VNU-HUS MAT1206E/3508: Introduction to AI

Reasoning with Uncertainty
In-class Discussion

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.

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

⇒ Require *different kinds of logic!*

- This example illustrates problems with traditional (classical) logic.
- It shows how standard logic struggles with exceptions and uncertain knowledge.
- Key idea: Logic derives that a penguin can fly, which is absurd!
- Leads to the need for *probabilistic* or *non-monotonic* logic.

Introduction

Additionally, *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!

Introduction

Reasoning with conditional probabilities

- *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

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

Example 1

- **Experiment:** Rolling a fair six-sided die.
- **Sample space:** $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Let E be the **event** "rolling an even number," so $E = \{2, 4, 6\}$ (a subset of Ω).
- If the die shows 4 (which is in E), then E has occurred.
- Let $A = \{1, 3, 5\}$ (odd numbers) and $B = \{4, 5, 6\}$ (numbers greater than 3).
- Then $A \cup B = \{1, 3, 4, 5, 6\}$ (union of the subsets), which is also an event meaning "rolling an odd number or a number greater than 3."
- **Elementary event:** The event "rolling a 3" is $\{3\}$. This is the smallest possible non-empty event, representing a single specific outcome.
- **Sure event:** Ω. This is the entire sample space, $\{1, 2, 3, 4, 5, 6\}$, which always occurs no matter what the die shows—you're guaranteed to roll one of these numbers.
- **Impossible event:** $\emptyset$ (the empty set). This represents something that can never happen, like "rolling a 7" on a six-sided die. The subset is empty because no outcome in Ω satisfies it.

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 2

Throwing a die, the probability for an even number is

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

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 $P(\textit{JohnCalls})$ instead of $P(\textit{JohnCalls} = t)$.

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).*

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.

Computing with Probabilities

Conditional Probability

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.

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

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

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

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

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

- 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*.

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

Theorem 3 (Bayes' Theorem)

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

Example 3

Leuko Leukocyte value higher than 10000

App Patient has appendicitis (appendix inflammation)

$\mathbf{P}(\textit{App}, \textit{Leuko})$	*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:

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

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

Example 3 (continued)

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

- 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 4

- 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 4

- 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!*

Example 4

- 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

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

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

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

The Principle of Maximum Entropy
An Inference Rule for Probabilities

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

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

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

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

- Solve it as far as is possible, we get

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

One equation is missing!

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*.

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

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

- 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.

Exercise 1

Do your own research on the relationship between *conditional probability* and *material implication* in the context of modeling reasoning.

- 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.

Exercise 2

Do your own research on the application of MaxEnt systems in the medical expert system LEXMED.

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.

- 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 5 (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:* $P(Bur \mid J \vee M)$, $P(J \mid Bur)$, $P(M \mid Bur)$

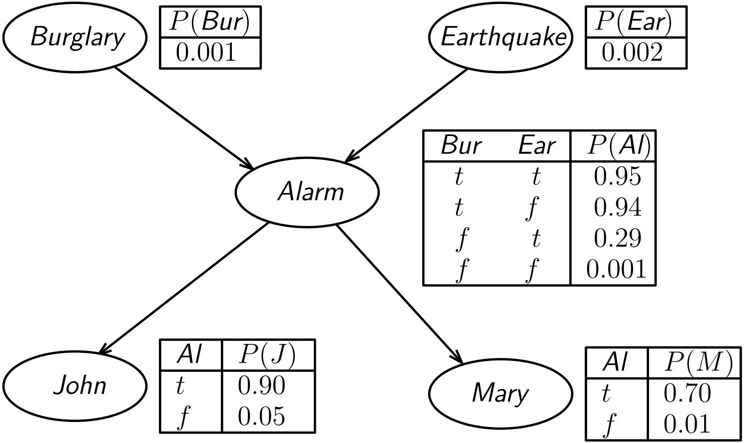
- A *Bayesian network* is a directed acyclic graph (DAG) in which
  - each node represents a random variable,
  - each edge $X_i \to X_j$ represents a *direct influence* of variable X_i on variable X_j, and
  - each node is associated with a conditional probability table (CPT) that quantifies the effects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

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

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

Reasoning with Bayesian Networks

Graphical Representation of Knowledge as a Bayesian Network



$P(Bur)$
0.001

$P(Ear)$
0.002

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

Al	$P(J)$
t	0.90
f	0.05

Al	$P(M)$
t	0.70
f	0.01

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

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 6 (Alarm-Example (cont.))

- John and Mary independently react to an alarm.
  $\mathbf{P}(J, M \mid \mathit{Al}) = \mathbf{P}(J \mid \mathit{Al}) \cdot \mathbf{P}(M \mid \mathit{Al})$.
- Thus, *given an alarm, two variables J and M are independent*.
- (Without any condition,) J and M are not independent, that is, $\mathbf{P}(J, M) \neq \mathbf{P}(J) \cdot \mathbf{P}(M)$. **[Why?]**
  - **Hint:** It 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 $P(\mathit{Al})$ using the given probabilities, marginalization, and independence of *Bur* and *Ear*.
    (Result: $P(\mathit{Al}) \approx 0.00252$.)
  - Then calculate $P(J)$ and $P(M)$ using conditional probabilities and the computed $P(\mathit{Al})$.
    (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 7 (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.

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

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

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

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*

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

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

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

Therefore,

$$
\begin{aligned}
P(J \mid \mathit{Bur}) &= P(J \mid \mathit{Al})P(\mathit{Al} \mid \mathit{Bur}) + P(J \mid \neg\mathit{Al})P(\neg\mathit{Al} \mid \mathit{Bur}) \\
&= 0.9 \cdot 0.94 + 0.05 \cdot 0.06 = 0.849.
\end{aligned}
$$

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

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

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

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.

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

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

Reasoning with Bayesian Networks
Practical Application

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

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

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

- 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%.

Construction of a Bayesian network

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

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

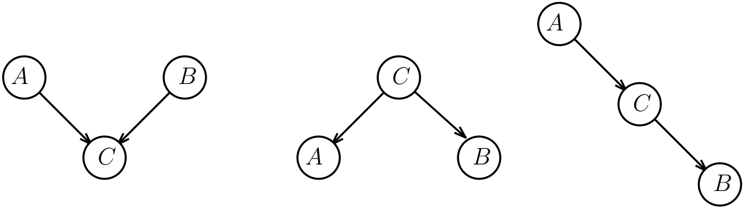
Construction of the network in the alarm example.

- **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*

Design of a Bayesian network structure

1. *Nodes:* Determine the set of random variables required to model the domain and fix an ordering $\{X_1, \ldots, X_n\}$, where, if possible, *causes precede effects* to obtain a more compact network.
2. *Links:* For each $i = 1, \ldots, n$ do:
   (a) Select a minimal parent set $\mathrm{Parents}(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ such that the conditional independence constraint holds:
   $$P\big(X_i \mid X_1, \ldots, X_{i-1}\big) \;=\; P\big(X_i \mid \mathrm{Parents}(X_i)\big).$$
   Minimality means no proper subset of $\mathrm{Parents}(X_i)$ satisfies the equality.
   (b) For each $X_j \in \mathrm{Parents}(X_i)$ insert the directed edge $X_j \to X_i$.
   (c) Specify the conditional probability table (CPT) for X_i:
   $$P\big(X_i \mid \mathrm{Parents}(X_i)\big),$$
   i.e., list $P(X_i = x \mid \mathrm{Parents}(X_i) = p)$ for every value x of X_i and every combination p of values of $\mathrm{Parents}(X_i)$.

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).

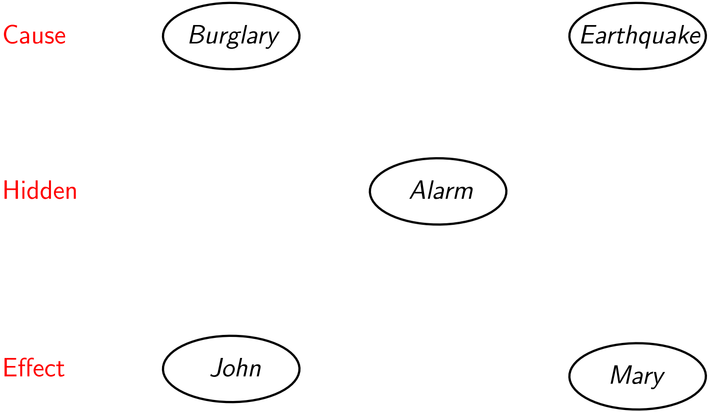


Figure: Stepwise construction of the alarm network considering causality

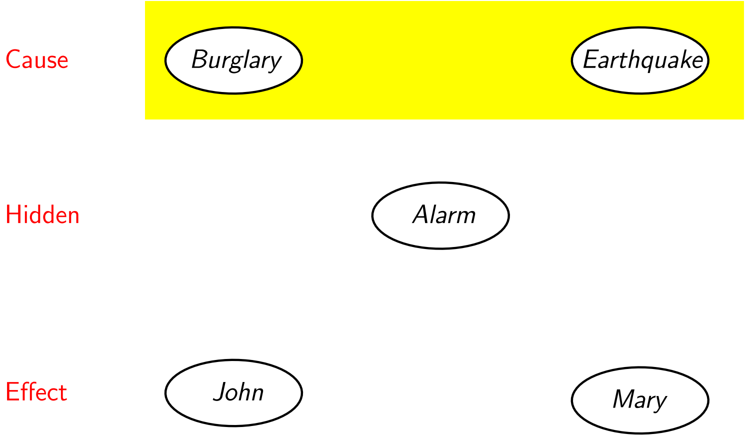


Figure: Stepwise construction of the alarm network considering causality

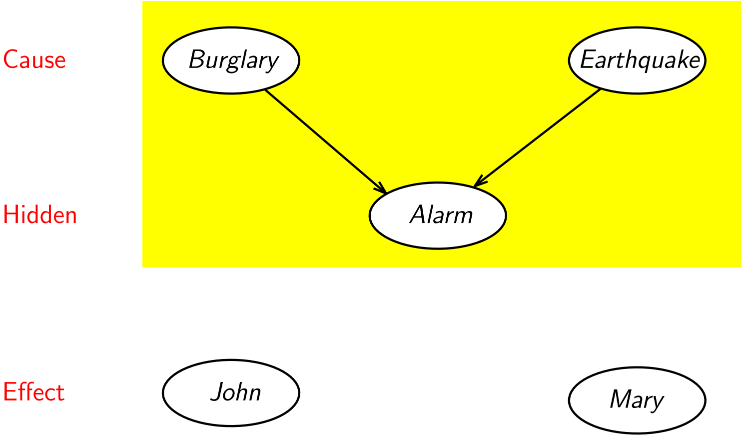


Figure: Stepwise construction of the alarm network considering causality

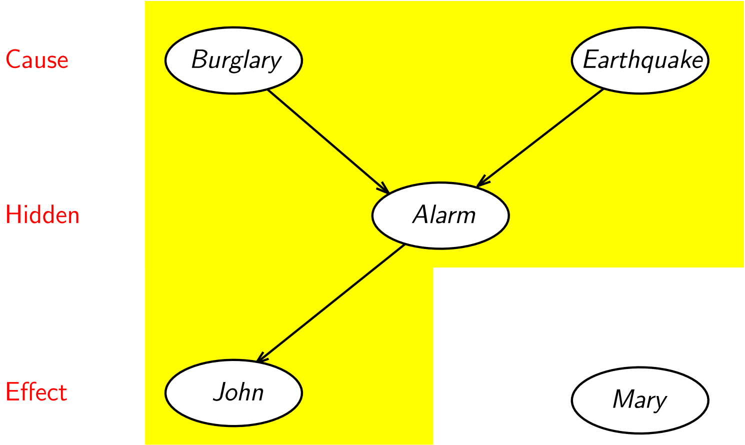


Figure: Stepwise construction of the alarm network considering causality

Reasoning with Bayesian Networks
Development of Bayesian Networks

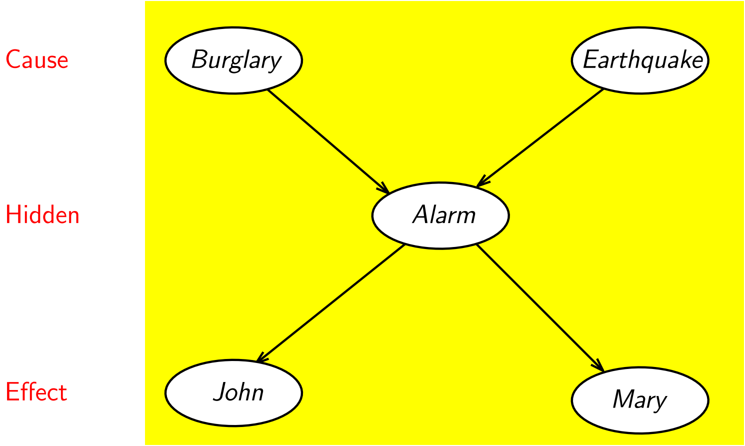


Figure: Stepwise construction of the alarm network considering causality

- A Bayesian network is often *far more compact than the full joint distribution*. This compactness makes it *feasible to handle domains with many variables*.
- **Locally structured (sparse) systems:** *each subcomponent interacts directly with only a bounded number of other components, independent of the total number of components*. This typically yields linear rather than exponential growth in complexity.
- If each random variable is *directly influenced* by *at most k others* (for some constant k) and we assume n Boolean variables, then:
  - each conditional probability table (CPT) needs at most 2^k numbers,
  - the whole network can be specified by at most $n2^k$ numbers,
  - whereas the full joint distribution requires 2^n numbers.
  - Concrete example: $n = 30$, $k = 5$:

$$n2^k = 30 \cdot 2^5 = 30 \cdot 32 = 960, \qquad 2^n = 2^{30} \approx 1.07 \times 10^9.$$

- If a network is *fully connected* (every variable directly influenced by all others) the CPT specification cost approaches that of the joint distribution.
- **Practical tradeoff:** small, tenuous dependencies can be omitted to avoid large increases in model complexity. Example: one might add edges Ear $\rightarrow$ J and Ear $\rightarrow$ M, but only if the gain in accuracy justifies the extra parameters.

- 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

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

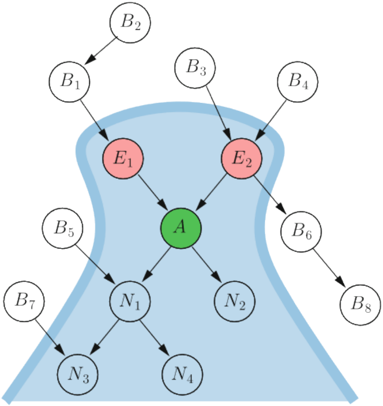
$\Leftrightarrow$ An arbitrary variable X_i in a Bayesian network is conditionally independent of its ancestors, given its parents.

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.



Reasoning with Bayesian Networks

Semantics of Bayesian Networks

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

References

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.