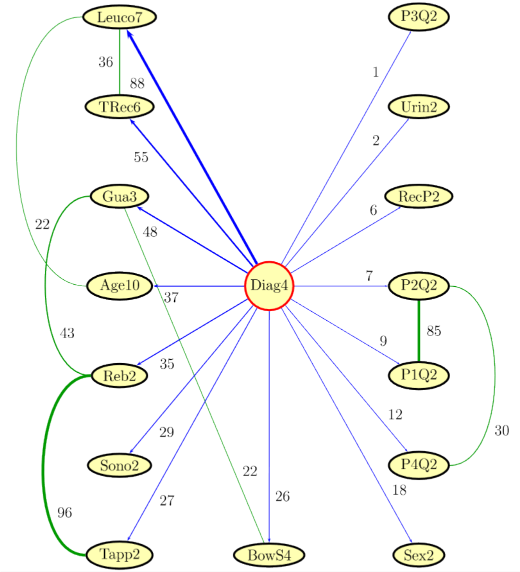
Learning of Rules by Statistical Induction

- *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.

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.)

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

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

Some of the LEXMED rules with probability intervals written in PIT syntax. “∗” stands for “∧” 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];
```

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.

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.

The complete probability model is calculated with the method of maximum entropy by the PIT-system.

LEXMED
Function of LEXMED

- 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).

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.

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

Cost matrix in the binary case

- Average additional cost for the two possible treatments:

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

- Multiply the vector $(k_2 p_2, k_1 p_1)^T$ by any scalar, say $1/k_1$, 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.

Exercise 11 ([Ertel 2025], Exercise 7.9, p. 173)

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.

To simplify the representation and make for a better comparison to similar studies, LEXMED was restricted to the two-value distinction between appendicitis and *NSAP*, as described before.

For each k $(0 \leq k < \infty)$, the *sensitivity* and *specificity* are measured against the test data

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

LEXMED

Performance

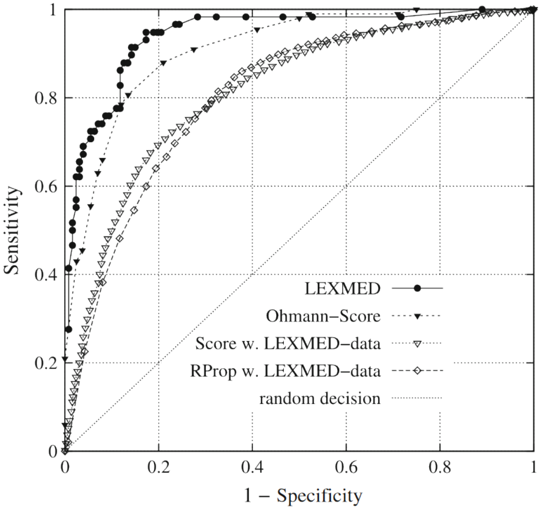


Figure: ROC curve from LEXMED compared with the Ohmann score and two additional models

- 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
- Diagnosis quality is comparable to an experienced surgeon
- Commercial marketing very difficult
- Wrong time?
- Wish of patients for personal care!
- Since de Dombal 1972, 39 years passed. Will it take another 39 years to make computer diagnostics become an established medical tool?

Reasoning with Bayesian Networks

- d variables $X_1, \ldots, X_d$ with n values each
- Probability distribution has $n^d - 1$ values.
- In practice *the distribution contains many redundancies*.
  $\Rightarrow$ It can be heavily reduced with the appropriate methods.
- *Bayesian networks* utilize knowledge about the independence of variables to simplify the model.

Reasoning with Bayesian Networks
Independent Variables

- Simplest case: all variables are pairwise independent

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

- Conditional probabilities become trivial:[1]

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

- 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] In the naive Bayes method, the independence of all attributes is assumed, and this method has been successfully applied to text classification.

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.

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

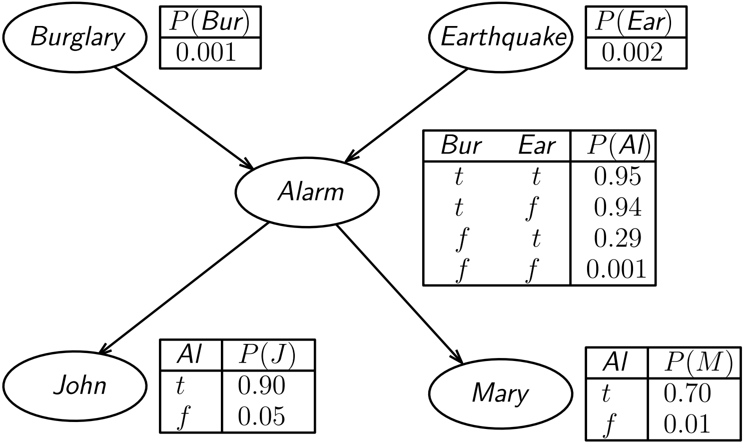
  - A priori probabilities: $P(Bur) = 0.001$, $P(Ear) = 0.002$. (*Bur* and *Ear* are independent.)
- *Requests:* $P(Bur \mid J \vee M)$, $P(J \mid Bur)$, $P(M \mid Bur)$

- A *Bayesian network* is a directed acyclic graph (DAG) in which
  - each node represents a random variable,
  - each edge $X_i \to X_j$ represents a *direct influence* of variable X_i on variable X_j, and
  - each node is associated with a conditional probability table (CPT) that quantifies the effects that the parents have on the node.
- 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 $X_1, \ldots, X_d$ can be expressed as

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

where $\text{Parents}(X_i)$ denotes the set of parent nodes of X_i in the graph.

Reasoning with Bayesian Networks



$P(Bur)$
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.

Definition

Two variables A and B are called *conditionally independent*, given C if

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

(This equation is *true for all combinations of values for all three variables* (that is, for the distribution).)

Remark

- independent $\not\Rightarrow$ conditional independent.
- conditional 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 coins: tossing a coin result heads 50% of the time. Toss these two 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**

- $\mathbf{P}(F, S) = \mathbf{P}(F) \cdot \mathbf{P}(S)$
  - As we toss two coins at the same time, the result of the first coin does not affect the result of the second coin and vice versa.
- $\mathbf{P}(F, S \mid E) \neq \mathbf{P}(F \mid E) \cdot \mathbf{P}(S \mid E)$
  - When you know that, say $E = t$ (i.e., the result of the two coins, by some magic power, must be the same), then knowing the result of the first coin tells you exactly the result of the second coin.

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

There are two biased coins: tossing *coin1* results heads 99% of the time and tossing *coin2* results tails 99% of the time. Choose one coin at random and toss it twice.

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**

- $\mathbf{P}(F, S \mid C) = \mathbf{P}(F \mid C) \cdot \mathbf{P}(S \mid C)$
  - If you already know which coin is taken, then knowing the result of the first toss does not help predicting the result of the second toss.
- $\mathbf{P}(F, S) \neq \mathbf{P}(F) \cdot \mathbf{P}(S)$
  - If you do not know which coin is taken, then knowing the result of the first toss is useful. For example, if the result of the first toss is head, then it is a strong evidence that you take *coin1*, and thus the second toss is unlikely to result head.

Example 17 (Alarm-Example (cont.))

- John and Mary independently react to an alarm.
  $\mathbf{P}(J, M \mid Al) = \mathbf{P}(J \mid Al) \cdot \mathbf{P}(M \mid Al)$.
- 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 suffices to show that the equation does not hold for one combination of values of J and M, say $P(J, M) \neq P(J) \cdot P(M)$. (More precisely, $P(J = t, M = t) \neq P(J = t) \cdot P(M = t)$.)
  - Calculate $P(Al)$ using the given probabilities, marginalization, and independence of *Bur* and *Ear*.
    (Result: $P(Al) \approx 0.00252$.)
  - Then calculate $P(J)$ and $P(M)$ using conditional probabilities and the computed $P(Al)$.
    (Result: $P(J) = 0.052$ and $P(M) = 0.0117$.)
  - Similarly, calculate $P(J, M)$ using conditional probabilities, conditional independence of J and M given *Al*.
    (Result: $P(J, M) \approx 0.002086$.)
  - Compare $P(J, M)$ and $P(J) \cdot P(M)$.

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), $\mathbf{P}(A, B, C) = \mathbf{P}(A \mid C)P(B \mid C)\mathbf{P}(C)$. Comparing with (a).

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

Reasoning with Bayesian Networks
Practical Application

Now we turn again to the *alarm example* and show *how the Bayesian network can be used for reasoning*.

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

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

Chain rule

J and *Bur* are independent given *Al*

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

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

Similarly, $P(\neg \textit{Al} \mid \textit{Bur}) = 0.06$.

Reasoning with Bayesian Networks

Practical Application

Therefore,

$$P(J \mid \textit{Bur}) = P(J \mid \textit{Al})P(\textit{Al} \mid \textit{Bur}) + P(J \mid \neg\textit{Al})P(\neg\textit{Al} \mid \textit{Bur})$$
$$= 0.9 \cdot 0.94 + 0.05 \cdot 0.06 = 0.849.$$

Analogously, $P(M \mid \textit{Bur}) = 0.659$.

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

$$\begin{aligned} P(J, M \mid \textit{Bur}) &= P(J, M \mid \textit{Al})P(\textit{Al} \mid \textit{Bur}) \\ &+ P(J, M \mid \neg\textit{Al})P(\neg\textit{Al} \mid \textit{Bur}) \\ &= P(J \mid \textit{Al})P(M \mid \textit{Al})P(\textit{Al} \mid \textit{Bur}) \\ &+ P(J \mid \neg\textit{Al})P(M \mid \neg\textit{Al})P(\neg\textit{Al} \mid \textit{Bur}) \\ &= 0.9 \cdot 0.7 \cdot 0.94 + 0.05 \cdot 0.01 \cdot 0.06 = 0.5922. \end{aligned}$$

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

$$P(J \vee M \mid \textit{Bur}) = P(\neg(\neg J \wedge \neg M) \mid \textit{Bur})$$
$$= 1 - P(\neg J, \neg M \mid \textit{Bur})$$
$$P(\neg J, \neg M \mid \textit{Bur}) = P(\neg J \mid \textit{Al}) P(\neg M \mid \textit{Al}) P(\textit{Al} \mid \textit{Bur})$$
$$+ P(\neg J \mid \neg \textit{Al}) P(\neg M \mid \neg \textit{Al}) P(\neg \textit{Al} \mid \textit{Bur})$$
$$= 0.1 \cdot 0.3 \cdot 0.94 + 0.95 \cdot 0.99 \cdot 0.06 = 0.085.$$
$$P(J \vee M \mid \textit{Bur}) = 1 - P(\neg J, \neg M \mid \textit{Bur})$$
$$= 1 - 0.085 = 0.915.$$

Bob thus receives a notification from either John or Mary for about 92% of all burglaries

Reasoning with Bayesian Networks
Practical Application

$$P(\textit{Bur} \mid J) = \frac{P(J \mid \textit{Bur})P(\textit{Bur})}{P(J)} = \frac{0.849 \cdot 0.001}{0.052} = 0.016$$

$$P(\textit{Bur} \mid M) = \frac{P(M \mid \textit{Bur})P(\textit{Bur})}{P(M)} = \frac{0.659 \cdot 0.001}{0.0117} = 0.056$$

$$P(\textit{Bur} \mid J, M) = \frac{P(J, M \mid \textit{Bur})P(\textit{Bur})}{P(J, M)} = \frac{0.5922 \cdot 0.001}{0.002086} = 0.284.$$

- If John calls, the probability of a burglary is 1.6%. If Mary calls, it is 5.6%, which is about five times higher than John.
  $\Rightarrow$ Significantly higher confidence given a call from Mary.
- Bob should only be seriously concerned about his home if both of them call, as the probability of a burglary in that case is 28.4%.

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

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

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]);
```

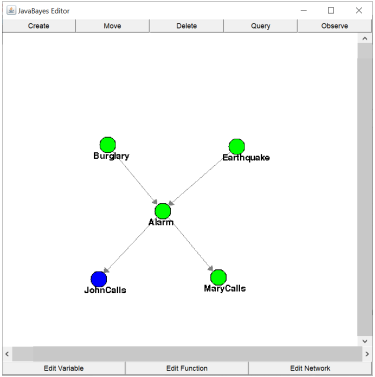
Response:

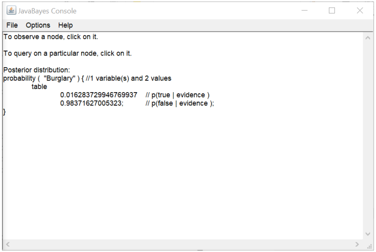
P([Burglary=t] | [John=t] AND [Mary=t]) = 0.2841.

- PIT is not a classical Bayesian network tool.
- PIT can take arbitrary conditional probabilities and queries as input and calculate correct results.
- On input of CPTs or equivalent rules, the MaxEnt principle implies the same conditional independences and thus also the same answers as a Bayesian network.
- Bayesian networks are thus a special case of MaxEnt.

Reasoning with Bayesian Networks

Software for Bayesian Networks





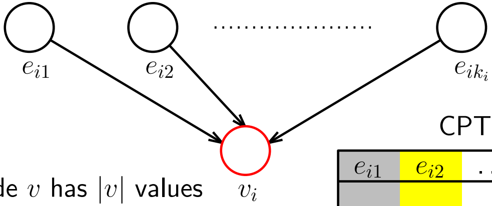
- A classic system is JavaBayes. Two windows: graphical editor + console
- With the graphical network editor, nodes and edges can be manipulated and the values in the CPTs edited.
- The values of variables can be assigned with "Observe" and the values of other variables called up with "Query". The answers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

- 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.

- For the variables $v_1, \ldots, v_n$ with $|v_1|, \ldots, |v_n|$ different values each, the distribution has a total of

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

independent entries.
- Alarm example: $2^5 - 1 = 31$ independent entries.



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)

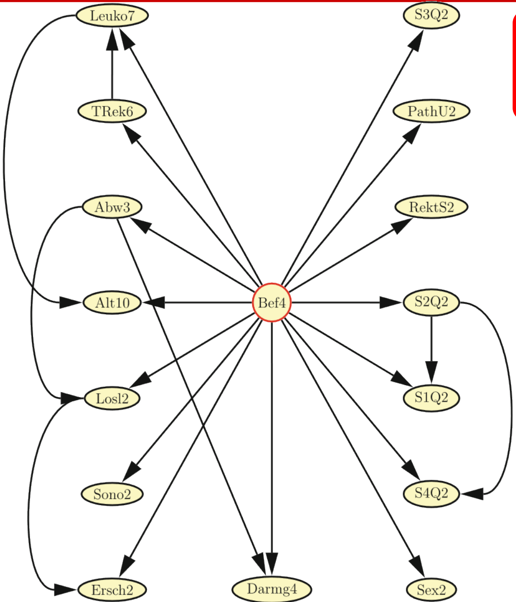
Alarm example: $2 + 2 + 4 + 1 + 1 = 10$ entries which uniquely describe the network.

Special case:

- n variables
- Equal number b of values
- Each node has k parent nodes
- All CPTs together have $n(b-1)b^k$ entries
- Complete distribution contains $b^n - 1$ entries
- Local connection
- Network becomes modularized $\Rightarrow$ reduction in complexity

Reasoning with Bayesian Networks

Development of Bayesian Networks



Bayesian network for the LEXMED application

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

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

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

$$\sum_{i=1}^{n}(|v_i|-1)\prod_{j=1}^{k_i}|e_{ij}| = 6\cdot 6\cdot 4 +$$

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

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

Reasoning with Bayesian Networks

Development of Bayesian Networks

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

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

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

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

$$9\cdot 7\cdot 4+$$

Reasoning with Bayesian Networks

Development of Bayesian Networks

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

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

$$9\cdot 7\cdot 4 + 1\cdot 3\cdot 4 +$$

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

Reasoning with Bayesian Networks

Development of Bayesian Networks

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

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

$$9\cdot 7\cdot 4+1\cdot 3\cdot 4+1\cdot 4+$$

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

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

$$9\cdot 7\cdot 4+1\cdot 3\cdot 4+1\cdot 4+$$

$$1\cdot 2\cdot 4+$$

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

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

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

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

Reasoning with Bayesian Networks

Development of Bayesian Networks

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

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

$$9\cdot 7\cdot 4 + 1\cdot 3\cdot 4 + 1\cdot 4 +$$

$$1\cdot 2\cdot 4 + 3\cdot 3\cdot 4 + 1\cdot 4 +$$

$$1\cdot 4\cdot 2 +$$

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

$$\sum_{i=1}^{n}(|v_i|-1)\prod_{j=1}^{k_i}|e_{ij}| = 6\cdot6\cdot4+5\cdot4+2\cdot4+$$
$$9\cdot7\cdot4+1\cdot3\cdot4+1\cdot4+$$
$$1\cdot2\cdot4+3\cdot3\cdot4+1\cdot4+$$
$$1\cdot4\cdot2+1\cdot4\cdot2+$$

Reasoning with Bayesian Networks

Development of Bayesian Networks

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

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

Reasoning with Uncertainty
Hoàng Anh Đức
Additional Materials
Introduction
Computing with Probabilities
Conditional Probability
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
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

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

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

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

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

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

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

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		$\boxed{2}$	$\boxed{P3Q2}$
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*	$\boxed{4}$	$\boxed{Diag4}$

- Size of the distribution: $20\,643\,839$ values.
- Size of the Bayesian network: 521 values. **[Why?]**

$$\sum_{i=1}^{n}(|v_i|-1)\prod_{j=1}^{k_i}|e_{ij}| = 6\cdot6\cdot4+5\cdot4+2\cdot4+ 9\cdot7\cdot4+1\cdot3\cdot4+1\cdot4+ 1\cdot2\cdot4+3\cdot3\cdot4+1\cdot4+ 1\cdot4\cdot2+1\cdot4\cdot2+1\cdot4+ 1\cdot4+1\cdot4+1\cdot4+1 = 521.$$

Symptom	Values	#	Short
Gender	*Male, female*	2	*Sex2*
	65–	10	*Age10*
		2	*P1Q2*
		2	*P2Q2*
		2	*P3Q2*
		2	*P4Q2*
		3	*Gua3*
		2	*Reb2*
		2	*Tapp2*
		2	*RecP2*
		4	*BowS4*
		2	*Sono2*
		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*

Construction of a Bayesian network

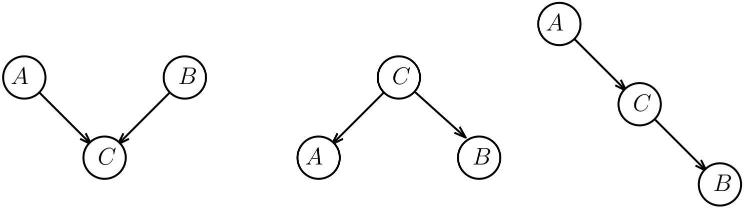
(1) Design of the network structure (usually performed manually)

(2) Entering the probabilities in the CPTs (usually automated)

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*

Reasoning with Bayesian Networks
Development of Bayesian Networks



A and B are independent

A and B are independent given C

Figure: There is no edge between A and B if they are independent (left) or conditionally independent (middle, right).

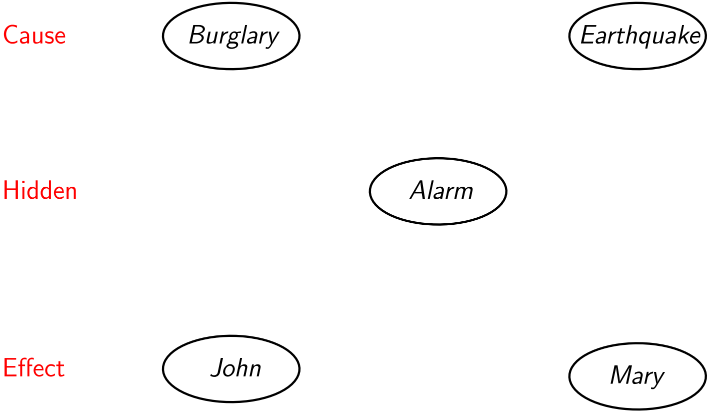


Figure: Stepwise construction of the alarm network considering causality

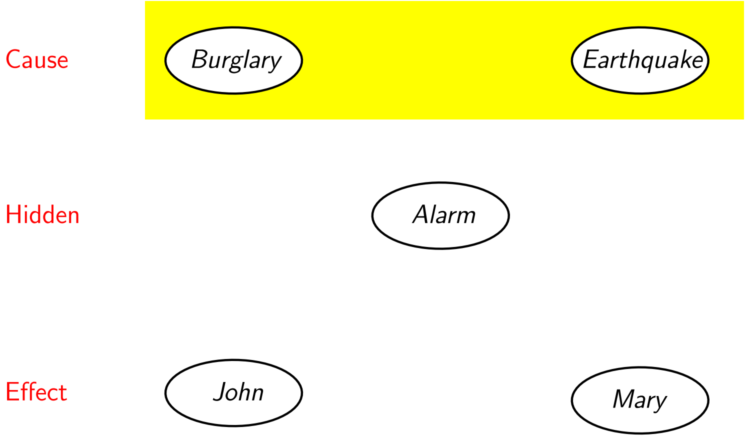


Figure: Stepwise construction of the alarm network considering causality

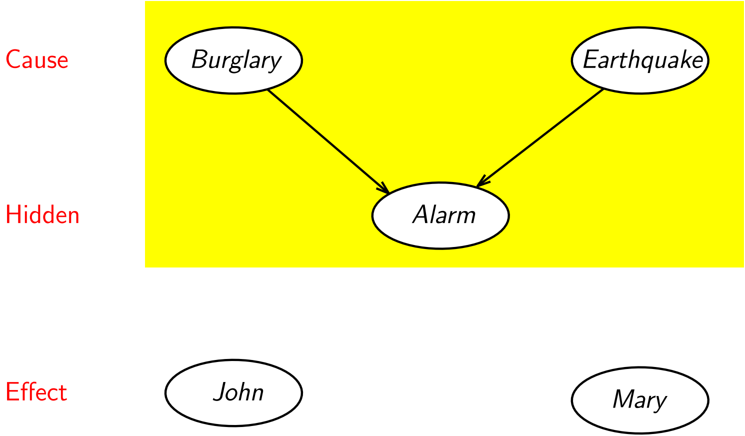


Figure: Stepwise construction of the alarm network considering causality

Reasoning with Bayesian Networks

Development of Bayesian Networks

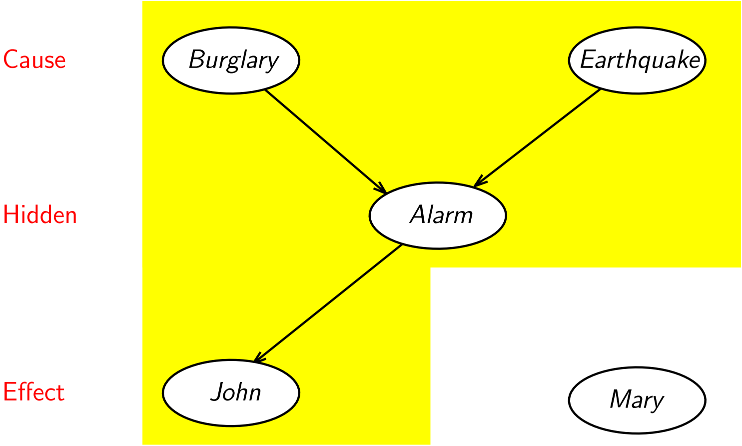


Figure: Stepwise construction of the alarm network considering causality

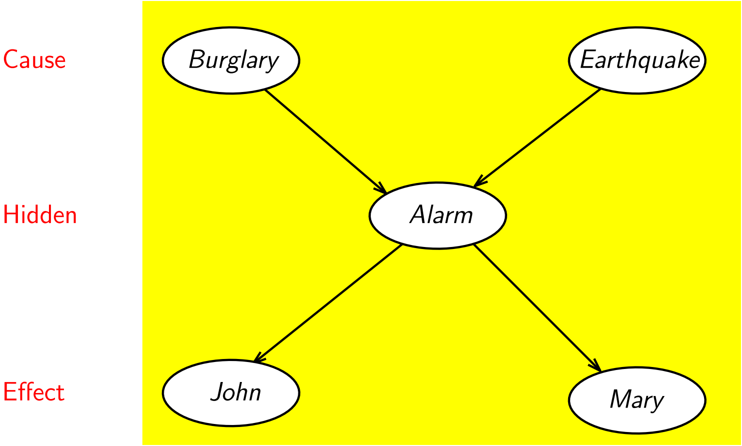


Figure: Stepwise construction of the alarm network considering causality

- The *structure of the Bayesian network* heavily *depends on the chosen variable ordering*.
- If the order of variables is chosen to reflect the causal relationship *beginning with the causes and proceeding to the diagnosis variables*, then the result will be *a simple network*.
- *Otherwise* the network may contain *significantly more edges*. Such non-causal networks are *often very difficult to understand* and have a *higher complexity for reasoning*.

Requirements

- Bayesian network has *no cycles*.
- The *variables are numbered* such that *no variable has a lower index than any variable that predecessor*.

It holds

$$\mathbf{P}(X_n \mid X_1, \ldots, X_{n-1}) = \mathbf{P}(X_n \mid \mathit{Parent}(X_n))$$

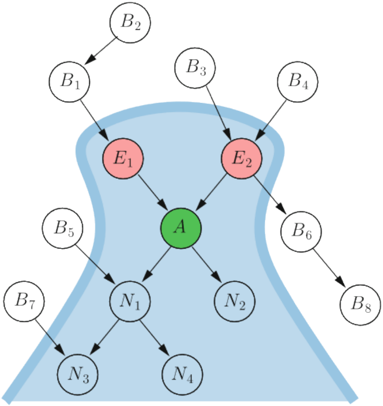
$\Leftrightarrow$ An arbitrary variable X_i in a Bayesian network is conditionally independent of its ancestors, given its parents.

More generally,

Theorem 7

A node in a Bayesian network is conditionally independent from all non-successor nodes, given its parents.

Example of conditional independence in a Bayesian network. If the parent nodes E_1 and E_2 are given, then all non-successor nodes $B_1, \ldots, B_8$ are independent of A.



- Chain rule for Bayesian network

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

- Using this rule in the alarm example,

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

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). A DAG is a graph without cycles, that is, without paths of the form $(A, \ldots, A)$.
  - For every variable A the CPT (that is, the table of conditional probabilities $P(A \mid \textit{Parents}(A))$) is given.
- Two variables A and B are called *conditionally independent* given C if $\mathbf{P}(A, B \mid C) = \mathbf{P}(A \mid C)\mathbf{P}(B \mid C)$ or, equivalently, if $\mathbf{P}(A \mid B, C) = \mathbf{P}(A \mid C)$.

Basics of Bayesian Networks (cont.)

- Besides the foundational rules of computation for probabilities, the following rules are also true:

  Bayes' Theorem $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$.

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

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

Basics of Bayesian Networks (cont.)

- *A variable in a Bayesian network is conditionally independent of all non-successor variables given its parent variables*. If $X_1, \ldots, X_{n-1}$ are no successors of X_n, we have $P(X_n \mid X_1, \ldots, X_{n-1}) = P(X_n \mid \mathit{Parents}(X_n))$. This condition must be honored during the construction of a network.
- During construction of a Bayesian network *the variables should be ordered according to causality*. First the causes, then the hidden variables, and the diagnosis variables last.
- *Chain rule:* $\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid \mathit{Parent}(X_i))$.

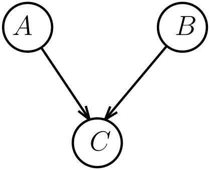
Exercise 12 ([Ertel 2025], Exercise 7.10, p. 173)

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



Exercise 13 ([Ertel 2025], Exercise 7.11, p. 173)

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

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

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

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

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

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

(f) Check all of your results with JavaBayes and with PIT.

(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 nodes is removed. Which CPTs does this change? (Why these in particular?)

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

Exercise 14 ([Ertel 2025], Exercise 7.12, p. 173)

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	*Value*
Li	Light is on	t/f
Str	Street condition	*dry*, *wet*, *snow_covered*
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:

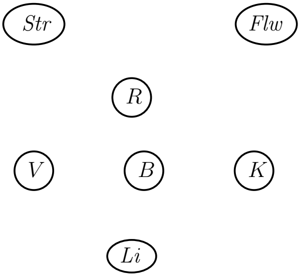
$$P(\textit{Li} \mid V, R) = P(\textit{Li} \mid V)$$
$$P(V \mid R, \textit{Str}) = P(V \mid R)$$
$$P(V \mid R, \textit{Flw}) = P(V \mid R)$$

(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 $(\textit{Str}, \textit{Li})$.

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



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

- Combining MaxEnt and Bayesian networks
  - Building a network according to the Bayesian methodology, enter all the edges accordingly and 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 $P(A) = 0.8$.
- PIT recognizes inconsistency.
- In some cases reasoning is possible anyway.

Summary

- Medial expert system LEXMED
  - can be modeled and implemented using MaxEnt and Bayesian networks
  - can replace the well-established, but too weak linear scoring systems used in medicine
- Better than linear score systems.
- 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

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.

References

- Ertel, Wolfgang (2025). *Introduction to Artificial Intelligence*. 3rd. Springer. DOI: 10.1007/978-3-658-43102-0.
- 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.
- Shortliffe, Edward (1976). *Computer-based medical consultations: MYCIN*. Elsevier.