Introduction to Artificial Intelligence

George Voutsadakis[1]

[1]Mathematics and Computer Science
Lake Superior State University

LSSU Math 400

Subsection 1

The Flying Penguin

Tweety, the Flying Penguin

- Consider the statements:
  1. Tweety is a penguin;
  2. Penguins are birds;
  3. Birds can fly.
- Formalized in PL1, they yield the knowledge base KB:

$$\begin{aligned} &\text{penguin}(\text{tweety}) \\ &\text{penguin}(x) \Rightarrow \text{bird}(x) \\ &\text{bird}(x) \Rightarrow \text{fly}(x) \end{aligned}$$

- With resolution, KB $\vdash$ fly(tweety).
- This formalization of the flight attributes of penguins is insufficient.
- If we add the statement "Penguins cannot fly", i.e.,

$$\text{penguin}(x) \Rightarrow \neg\text{fly}(x),$$

  $\neg$fly(tweety) can be derived, but fly(tweety) is still true.
- The knowledge base becomes therefore inconsistent.

Monotonic Logic

- The logic PL1 is monotonic: Although a formula explicitly stating that penguins cannot fly was added, the opposite can still be derived.

Definition of Monotonicity

A logic is called **monotonic** if, for an arbitrary knowledge base KB and an arbitrary formula φ, the set of formulas derivable from KB is a subset of the formulas derivable from $\text{KB} \cup \{\varphi\}$.

- After extending a set of formulas all previously derivable statements can still be proved and, potentially, additional statements.
- For the Tweety example this means that the extension of the knowledge base will never achieve the desired goal.

Tweety Logically Accommodated...

- We modify KB by replacing the obviously false statement "(all) birds can fly" with the more exact statement "(all) birds except penguins can fly" and obtain KB_2:

$$\begin{aligned}
&\text{penguin}(\text{tweety}) \\
&\text{penguin}(x) \Rightarrow \text{bird}(x) \\
&\text{bird}(x) \wedge \neg\text{penguin}(x) \Rightarrow \text{fly}(x) \\
&\text{penguin}(x) \Rightarrow \neg\text{fly}(x)
\end{aligned}$$

- Now we can derive $\neg\text{fly}(\text{tweety})$, but not $\text{fly}(\text{tweety})$, because for that we would need $\neg\text{penguin}(\text{tweety})$, which is not derivable.

...Normal Birds Left Logically Destitute

- As long as there are only penguins in this world, peace reigns. Every normal bird, however, immediately causes problems.
- We add the sea duck Seamore and get KB_3.
- Flight attributes of Seamore are a mystery because we forgot to say "sea ducks are not penguins", so we extend to KB_4:

$$\begin{aligned}
&\text{seaduck}(\text{seamore})\\
&\text{seaduck}(x) \Rightarrow \text{bird}(x)\\
&\text{seaduck}(x) \Rightarrow \neg\text{penguin}(x)\\
&\text{penguin}(\text{tweety})\\
&\text{penguin}(x) \Rightarrow \text{bird}(x)\\
&\text{bird}(x) \wedge \neg\text{penguin}(x) \Rightarrow \text{fly}(x)\\
&\text{penguin}(x) \Rightarrow \neg\text{fly}(x)
\end{aligned}$$

- The fact that sea ducks are not penguins, which is self-evident to humans, must be explicitly added.
- In fact, for every type of bird (except for penguins) we must say that it is not a member of penguins.

More Idiosyncratic Logics

- For every object in KB, in addition to its attributes, all of the attributes it does not have must be listed.
  - To solve this problem non-monotonic logics have been developed.
  - Default logics allow objects to be assigned attributes which are valid as long as no other rules are available.
- Monotony can be especially inconvenient in complex planning problems in which the world can change.
  - If for example a blue house is painted red, then afterwards it is red. A knowledge base such as

    $$\text{color}(\text{house}, \text{blue}), \quad \text{paint}(\text{house}, \text{red}), \quad \text{paint}(x, y) \Rightarrow \text{color}(x, y),$$

    leads to the conclusion that, after painting, the house is red and blue.
- This problem in planning is known as the frame problem.
- A solution for this is the situation calculus.
- An interesting approach for modeling problems such as the Tweety example is probabilistic logic based on probability theory.
  - The statement "all birds can fly" is false.
  - A statement like "almost all birds can fly" is correct.

Subsection 2

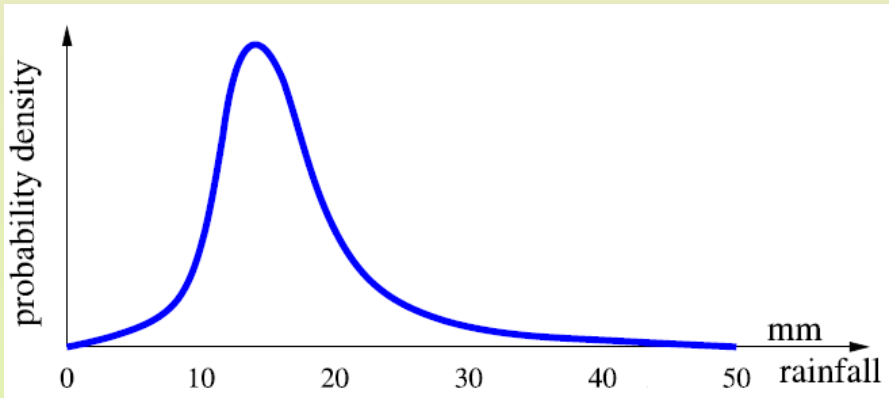
Modeling Uncertainty

Probabilistic Statements

- Two-valued logic can and should only model circumstances in which the only relevant truth values are true and false.
- For many tasks in everyday reasoning, two-valued logic is not expressive enough.
  - The rule $\text{bird}(x) \Rightarrow \text{fly}(x)$ is true for almost all birds, but for some it is false.
- Working with probabilities allows exact formulation of uncertainty.
  - The statement "99% of all birds can fly" can be formalized by the expression $P(\text{bird}(x) \Rightarrow \text{fly}(x)) = 0.99$.
- Later, we will see that, in this case, it is preferable to work with conditional probabilities such as $P(\text{fly}|\text{bird}) = 0.99$.
- With the help of Bayesian networks, complex applications with many variables can also be modeled.

Fuzzy Statements and Probability Densities

- A different model is needed for the statement "The weather is nice", since it makes little sense to speak in terms of true and false.
- The variable weather_is_nice should not be modeled as binary, but rather continuously with values, for example, in the interval $[0, 1]$: weather_is_nice $= 0.7$ means "The weather is fairly nice".
- Fuzzy logic was developed for this type of continuous (fuzzy) variable.
- Probability theory also offers the possibility of making statements about the probability of continuous variables.
  - A statement "There is a high probability that there will be some rain" could be formulated as a probability density $P(\text{rainfall} = X) = Y$:



Hierarchy of Propositional Formalisms

- Probabilistic and fuzzy logic, combined with inductive statistics and the theory of Bayesian networks, make it possible to answer arbitrary probabilistic queries.
- Probability theory as well as fuzzy logic are not directly comparable to predicate logic because they do not allow variables or quantifiers.
- They can thus be seen as extensions of propositional logic:

Formalism	# of Truth Values	Probability?
Propositional logic	2	no
Fuzzy logic	∞	no
Discrete probabilistic logic	n	yes
Continuous probabilistic logic	∞	yes

Uncertain and Incomplete Knowledge

- Reasoning under uncertainty with limited resources plays a key role in both everyday situations and in many technical applications of AI.
- In these areas heuristic processes are very important as, e.g., in looking for a parking space in city traffic.
- Heuristics alone are often not enough, especially when a quick decision is needed given incomplete knowledge.
- Example: A pedestrian crosses the street and an auto quickly approaches. To prevent a serious accident, the pedestrian must react quickly. He is not capable of worrying about complete information about the state of the world. He must come to an optimal decision under the given constraints (little time and little, potentially uncertain, knowledge).
- In similar situations a method for reasoning with uncertain and incomplete knowledge is needed.

Uncertainty and Incompleteness in Medical AI

- How could we reason under uncertainty in simple medical diagnosis?
- 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. Using propositional logic:

$$\text{Stomach\_pain\_right\_lower} \wedge \text{Leukocytes} > 10000 \rightarrow \text{Appendicitis}$$

- If we know that Stomach_pain_right_lower $\wedge$ Leukocytes > 10000 is true, then we can use modus ponens to derive Appendicitis.
- This model is clearly too coarse.
- In 1976, Shortliffe and Buchanan recognized this when building their medical expert system MYCIN. They developed a calculus using so-called certainty factors, which allowed the certainty of facts and rules to be represented.
- A rule $A \rightarrow B$ is assigned a certainty factor β, written $A \rightarrow_\beta B$.
- The semantics of a rule $A \rightarrow_\beta B$ is defined via the conditional probability $P(B|A) = \beta$.

Uncertainty and Incompleteness in Medical AI (Cont'd)

- In the above example, the rule could then read

$$\text{Stomach\_pain\_right\_lower} \wedge \text{Leukocytes} > 10000 \rightarrow_{0.6} \text{Appendicitis}.$$

- For reasoning with this kind of formulas, they used a calculus for connecting the factors of rules.
- It was shown that with this calculus inconsistent results could be derived.
- There were attempts to solve this problem by using non-monotonic logic and default logic, which proved unsuccessful.
- The Dempster-Schäfer theory assigns a belief function Bel(A) to a proposition A, whose value gives the degree of evidence for the truth of A. But even this formalism has weaknesses.
- Even fuzzy logic, with its successes in control theory, demonstrates considerable weaknesses when reasoning under uncertainty in more complex applications.

Probability Theory Offers Solutions

- Probability theory has had more and more influence in AI.
- In reasoning with Bayesian networks, or subjective probability, it has become an indispensable tool in the AI toolbox.
- Rather than implication as it is known in logic (material implication), conditional probability is used, which models everyday causal reasoning significantly better.
- Reasoning with probability profits heavily from the fact that probability theory is a well developed old branch of mathematics.
- We
  - introduce the foundations needed for reasoning with probability;
  - present an example for reasoning with uncertain and incomplete knowledge;
  - blend in, in a natural way, the method of maximum entropy (MaxEnt);
  - study reasoning with Bayesian networks;
  - show the relationship between the two methods.

Subsection 3

Probability: The Basics

Sample Space, Events and Elementary Events

- For a single roll of a gaming die (experiment), the probability of the event "rolling a six" equals $\frac{1}{6}$, whereas the probability of the occurrence "rolling an odd number" is equal to $\frac{1}{2}$.

Definition of Events and Elementary Events

The **sample space** Ω of an experiment is the finite set of all possible outcomes of the experiment. Each individual outcome $\omega \in \Omega$ (viewed as a subset $\{\omega\} \subseteq \Omega$) is called an **elementary event**. An **event** is any subset (containing, possibly, many outcomes) of the sample space.

- Example (cont'd): For a single roll of one gaming die the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. So "rolling a 6" ($\{6\}$) is an elementary event. "Rolling an even number" ($\{2, 4, 6\}$) is an event, but not an elementary event. "Rolling a number smaller than five" ($\{1, 2, 3, 4\}$) is also an event that is not elementary.

Set Operations and Logical Connectives

- Given two events A and B, $A \cup B$ is also an event.
- Ω itself is the **certain event** (it always occurs), and the empty set $\emptyset$ is the **impossible event** (it never occurs).
- If we view the events $A, B \subseteq \Omega$ as propositions "A occurs", "B occurs", then, instead of $A \cap B$, we may write $A \wedge B$.
- Think, semantically, of the intersection of two sets being defined

$$x \in A \cap B \quad \text{iff} \quad x \in A \wedge x \in B.$$

- Similarly for other 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

- The variables are called **random variables** in probability theory.

Discrete Probabilities

- The probability of "rolling a five or a six" is equal to $\frac{1}{3}$. This can be described by $P(\text{facenumber} \in \{5, 6\}) = P(\text{face\_number} = 5 \vee \text{face\_number} = 6) = \frac{1}{3}$.

Definition of Discrete Probability

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a finite sample space. If all elementary events are equiprobable, then the probability $P(A)$ of the event A is $P(A) = \frac{|A|}{|\Omega|}$.

- So every elementary event has the probability $\frac{1}{|\Omega|}$.
- To describe events we use variables with appropriate values.
- For example, a variable eye_color can take on the values green, blue, brown. eye_color = blue then describes an event.
- For binary (boolean) variables, we usually write $P(\text{JohnCalls})$ instead of (the formally correct) $P(\text{JohnCalls} = t)$.

Properties of Discrete Probabilities

- The probability of rolling an even number is

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

- The following important rules follow directly from the definition:

Theorem (Properties of Discrete Probabilities)

1. $P(\Omega) = 1$.
2. $P(\emptyset) = 0$.
3. For pairwise exclusive events A and B, $P(A \vee B) = P(A) + P(B)$.
4. $P(A) + P(\neg A) = 1$.
5. For arbitrary events A and B, $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$.
6. For $A \subseteq B$, $P(A) \leq P(B)$.
7. If $A_1, \ldots, A_n$ are pairwise disjoint and $\bigcup_{i=1}^{n} A_i = \Omega$, then $\sum_{i=1}^{n} P(A_i) = 1$.

Joint Probability Distributions

- We write $P(A, B)$ for $P(A \wedge B)$.
- We call the vector $(P(A, B), P(A, \neg B), P(\neg A, B), P(\neg A, \neg B))$ consisting of the probabilities of all combinations of truth values for A, B, a **distribution** or **joint probability distribution of** A **and** B. A shorthand for this is $\mathbf{P}(A, B)$.
- In the case of two variables:

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

- For the d variables $X_1, \ldots, X_d$ with n values each, the distribution has the values $P(X_1 = x_1, \ldots, X_d = x_d)$, where $x_1, \ldots, x_d$ can be any of the n different values. The distribution can therefore be represented as a d-dimensional matrix with a total of n^d elements. One of these n^d values is redundant (why?) and the distribution is characterized by $n^d - 1$ unique values.

Subsection 4

Conditional Probability

Conditional Probability: An Example

- The speed of 100 vehicles is measured. For each measurement it is also noted whether the driver is a student.

Event	Frequency	Relative frequency
Vehicle observed	100	1
Driver is a student (S)	30	0.3
Speed too high (H)	10	0.1
Student and speeding ($S \wedge H$)	5	0.05

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

The answer is given by the conditional probability

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

This is different from the a priori probability $P(H) = 0.1$ of speeding.

Conditional Probability and Independence

Definition of Conditional Probability

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

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

- The conditional probability $P(A|B)$ can be understood as the probability of $A \wedge B$ when we only look at the event B, i.e., $P(A|B) = \frac{|A \wedge B|}{|B|}$.
- In fact, $P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{\frac{|A \wedge B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{|A \wedge B|}{|B|}$.

Independence

Definition of Independent Events

If, for two events A and B, $P(A|B) = P(A)$, then these events are called **independent**.

- Thus A and B are independent if the probability of the event A is not influenced by the event B.

Theorem (Characterization of Independence)

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

- Example: For a roll of two independent dice, the probability of "rolling two sixes" is

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

The first equation is true only when the two dice are independent. If not, and, say, die 2 is bound to be always the same as die 1, then $P(D_1 = 6 \wedge D_2 = 6) = \frac{1}{6}$.

The Product Rule and the Chain Rule

- Note that by the definition of conditional probability, we get the **Product Rule**:

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

- For n variables, this yields the **Chain Rule**:

$$\begin{aligned}
&P(X_1, \ldots, X_n) \\
&= P(X_n|X_1, \ldots, X_{n-1}) \cdot P(X_1, \ldots, X_{n-1}) \\
&= P(X_n|X_1, \ldots, X_{n-1}) \cdot P(X_{n-1}|X_1, \ldots, X_{n-2}) \cdot P(X_1, \ldots, X_{n-2}) \\
&= P(X_n|X_1, \ldots, X_{n-1}) \cdot P(X_{n-1}|X_1, \ldots, X_{n-2}) \cdot \cdots \cdot P(X_2|X_1) \cdot P(X_1) \\
&= \sum_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}).
\end{aligned}$$

- Thus, we can represent a distribution as a product of conditional probabilities.

Marginalization

- Because $A \Leftrightarrow (A \wedge B) \vee (A \wedge \neg B)$ is true for binary variables A and B, and the events $A \wedge B$ and $A \wedge \neg B$ are disjoint,

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

- For arbitrary variables $X_1, \ldots, X_d$, a variable, say X_d, can be eliminated by summation over all of its values:

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

- The application of this formula is called **marginalization**.
- This summation can continue with the variables $X_1, \ldots, X_{d-1}$ until just one variable is left.
- Marginalization can also be applied to the distribution $P(X_1, \ldots, X_d)$. The resulting distribution $P(X_1, \ldots, X_{d-1})$ is called the **marginal distribution**.

Marginalization: An Example

- We observe the set of all patients who come to the doctor with acute stomach pain. For each patient the leukocyte value (abundance of white blood cells) is measured. The variable Leuko is true if and only if the leukocyte exceeds 10,000. This indicates an infection. The variable App tells us whether the patient has appendicitis. The distribution $P(\text{App}, \text{Leuko})$ is given by:

$\mathbf{P}(\text{App}, \text{Leuko})$	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

The sums of last row and column are arrived at by marginalization:

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

Since $P(\text{Leuko}|\text{App}) = \frac{P(\text{Leuko},\text{App})}{P(\text{App})} = \frac{0.23}{0.28} = 0.82$, about 82% of all appendicitis cases lead to a high leukocyte value.

Bayes' Theorem

- Swapping A and B in the definition of conditionals yields

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A \wedge B)}{P(A)}.$$

- Solving the two equations for $P(A \wedge B)$ and setting them equal:

Bayes' Theorem

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

- Example: Applying this to the appendicitis problem, we get

$$P(\text{App}|\text{Leuko}) = \frac{P(\text{Leuko}|\text{App}) \cdot P(\text{App})}{P(\text{Leuko})} = \frac{0.82 \cdot 0.28}{0.54} = 0.43.$$

- Applying Bayes we can calculate $P(A|B)$ if we know $P(B|A)$.
- The probabilistic inference mechanism associated with Bayes' Theorem is the Principle of Maximum Entropy.

Subsection 5

The Principle of Maximum Entropy

Idea of Maximization of Entropy

- We show, using an inference example, that a calculus for reasoning under uncertainty can be realized using probability theory.
- However, there are many limitations.
- When too little knowledge is available to solve the necessary equations, new ideas are needed.
- The American physicist E.T. Jaynes in the 50's claimed that given missing knowledge, one can maximize the entropy of the desired probability distribution.
- He applied this principle to many examples and the method, as was later further developed, has now many successful technological applications.

Probabilistic Inference

- We want to derive an inference rule for uncertain knowledge that is analogous to the modus ponens.
- From the knowledge of A and the rule $A \Rightarrow B$, the conclusion B shall be reached: $\frac{A, A \Rightarrow B}{B}$.
- Adapting to probability, we get

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

- If α, β are given, what should $P(B)$ be?
- By marginalization we obtain

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

- $P(A), P(\neg A), P(B|A)$ are known, but $P(B|\neg A)$ is not.
- We cannot make an exact statement about $P(B)$, but we can estimate $P(B) \geq P(B|A) \cdot P(A)$.

Estimating Missing Probabilities

- We now consider the distribution

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

- Introduce for shorthand the four unknowns

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

- To calculate the four unknowns, four equations are needed:
  - One equation is already known in the form of the normalization condition $p_1 + p_2 + p_3 + p_4 = 1$.
  - Since $P(A) = \alpha$, $P(B|A) = \beta$, we get $P(A, B) = P(B|A) \cdot P(A) = \alpha\beta$.
  - Also $P(A) = P(A, B) + P(A, \neg B) = p_1 + p_2$.
- We get

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

- We obtain $p_2 = \alpha(1 - \beta),\ p_3 + p_4 = 1 - \alpha$.

What to do about p_3 and p_4?

- To obtain a definite solution despite this missing knowledge, we use the given equation as a constraint for the solution of an optimization problem.
- We are looking for a distribution $\mathbf{p}$ (for 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 exactly should the entropy function be maximized? Instead of fixing an ad hoc value for p_3 or p_4 and solve for the other, it is better to determine the values p_3 and p_4 such that the information added is minimal.
- Maximization of entropy minimizes the information content of the distribution.

Maximizing the Entropy

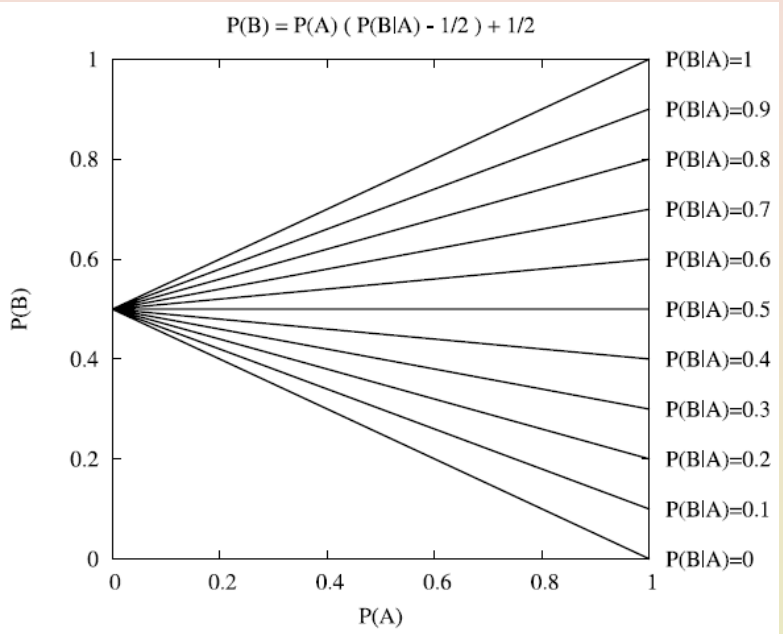
- To determine the maximum of the entropy under the constraint $p_3 + p_4 - 1 + \alpha = 0$, we use the method of Lagrange multipliers.
- The Lagrange function reads
  $L = -p_3 \ln p_3 - p_4 \ln p_4 + \lambda(p_3 + p_4 - 1 + \alpha)$.
- Taking the partial derivatives with respect to p_3 and p_4 we obtain

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

- So $p_3 = p_4 = \frac{1-\alpha}{2}$.
- This yields
  $P(B) = P(A,B) + P(\neg A, B) = p_1 + p_3 = \alpha\beta + \frac{1-\alpha}{2} = \alpha(\beta - \frac{1}{2}) + \frac{1}{2}$.
- Substituting in α and β yields $P(B) = P(A)(P(B|A) - \frac{1}{2}) + \frac{1}{2}$.

Looking at Some Specific Cases

- $P(B) = P(A)(P(B|A) - \frac{1}{2}) + \frac{1}{2}$.
- $P(B)$ is shown below for various values of $P(B|A)$.
- When $P(B)$ and $P(B|A)$ are $\{0, 1\}$-valued we obtain modus ponens. E.g., when A and $B|A$ are both true, B is also true.



- If $P(A) = 0$, $\neg A$ is true. Modus ponens cannot be applied, but our formula gives $\frac{1}{2}$ for $P(B)$ irrespective of $P(B|A)$. When A is false, we know nothing about B, which reflects our intuition exactly.
- If $P(A) = 1$ (A is true) and $P(B|A) = 0$ ($A \Rightarrow B$ false), then $A \Rightarrow \neg B$ is true, so B is false. Modus ponens holds again!
- The horizontal line in the figure means that we cannot make a prediction about B in the case of $P(B|A) = \frac{1}{2}$.

Maximum Entropy and Indifference

Theorem (Maximum Entropy)

Given a consistent set of linear probabilistic equations as constraints, there exists a unique maximum for the entropy function. The resulting MaxEnt distribution has minimum information content under the constraints.

- In the preceding calculation for $P(B)$, the two values p_3 and p_4 always occur symmetrically.

Definition of Indifference

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

- The indifference of p_3, p_4 leads to them being set equal by MaxEnt.

Theorem (Entropy for Indifferent Variables)

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

Max Entropy Under No Constraints

- What if only $p_1 + p_2 + \cdots + p_n = 1$ is known?
- All variables are indifferent. So $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$.
- This means that given a complete lack of knowledge, all worlds are equally probable, i.e., the distribution is uniform.
- Example: For two variables A and B it would be the case that $P(A, B) = P(A, \neg B) = P(\neg A, B) = P(\neg A, \neg B) = \frac{1}{4}$, from which $P(A) = P(B) = \frac{1}{2}$ and $P(B|A) = \frac{1}{2}$.
- If the value of a condition deviates from the one derived from the uniform distribution, the probabilities of the worlds shift.
- Example: If $P(B|A) = \beta$ is known, $P(A, B) = P(B|A)P(A) = \beta P(A)$. So $p_1 = \beta(p_1 + p_2)$ and we get the constraints:

$$\begin{aligned} \beta p_2 + (\beta - 1)p_1 &= 0 \\ p_1 + p_2 + p_3 + p_4 - 1 &= 0 \end{aligned}$$

The Lagrange equations are complicated, so a numeric solution yields $p_3 = p_4$, which was expected since p_3, p_4 are indifferent.

Conditional Probability Versus Material Implication

- The conditional probability and classical implication for the extreme cases of probabilities zero and one are compared:

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

- In both cases with false premises (which, intuitively, are critical cases), $P(B|A)$ is undefined, which makes sense.
- What is $P(B|A)$ when $P(A) = \alpha$ and $P(B) = \beta$ are given and no other information is known?

$P(B|A)$ when $P(A) = \alpha$ and $P(B) = \beta$

- We maximize entropy.
- We set
  $p_1 = P(A, B),\ p_2 = P(A, \neg B),\ p_3 = P(\neg A, B),\ p_4 = P(\neg A, \neg B).$
- Then, obtain as constraints

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

- We calculate using entropy maximization: $p_1 = \alpha\beta$, $p_2 = \alpha(1-\beta)$, $p_3 = \beta(1-\alpha)$, $p_4 = (1-\alpha)(1-\beta)$.
- From $p_1 = \alpha\beta$, it follows that $P(A, B) = P(A) \cdot P(B)$, which means that A and B are independent.
- Due to lack of constraints connecting A and B, the MaxEnt principle results in the independence of these variables.
- If $P(A) \neq 0$, from the definition $P(B|A) = \frac{P(A,B)}{P(A)}$ and the independence of A and B, it follows $P(B|A) = P(B)$.

MaxEnt Systems

- Due to the nonlinearity of the entropy function, MaxEnt optimization usually cannot be carried out symbolically.
- Two systems were developed for numerical entropy maximization
  - SPIRIT (Symmetrical Probabilistic Intensional Reasoning in Inference Networks in Transition) was built at FernUniversität Hagen.
  - PIT (Probability Induction Tool) was developed at the Munich Technical University.

The PIT MaxEnt System

- PIT uses the the sequential quadratic programming (SQP) method to find an extremum of the entropy function under the given constraints.
- As input, PIT expects data containing the constraints.
- The constraints $P(A) = \alpha$ and $P(B|A) = \beta$ have the form

$$\begin{aligned}&\texttt{var } A\{t, f\}, B\{t, f\};\\ &\texttt{P}([A = t]) = 0.6;\\ &\texttt{P}([B = t]|[A = t]) = 0.3;\\ &\texttt{QP}([B = t]);\\ &\texttt{QP}([B = t]|[A = t]);\end{aligned}$$

- The query $\texttt{QP}([B = t])$ indicates that $P(B)$ is the desired value.
- As a result we get

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

The Tweety Example in PIT

- We show, using Tweety, that probabilistic reasoning and, in particular, MaxEnt are non-monotonic and model everyday reasoning:
- We model the relevant rules:

$$P(\text{bird}|\text{penguin}) = 1 \quad \text{“penguins are birds”}$$
$$P(\text{flies}|\text{bird}) \in [0.95, 1] \quad \text{“(almost all) birds can fly”}$$
$$P(\text{flies}|\text{penguin}) = 0 \quad \text{“penguins cannot fly”}$$

- We input in PIT:

$\texttt{var}\ \text{penguin}\{\texttt{yes}, \texttt{no}\}, \text{bird}\{\texttt{yes}, \texttt{no}\}, \text{flies}\{\texttt{yes}, \texttt{no}\};$
$P([\text{bird} = \texttt{yes}]|[\text{penguin} = \texttt{yes}]) = 1;$
$P([\text{flies} = \texttt{yes}]|[\text{bird} = \texttt{yes}])\ \texttt{IN}\ [0.95, 1];$
$P([\text{flies} = \texttt{yes}]|[\text{penguin} = \texttt{yes}]) = 0;$
$\texttt{QP}([\text{flies} = \texttt{yes}]|[\text{penguin} = \texttt{yes}]);$

- We get back the correct answer

Nr.	Truthvalue	Probability	Query
1	UNSPECIFIED	$0.000e+00$	$\texttt{QP}([\text{penguin} = \texttt{yes}] - \mid > [\text{flies} = \texttt{yes}]);$

Subsection 6

Reasoning With Bayesian Networks

Why Bayesian Networks?

- If, in probability modeling, d variables $X_1, \ldots, X_d$ with n values each are used, then the associated probability distribution has n^d total values.
- This means that in the worst case the memory use and computation time for determining the specified probabilities grows exponentially with the number of variables.
- In practice the applications are usually very structured and the distribution contains many redundancies.
- So memory and time requirements can be heavily reduced with the appropriate methods.
- The use of Bayesian networks is one of the AI techniques which have been successfully used in practice to model applications with such redundancies.
- Bayesian networks utilize knowledge about the independence of (some pairs of) variables to simplify the model.

Independent Variables

- In the simplest case, all variables are pairwise independent and, therefore, $P(X_1, \ldots, X_d) = P(X_1) \cdot P(X_2) \cdot \cdots \cdot P(X_d)$.
- All entries in the distribution can thus be calculated from the d values $P(X_1), \ldots, P(X_d)$.
- Interesting applications, however, cannot be modeled because conditional probabilities become trivial: Since $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$, all conditional probabilities are reduced to the a priori probabilities.
- 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.
- We illustrate reasoning with Bayesian networks using a simple example by Judea Pearl, a UCLA pioneer in Bayesian Networks.

The Alarm Example I

- Bob has an alarm system installed in his house to protect against burglars. He cannot hear the alarm when he is working at the office, so he has asked his two neighbors John and Mary to call him at his office if they hear his alarm. After a few years Bob knows how reliable John and Mary are and models their calling behavior using conditional probability as follows:

$$P(\mathrm{J}|\mathrm{Al}) = 0.90, \quad P(\mathrm{M}|\mathrm{Al}) = 0.70,$$
$$P(\mathrm{J}|\neg\mathrm{Al}) = 0.05, \quad P(\mathrm{M}|\neg\mathrm{Al}) = 0.01.$$

- Mary is hard of hearing so she fails to hear the alarm more often than John.
- John sometimes mixes up the alarm at Bob's house with the alarms at other houses.

The Alarm Example II

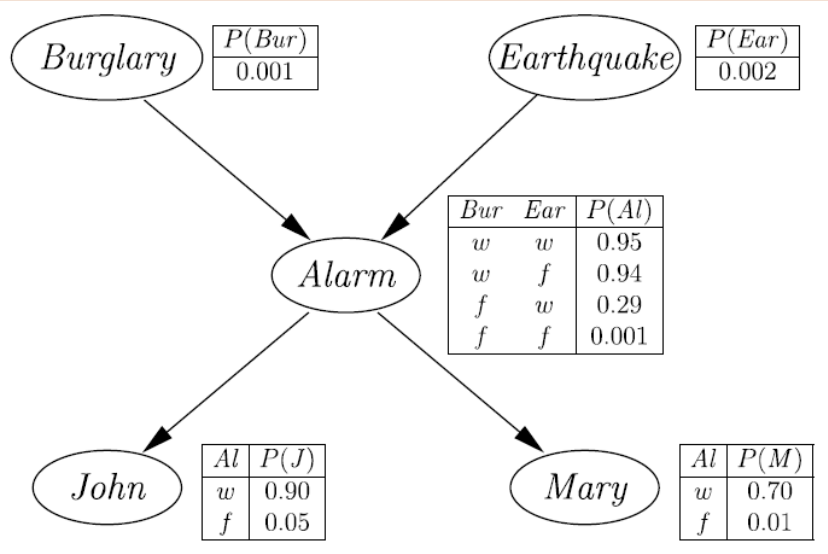
- The alarm is triggered by a burglary, but can also be triggered by a (weak) earthquake. These relationships are modeled by

$$P(\text{Al}|\text{Bur},\text{Ear}) = 0.95, \quad P(\text{Al}|\text{Bur},\neg\text{Ear}) = 0.94,$$
$$P(\text{Al}|\neg\text{Bur},\text{Ear}) = 0.29, \quad P(\text{Al}|\neg\text{Bur},\neg\text{Ear}) = 0.001,$$

  as well as the a priori probabilities $P(\text{Bur}) = 0.001$ and $P(\text{Ear}) = 0.002$.
- These two variables are independent:
  - Earthquakes do not depend on the habits of burglars.
  - Burglars do not plan based on earthquake predictions.
- Queries are now made against this knowledge base. For example, Bob might be interested in $P(\text{Bur}|\text{J} \vee \text{M})$, $P(\text{J}|\text{Bur})$ or $P(\text{M}|\text{Bur})$.

Graphical Representation

- A graphical representation of the Bayesian network:



- Each node represents a variable and every directed edge a statement of conditional probability.
- The edge from Al to J represents the two values $P(\text{J}|\text{Al})$ and $P(\text{J}|\neg\text{Al})$, which is given in the form of a **CPT** (**conditional probability table**).
- The CPT of a node lists all the conditional probabilities of the node's variable conditioned on all the nodes connected by incoming edges.
- Why are there no other edges included besides the four that are drawn in?
  - Nodes Bur and Ear are not linked since the variables are independent.
  - All other nodes have a parent node, which makes the reasoning a little more complex.

Conditional Independence

Conditionally Independent Random Variables

Two variables A and B are called **conditionally independent given** C if $P(A, B|C) = P(A|C) \cdot P(B|C)$.

- Example: Look at nodes J and M in the alarm example, which have the common parent node Al. If John and Mary independently react to an alarm, then the two variables J and M are independent given Al, that is: $P(\mathrm{J}, \mathrm{M}|\mathrm{Al}) = P(\mathrm{J}|\mathrm{Al}) \cdot P(\mathrm{M}|\mathrm{Al})$.
  - Because of the conditional independence of the two variables J and M, no edge between these two nodes is added.
  - However, J and M are not independent (unconditionally).
- The relationship between J and Bur is similar, because John does not react to a burglary, but only to the alarm. Thus J and Bur are independent given Al and $P(\mathrm{J}, \mathrm{Bur}|\mathrm{Al}) = P(\mathrm{J}|\mathrm{Al}) \cdot P(\mathrm{Bur}|\mathrm{Al})$.
- Given an alarm, the variables J and Ear, M and Bur, as well as M and Ear are also independent.

Characterizing Conditional Independence

Theorem (Characterization of Conditional Independence)

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

$$\begin{aligned} P(A, B|C) &= P(A|C) \cdot P(B|C), \\ P(A|B, C) &= P(A|C), \\ P(B|A, C) &= P(B|C). \end{aligned}$$

- On one hand, using conditional independence we can conclude that
$$P(A, B, C) = P(A, B|C)P(C) = P(A|C)P(B|C)P(C).$$
- On the other hand, the product (chain) rule gives us
$$P(A, B, C) = P(A|B, C)P(B|C)P(C).$$
- Thus $P(A|B, C) = P(A|C)$ is equivalent to the first equation.
- The last equation is obtained similarly.

Sensitivity of John and Mary I

- Bayesian networks can be used for reasoning.
- Bob can evaluate the sensitivity of John and Mary's reporting.
- Using the product rule and the conditional independence of J and Bur given Al:

$$P(\mathrm{J}|\mathrm{Bur}) = \frac{P(\mathrm{J},\mathrm{Bur})}{P(\mathrm{Bur})} = \frac{P(\mathrm{J},\mathrm{Bur},\mathrm{Al}) + P(\mathrm{J},\mathrm{Bur},\neg\mathrm{Al})}{P(\mathrm{Bur})}$$

and

$$\begin{aligned} P(\mathrm{J},\mathrm{Bur},\mathrm{Al}) &= P(\mathrm{J}|\mathrm{Bur},\mathrm{Al})P(\mathrm{Al}|\mathrm{Bur})P(\mathrm{Bur}) \\ &= P(\mathrm{J}|\mathrm{Al})P(\mathrm{Al}|\mathrm{Bur})P(\mathrm{Bur}). \end{aligned}$$

- Now, we get $P(\mathrm{J}|\mathrm{Bur})$

$$\begin{aligned} &= \frac{P(\mathrm{J}|\mathrm{Al})P(\mathrm{Al}|\mathrm{Bur})P(\mathrm{Bur}) + P(\mathrm{J}|\neg\mathrm{Al})P(\neg\mathrm{Al}|\mathrm{Bur})P(\mathrm{Bur})}{P(\mathrm{Bur})} \\ &= P(\mathrm{J}|\mathrm{Al})P(\mathrm{Al}|\mathrm{Bur}) + P(\mathrm{J}|\neg\mathrm{Al})P(\neg\mathrm{Al}|\mathrm{Bur}). \end{aligned}$$

- Here $P(\mathrm{Al}|\mathrm{Bur})$ and $P(\neg\mathrm{Al}|\mathrm{Bur})$ are missing.

Sensitivity of John and Mary II

- $P(\text{Al}|\text{Bur})$ and $P(\neg\text{Al}|\text{Bur})$ are missing. Therefore we calculate

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

  as well as $P(\neg\text{Al}|\text{Bur}) = 0.06$ and use this to get $P(\text{J}|\text{Bur}) = 0.90.94 + 0.050.06 = 0.849$.
- Analogously we calculate $P(\text{M}|\text{Bur}) = 0.659$.
- We now know that John calls for about 85% of all burglaries and Mary for about 66% of all burglaries.
- The probability of both of them calling is calculated, due to conditional independence, as $P(\text{J},\text{M}|\text{Bur}) = P(\text{J}|\text{Bur})P(\text{M}|\text{Bur}) = 0.849 \cdot 0.659 = 0.559$.

Probability that John or Mary Will Report

- For the probability of a call from John or Mary

$$\begin{aligned} P(\mathrm{J} \vee \mathrm{M}|\mathrm{Bur}) &= P(\neg(\neg\mathrm{J}, \neg\mathrm{M})|\mathrm{Bur}) = 1 - P(\neg\mathrm{J}, \neg\mathrm{M}|\mathrm{Bur}) \\ &= 1 - P(\neg\mathrm{J}|\mathrm{Bur})P(\neg\mathrm{M}|\mathrm{Bur}) = 1 - 0.051 = 0.948. \end{aligned}$$

  Bob thus receives notification for about 95% of all burglaries. Now to calculate $P(\mathrm{Bur}|\mathrm{J})$, we apply the Bayes formula, which gives us

$$P(\mathrm{Bur}|\mathrm{J}) = \frac{P(\mathrm{J}|\mathrm{Bur})P(\mathrm{Bur})}{P(\mathrm{J})} = \frac{0.849 \cdot 0.001}{0.052} = 0.016.$$

  Only about 1.6% of all calls from John are actually due to a burglary.
- Because the probability of false alarms is five times smaller for Mary, with $P(\mathrm{Bur}|\mathrm{M}) = 0.056$ we have significantly higher confidence given a call from Mary.
- Bob should only be seriously concerned about his home if both of them call, because $P(\mathrm{Bur}|\mathrm{J}, \mathrm{M}) = 0.284$.

Conditioning: "Sliding in" a New Variable

- We showed with

$$P(\text{J}|\text{Bur}) = P(\text{J}|\text{Al})P(\text{Al}|\text{Bur}) + P(\text{J}|\neg\text{Al})P(\neg\text{Al}|\text{Bur})$$

  how we can "slide in" a new variable.
- This relationship holds in general for two variables A and B given the introduction of an additional variable C and is called **conditioning**:

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

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

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

PIT, Bayesian Networks and MaxEnt

- Inputting the following in PIT:

```
 1 var Alarm{t,f}, Burglary{t,f}, Earthquake{t,f}, John{t,f}, Mary{t,f};
 2
 3 P([Earthquake=t]) = 0.002;
 4 P([Burglary=t]) = 0.001;
 5 P([Alarm=t] | [Burglary=t] AND [Earthquake=t]) = 0.95;
 6 P([Alarm=t] | [Burglary=t] AND [Earthquake=f]) = 0.94;
 7 P([Alarm=t] | [Burglary=f] AND [Earthquake=t]) = 0.29;
 8 P([Alarm=t] | [Burglary=f] AND [Earthquake=f]) = 0.001;
 9 P([John=t] | [Alarm=t]) = 0.90;
10 P([John=t] | [Alarm=f]) = 0.05;
11 P([Mary=t] | [Alarm=t]) = 0.70;
12 P([Mary=t] | [Alarm=f]) = 0.01;
13
14 QP([Burglary=t] | [John=t] AND [Mary=t]);
```

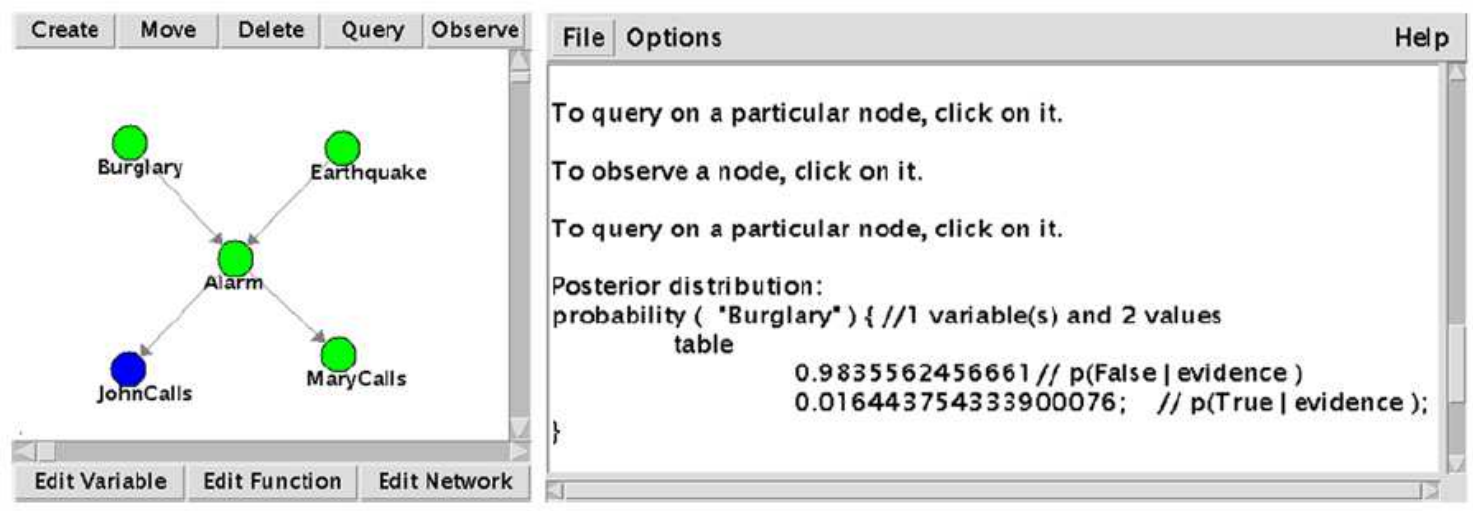
  we receive the answer:

$$P([\text{Burglary} = t]|[\text{John} = t] \texttt{ AND } [\text{Mary} = t]) = 0.2841.$$

- It can be shown that on input of CPTs or equivalent rules, the MaxEnt principle implies the same conditional independencies and, thus, also the same answers as a Bayesian network.
- Therefore, Bayesian networks are a special case of MaxEnt.

JavaBayes and Hugin

- JavaBayes has the graphical interface shown below:



  - With the graphical 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 window.
- The professional, commercial system Hugin is much more powerful:
  - It can use continuous variables in addition to discrete variables.
  - It can also learn Bayesian networks, that is, generate the network fully automatically from statistical data.

Probability Distributions vs Bayesian Networks

- A compact Bayesian network is very clear and significantly more informative for the reader than a full probability distribution.
- In addition, it requires much less memory.
- 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.
- In the alarm example the variables are all binary, so, for all variables $|v_i| = 2$, and the distribution has $2^5 - 1 = 31$ independent entries.
- For the number of independent entries for a Bayesian network: For a node v_i with k_i parent nodes $e_{i_1}, \ldots, e_{i_{k_i}}$, the associated CPT has $(|v_i| - 1) \prod_{j=1}^{k_i} |e_{ij}|$ entries. Then all CPTs have $\sum_{i=1}^{n} (|v_i| - 1) \prod_{j=1}^{k_i} |e_{ij}|$ entries.
- For the alarm example the result is then $2 + 2 + 4 + 1 + 1 = 10$.

A Rough Comparison

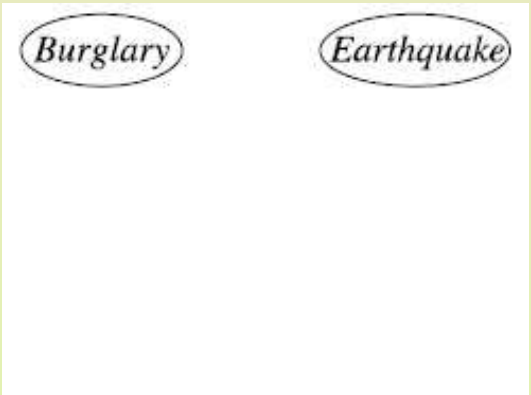
- We have

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

- The comparison in memory complexity between the full distribution and the Bayesian network becomes clearer if we assume all n variables have the same number b of values and each node has k parent nodes.
- Then the Bayesian Net equation can be simplified and all CPTs together have $n(b-1)b^k$ entries.
- The full distribution contains $b^n - 1$ entries.
- A significant gain is only made if the average number of parent nodes is much smaller than the number of variables. This means that the nodes are only locally connected.
- Because of the local connection, the network becomes modularized, which - as in software engineering - leads to a reduction in complexity.

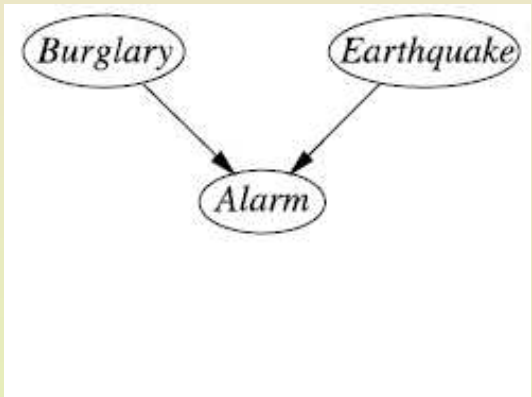
Causality and Network Structure I

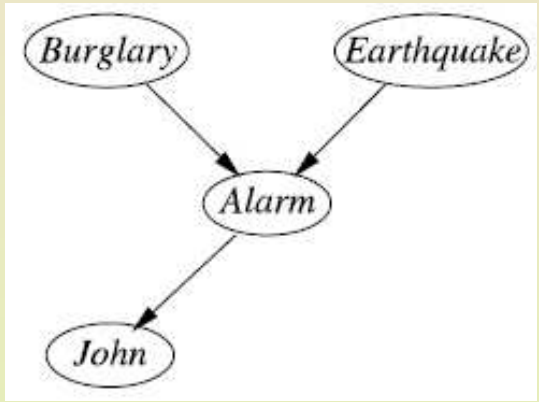
- Construction of a Bayesian network usually proceeds in two stages:
  1. Design of the network structure: usually performed manually.
  2. Entering the probabilities in the CPTs: Manually entering the values is very tedious. If a database is available, this step can be automated through estimation by counting frequencies.
- The alarm example: At the beginning we know the two causes Burglary and Earthquake and the two symptoms John and Mary.
- However, because John and Mary do not directly react to a burglar or earthquake, rather only to the alarm, we add Alarm.
- Adding edges starts with the causes (no parent nodes). In this case, Burglary & Earthquake:



Causality and Network Structure II

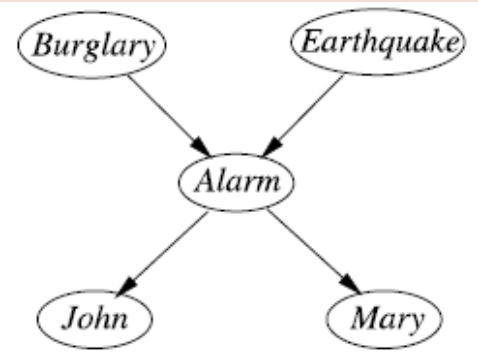
- Now we must check whether Earthquake is independent of Burglary. This is given, and thus no edge is added from Burglary to Earthquake.
- Because Alarm is directly dependent on Burglary and Earthquake, these variables are chosen next and an edge is added from both Burglary and Earthquake to Alarm.

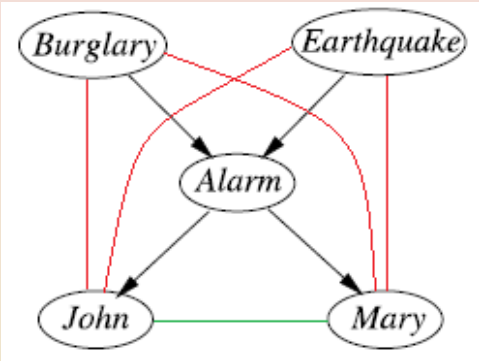




- Then we choose John. Because Alarm and John are not independent, an edge is added from Alarm to John. The same is true for Mary.

Causality and Network Structure III

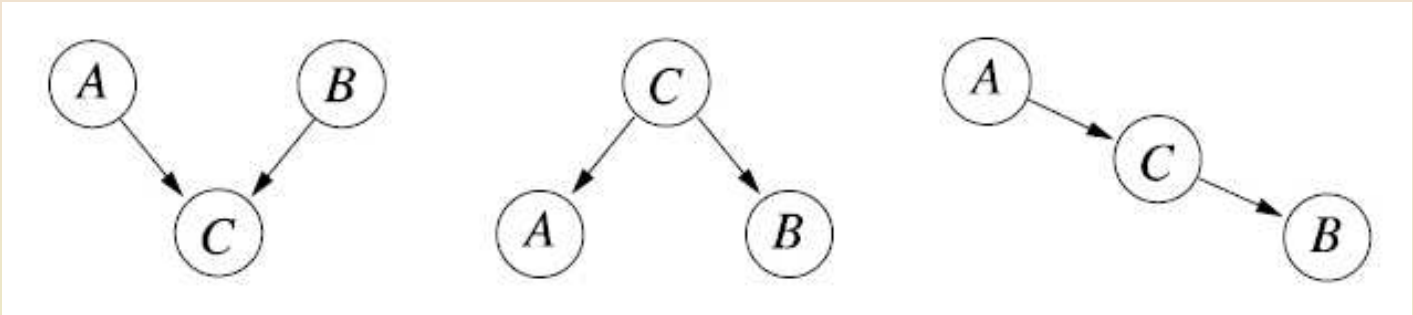




- Now we must check whether John is conditionally independent of Burglary given Alarm. If this is not the case, then another edge must be inserted from Burglary to John.
- We must also check whether edges are needed from Earthquake to John and from Burglary or Earthquake to Mary. Because of conditional independence, these four edges are not necessary.
- Edges between John and Mary are also unnecessary because John and Mary are conditionally independent given Alarm.
- The structure of the Bayesian network heavily depends on the chosen variable ordering: The order should reflect the causal relationship from causes towards diagnosis variables to get a simple network.

Semantics of the Networks

- No edge is added to a Bayesian network between A and B when they are independent or conditionally independent given a third variable C.



- Suppose the Bayesian network has no cycles and the variables are numbered such that no variable has a lower number than any variable that precedes it.
- Then, using all conditional independencies, we have

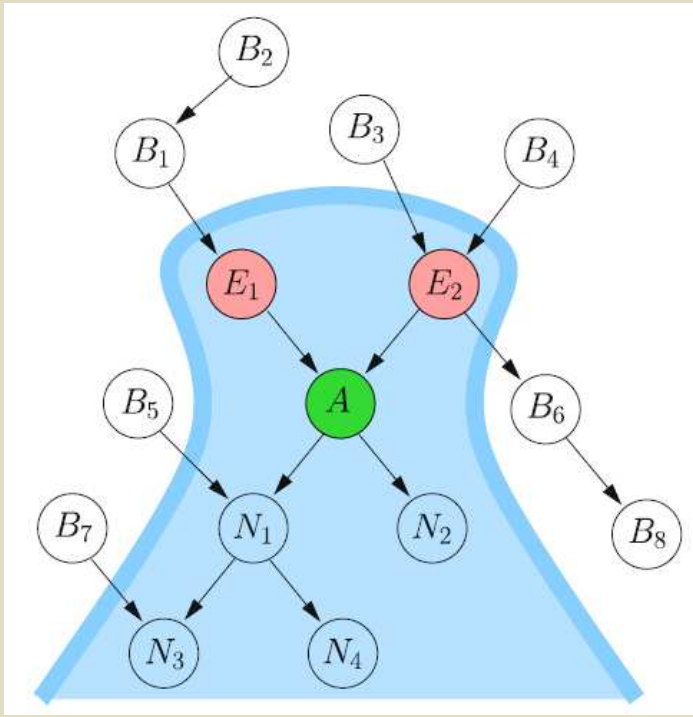
$$P(X_n|X_1,\ldots,X_{n-1}) = P(X_n|\text{Parents}(X_n)).$$

- This equation expresses that an arbitrary variable X_i is conditionally independent of its ancestors, given its parents.

Fundamental Theorem of Bayesian Networks

Theorem (Conditional Independence in Bayesian Networks)

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



- The chain rule now simplifies:

$$P(X_1, \ldots, X_n) = \sum_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) = \sum_{i=1}^{n} P(X_i | \text{Parents}(X_i)).$$

Definition of Bayesian Networks

- We now know the most important concepts and foundations of Bayesian networks:

Definition of Bayesian Network

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.
- For every variable A the CPT (table of conditional probabilities $P(A|\text{Parents}(A))$) is given.

- Two variables A and B are **conditionally independent given** C if $P(A, B|C) = P(A|C) \cdot P(B|C)$ or, equivalently, $P(A|B, C) = P(A|C)$.

The Properties of Bayesian Networks

- The following rules are true:
  - **Bayes' Theorem**: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
  - **Marginalization**:
    $P(B) = P(A, B) + P(\neg A, B) = P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)$
  - **Conditioning**: $P(A|B) = \sum_c P(A|B, C = c)P(C = c|B)$
- 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 not successors of X_n, then

    $$P(X_n|X_1, \ldots, X_{n-1}) = P(X_n|\text{Parents}(X_n)).$$

    This condition must be honored during the construction of a network.
- During construction the variables should be ordered according to causality. First the causes, then the hidden variables, and, finally, the diagnosis variables.
- **Chain rule**: $P(X_1, \ldots, X_n) = \sum_{i=1}^{n} P(X_i|\text{Parents}(X_i))$