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

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

Laplace Probability

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

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

Example 1

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

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

Quiz

Two fair dice are rolled. Let S be the sum of their results. A student applies the Laplace formula to the possible values $\{2, 3, \ldots, 12\}$ and concludes that $P(S = 7) = 1/11$. Which statement is correct?

(a) The conclusion is correct because S has 11 possible values.

(b) The conclusion is incorrect because the possible values of S are not equally likely.

(c) The conclusion is incorrect because the Laplace formula applies only to elementary events.

Definition

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

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

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

Under the *Laplace assumption*, this can also be written as

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

Definition

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

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

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

Theorem 1

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

- *Product Rule:*

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

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

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

Quiz

Let A and B be events with $P(A) > 0$. Suppose that

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

Which conclusion must be true?

(a) A causes B.

(b) B causes A.

(c) A and B are not independent.

Quiz

Let A be any event, and suppose that $P(B) = 0$. Which statement matches the definition used in this lecture?

(a) $P(A \mid B) = P(A)$, so A and B are independent.

(b) $P(A \mid B)$ is undefined, but A and B are independent.

(c) $P(A \mid B)$ is undefined, so A and B are dependent.

Example 2 (Simplified Alarm Model)

Al The alarm sounds.

Bur A burglary occurs.

$\mathbf{P}(Bur, Al)$	Bur	$\neg Bur$	Total
Al	0.0099	0.0990	0.1089
$\neg Al$	0.0001	0.8910	0.8911
Total	0.0100	0.9900	1

For example, it holds:

$$\begin{aligned} P(Al) &= P(Al, Bur) + P(Al, \neg Bur) \\ &= 0.0099 + 0.0990 = 0.1089, \end{aligned}$$

$$P(Al \mid Bur) = \frac{P(Al, Bur)}{P(Bur)} = \frac{0.0099}{0.0100} = 0.99.$$

Theorem 2 (Bayes' Theorem)

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

Example 2 (continued)

$$P(\mathit{Bur} \mid \mathit{Al}) = \frac{P(\mathit{Al} \mid \mathit{Bur})P(\mathit{Bur})}{P(\mathit{Al})} = \frac{0.99 \cdot 0.0100}{0.1089}$$

$$= \frac{0.0099}{0.1089} \approx 0.0909.$$

- $P(\mathit{Al} \mid \mathit{Bur})$: known burglary $\to$ chance that the alarm sounds; the alarm's *sensitivity*.
- $P(\mathit{Bur} \mid \mathit{Al})$: known alarm $\to$ chance of a burglary; *depends on the local burglary rate and false alarms*.
- Here 0.0099 is the $(\mathit{Al}, \mathit{Bur})$ cell, while 0.1089 is the *Al* row total.

Quiz

The same alarm system is used in two areas. For each $i \in \{1, 2\}$, $P_i(\mathit{Al} \mid \mathit{Bur}) = 0.99$ and $P_i(\mathit{Al} \mid \neg \mathit{Bur}) = 0.10$. However, $P_1(\mathit{Bur}) = 0.01$ and $P_2(\mathit{Bur}) = 0.001$. Consider the probability of burglary given an alarm, $P_i(\mathit{Bur} \mid \mathit{Al})$. Which statement is correct?

(a) It is the same in both areas.

(b) It is higher in Area 1.

(c) The alarm is less sensitive in Area 2.

The Principle of Maximum Entropy

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

Definition

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

- Entropy is a *measure of the uncertainty* associated with a random variable.
  - The *higher the entropy*, the *more uncertain or unpredictable the variable* is.
  - If one outcome has probability 1 and all others 0, then the entropy is 0 (no uncertainty).
  - If all outcomes are equally likely, then the entropy is maximized (maximum uncertainty).
- The *maximum-entropy principle (MaxEnt)* selects, among all distributions satisfying the stated constraints, a distribution that maximizes H [Ertel 2025], p. 141.

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.

- 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!

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

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

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

The Principle of Maximum Entropy
An Inference Rule for Probabilities

- The *full entropy* is

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

- Define the free part

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

  Then $H(\mathbf{p}) = H_{\text{free}}(p_3, p_4) + C$, where the constant $C = -p_1 \ln p_1 - p_2 \ln p_2$ does *not* depend on p_3, p_4.

- Therefore maximizing $H(\mathbf{p})$ over p_3, p_4 is the same as maximizing $H_{\text{free}}(p_3, p_4)$ under $p_3 + p_4 = 1 - \alpha$.

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

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

- Taking the partial derivatives with respect to p_3 and p_4

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

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

Quiz

Suppose $P(A) = 0.4$ and $P(B \mid A) = 0.75$. Under the MaxEnt solution above, what is $P(B \mid \neg A)$?

(a) 0.30

(b) 0.50

(c) 0.75

The Principle of Maximum Entropy
Maximum Entropy Without Explicit Constraints

- No knowledge is given, so no world is distinguished from the others.
- No constraints besides 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.

Quiz

For $\mathbf{P}(A, B) = (p_1, p_2, p_3, p_4)$, suppose the only constraint besides normalization is $P(B \mid A) = 1$. Which distribution is selected by MaxEnt?

(a) $\mathbf{P}(A, B) = (1/4, 1/4, 1/4, 1/4)$.

(b) $\mathbf{P}(A, B) = (1/3, 0, 1/3, 1/3)$.

(c) $\mathbf{P}(A, B) = (1/2, 0, 0, 1/2)$.

A *Bayesian network* is a DAG with one CPT for each node; its graph encodes conditional-independence assumptions. In Ertel's presentation, a fully specified Bayesian network is a *structured special case of MaxEnt*: it adds independence assumptions and complete CPTs [Ertel 2025], p. 172.

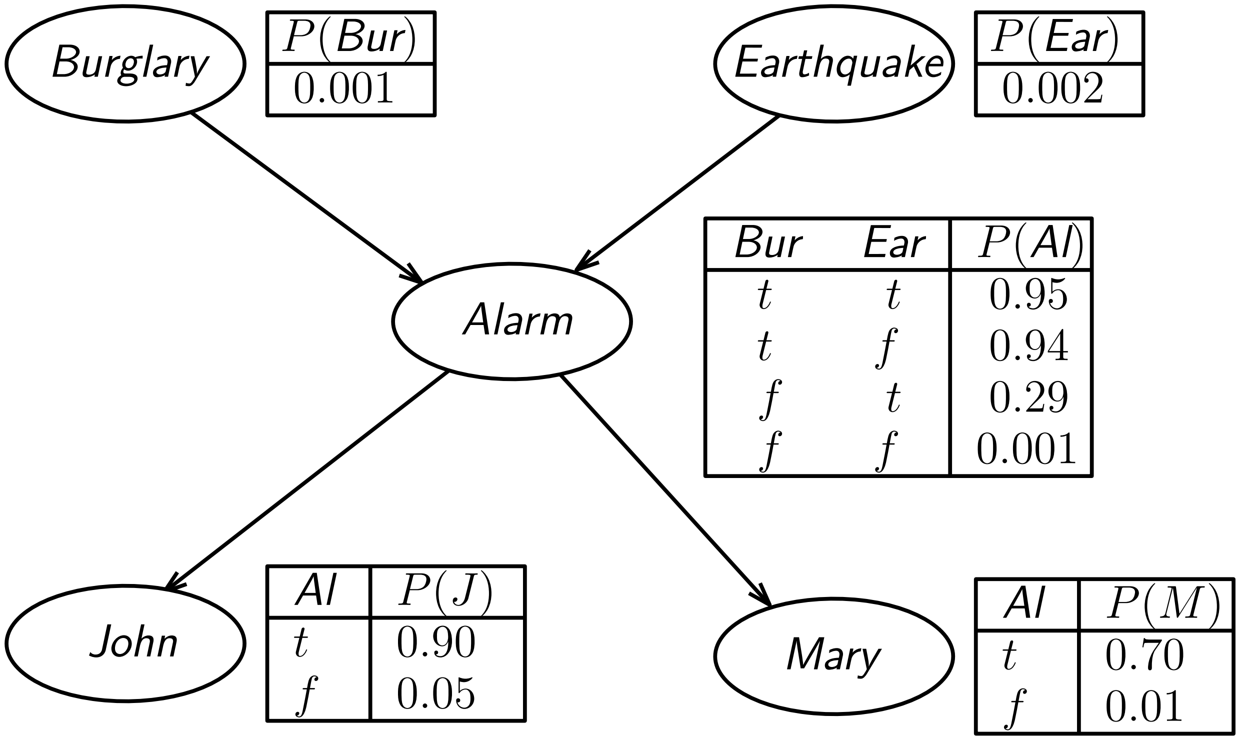


Figure: Alarm network and its CPTs.

Definition

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

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

Remark

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

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

Example 3 (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*.

Quiz

In the alarm example, *Bur* and *Ear* are independent. Suppose it is known that the alarm was triggered. Which relation is correct?

(a) $P(\mathit{Bur} \mid \mathit{Al}, \mathit{Ear}) = P(\mathit{Bur} \mid \mathit{Al}, \neg \mathit{Ear})$.

(b) $P(\mathit{Bur} \mid \mathit{Al}, \mathit{Ear}) > P(\mathit{Bur} \mid \mathit{Al}, \neg \mathit{Ear})$.

(c) $P(\mathit{Bur} \mid \mathit{Al}, \mathit{Ear}) < P(\mathit{Bur} \mid \mathit{Al}, \neg \mathit{Ear})$.

Since *Bur* and *Ear* are independent, conditional marginalization gives

$$
\begin{aligned}
P(\mathit{Al} \mid \mathit{Bur}) &= P(\mathit{Al} \mid \mathit{Bur}, \mathit{Ear}) P(\mathit{Ear}) + P(\mathit{Al} \mid \mathit{Bur}, \neg \mathit{Ear}) P(\neg \mathit{Ear}) \\
&= 0.95(0.002) + 0.94(0.998) = 0.94002, \\
P(\neg \mathit{Al} \mid \mathit{Bur}) &= 1 - P(\mathit{Al} \mid \mathit{Bur}) = 1 - 0.94002 = 0.05998.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P(J \mid \mathit{Bur}) &= P(J \mid \mathit{Al}) P(\mathit{Al} \mid \mathit{Bur}) + P(J \mid \neg \mathit{Al}) P(\neg \mathit{Al} \mid \mathit{Bur}) \\
&= 0.9 \cdot 0.94002 + 0.05 \cdot 0.05998 = 0.849017, \\
P(M \mid \mathit{Bur}) &= P(M \mid \mathit{Al}) P(\mathit{Al} \mid \mathit{Bur}) + P(M \mid \neg \mathit{Al}) P(\neg \mathit{Al} \mid \mathit{Bur}) \\
&= 0.7 \cdot 0.94002 + 0.01 \cdot 0.05998 = 0.6586138.
\end{aligned}
$$

Using the same marginalization,

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

Quiz

Use the practical calculations above to test whether J and M are independent given *Bur*. Which relation is correct?

(a) $P(J, M \mid \textit{Bur}) = P(J \mid \textit{Bur})P(M \mid \textit{Bur})$.

(b) $P(J, M \mid \textit{Bur}) > P(J \mid \textit{Bur})P(M \mid \textit{Bur})$.

(c) $P(J, M \mid \textit{Bur}) < P(J \mid \textit{Bur})P(M \mid \textit{Bur})$.

Construction of a Bayesian network

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

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

Quiz

For the alarm example, use the altered variable ordering $(M, \mathit{Al}, \mathit{Ear}, \mathit{Bur}, J)$. When *Bur* is added, which parent set satisfies the parent-selection condition and is **inclusion-minimal** (no selected parent can be removed)?

(a) $\mathrm{Parents}(\mathit{Bur}) = \{\mathit{Al}\}$.

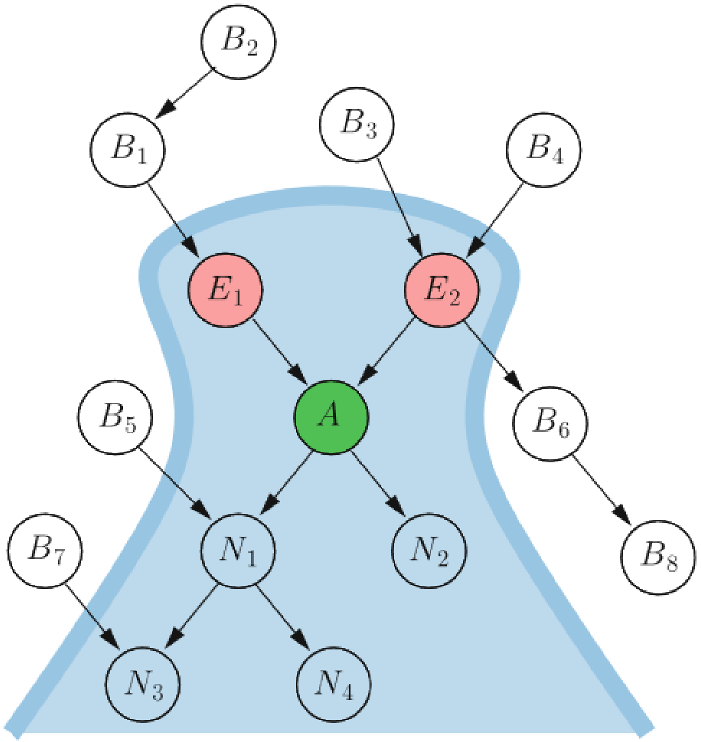
(b) $\mathrm{Parents}(\mathit{Bur}) = \{\mathit{Al}, \mathit{Ear}\}$.

(c) $\mathrm{Parents}(\mathit{Bur}) = \{M, \mathit{Al}, \mathit{Ear}\}$.

Theorem 3

Given its parents, a node in a Bayesian network is conditionally independent of the set of all other nodes that are neither its descendants nor its parents [Ertel 2025], p. 170.

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



Quiz

For the completed alarm network, which equality follows from Theorem 3?

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

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

(c) $\mathbf{P}(\mathit{Bur} \mid \mathit{Al}, \mathit{Ear}) = \mathbf{P}(\mathit{Bur} \mid \mathit{Al})$.

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

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

- Applying this factorization to the completed alarm network gives

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

Quiz

In the completed alarm network, which expression computes $P(J, M, \textit{Bur})$? Here $a \in \{\textit{Al}, \neg\textit{Al}\}$ and $e \in \{\textit{Ear}, \neg\textit{Ear}\}$.

(a) $P(\textit{Bur})P(J \mid \textit{Bur})P(M \mid \textit{Bur})$.

(b) $P(\textit{Bur})P(\textit{Ear})P(\textit{Al} \mid \textit{Bur}, \textit{Ear})P(J \mid \textit{Al})P(M \mid \textit{Al})$.

(c) $\sum_{a,e} P(\textit{Bur})P(e)P(a \mid \textit{Bur}, e)P(J \mid a)P(M \mid a)$.

Conditional Marginalization

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

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

- If C and B are independent, then $P(c \mid B) = P(c)$.

Quiz

In the alarm network, remove *Ear* while keeping all probabilities involving only *Bur*, *Al*, J, M unchanged. Recall that *Bur* and *Ear* are independent. For $b \in \{\textit{Bur}, \neg\textit{Bur}\}$ and $e \in \{\textit{Ear}, \neg\textit{Ear}\}$, which update is correct?

(a) Only the alarm CPT changes, using $P(\textit{Al} \mid b) = P(\textit{Al} \mid b, \neg\textit{Ear})$.

(b) Only the alarm CPT changes, using $P(\textit{Al} \mid b) = \sum_e P(\textit{Al} \mid b, e)P(e)$.

(c) The alarm, John, and Mary CPTs must all be recomputed by marginalizing over *Ear*.

References

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