Artificial Intelligence

15. Probabilistic Reasoning, Part I: Basics

(Our Machinery for) Thinking About What is Likely to be True

Jana Koehler Álvaro Torralba

SAARLAND UNIVERSITY
COMPUTER SCIENCE

Summer Term 2019

Thanks to Prof. Hoffmann for slide sources

Agenda

1. Introduction
2. Unconditional Probabilities
3. Conditional Probabilities
4. Independence
5. Basic Probabilistic Reasoning Methods
6. Bayes' Rule
7. Conditional Independence
8. Conclusion

Decision-Making Under Uncertainty

Example

Giving a lecture:

- Goal: Be in lecture hall at 10:15
- Possible plans:
  - P_1: *Get up at 8:00, leave at 8:40, arrive at 9:00.*
  - P_2: *Get up at 9:50, leave at 10:05, arrive at 10:15.*
- Decision: Both plans are correct, but P_2 succeeds only with probability 50%, and giving a lecture is important, so P_1 is the plan of choice.

Better Example: Which train to take to Frankfurt airport?

Uncertainty and Logic

Diagnosis: We want to build an expert dental diagnosis system, that deduces the cause (the disease) from the symptoms.

→ Can we base this on logic?

Attempt 1: Say we have a toothache. How's about:

$$\forall p[Symptom(p, toothache) \rightarrow Disease(p, cavity)]$$

→ Is this rule correct? No, toothaches may have different causes ("cavity" = "Loch im Zahn").

Attempt 2: So what about this:

$$\forall p[Symptom(p, toothache) \rightarrow \\ Disease(p, cavity) \vee Disease(p, gum_disease) \vee \ldots]$$

→ We don't know all possible causes.
→ And we'd like to be able to deduce which causes are more plausible!

Uncertainty and Logic, ctd.

Attempt 3: Perhaps a *causal* rule is better?

$$\forall p[Disease(p, cavity) \rightarrow Symptom(p, toothache)]$$

- Is this rule correct? No, not all cavities cause toothaches.
- Does this rule allow to deduce a cause from a symptom? No, setting $Symptom(p, toothache)$ to true here has no consequence on the truth of $Disease(p, cavity)$. [Note: If $Symptom(p, toothache)$ is *false*, we would conclude $\neg Disease(p, cavity)$ ... which would be incorrect, cf. previous question.]
- Anyway, this still doesn't allow to compare the plausibility of different causes.

$\rightarrow$ Logic does not allow to weigh different alternatives, and it does not allow to express incomplete knowledge ("cavity does not always come with a toothache, nor vice versa").

Unreliable Sensors

Robot Localization: Suppose we want to support localization using landmarks to narrow down the area.

→ "If you see the Eiffel tower, then you're in Paris."

Difficulty: Sensors perceptions can be unreliable.

- Even if a landmark is perceived, we cannot conclude with certainty that the robot is at that location. ("This is the half-scale Las Vegas copy, you dummy.")
- Even if a landmark is *not* perceived, we cannot conclude with certainty that the robot is *not* at that location. ("Top of Eiffel tower hidden in the clouds.")

→ Only the probability of being at a location increases or decreases.

Beliefs and Probabilities

What do we model with probabilities? Incomplete knowledge! We are not 100% sure, but we *believe to a certain degree* that something is true.

→ Probability ≈ Our degree of belief, given our current knowledge.

Example (Diagnosis)

- $Symptom(p, toothache) \rightarrow Disease(p, cavity)$ with 80% probability.
- But, for any given p, in reality we do, or do not, have cavity: 1 or 0! → The "probability" depends on our knowledge! The "80%" refers to the fraction of cavity, within the set of all p' that are indistinguishable from p based on our knowledge.
- If we receive new knowledge (e.g., $Disease(p, gum_disease)$), the probability changes!

→ Probabilities represent and measure the uncertainty that stems from lack of knowledge.

Questionnaire

Question!

What are sources of uncertainty in your life?

(A): Not knowing whether a train will be late.

(B): Not knowing what the exam questions will be.

(C): Not knowing whether the road can safely be crossed.

(D): Not knowing the outcome of a dice throw.

→ (A): Yes.

→ (B): Yes. Note that this depends on the agent's perspective/knowledge (there's no uncertainty here for *me*).

→ (C): In the usual sense of the wording: No, because you got full observability on this one. If we take into account exceptional cases, like a Meteorite that may hit you on the head, or a pit hidden beneath the asphalt that you may fall into, then yes.

→ (D): Yes. Note that one could argue this is due to partial observability: if you were able to observe all the relevant physical details ...

How to Obtain Probabilities?

Assessing probabilities through statistics:

- The agent is 90% convinced by its sensor information
  $:=$ in 9 out of 10 cases, the information is correct.
- $Disease(p, cavity) \rightarrow Symptom(p, toothache)$ with 80% probability
  $:=$ 8 out of 10 persons with a cavity have toothache.

$\rightarrow$ The process of estimating a probability P using statistics is called assessing P. Assessing even a single P can require huge effort! (Eg. "The likelihood of making it to the university within 10 minutes")

What is probabilistic reasoning? Deducing probabilities from knowledge about *other* probabilities.

$\rightarrow$ Probabilistic reasoning determines, based on probabilities that are (relatively) easy to assess, probabilities that are difficult to assess.

(Uncertainty and Rational Decisions)

Here: We're only concerned with deducing the likelihood of facts, not with action choice. In general, selecting actions is of course important.

Rational Agents:

- We have a choice of actions (go to FRA early, go to FRA just in time).
- These can lead to different solutions with different probabilities.
- The actions have different costs.
- The results have different utilities (safe timing/dislike airport food).

$\rightarrow$ A rational agent chooses the action with the maximum expected utility.

$\rightarrow$ Decision Theory = Utility Theory + Probability Theory.

(Decision-Theoretic Agent)

A particular kind of utility-based agent:

function DT-AGENT(*percept*) **returns** an *action*
  persistent: *belief_state*, probabilistic beliefs about the current state of the world
    action, the agent's action

  update *belief_state* based on *action* and *percept*
  calculate outcome probabilities for actions,
    given action descriptions and current *belief_state*
  select *action* with highest expected utility
    given probabilities of outcomes and utility information
  return *action*

Our Agenda for This Topic

→ Our treatment of the topic "Probabilistic Reasoning" consists of Chapters 15 and 16.

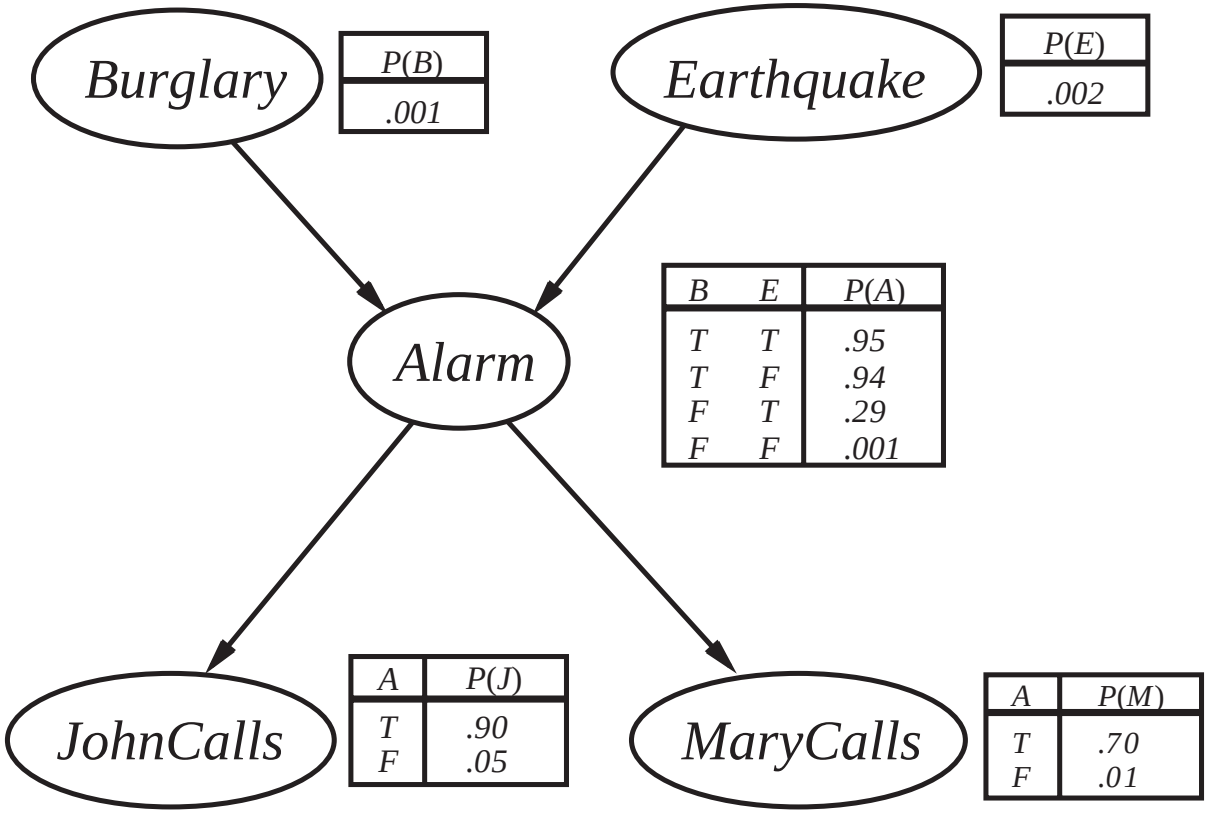
- **This Chapter:** All the basic machinery at use in Bayesian networks.

  → Sets up the framework and basic operations.

- **Chapter 16:** Bayesian networks: What they are, how to build them, how to use them.

  → The most wide-spread and successful practical framework for probabilistic reasoning.

This is Where We're Headed . . .



Our Agenda for This Chapter

- **Unconditional Probabilities** and **Conditional Probabilities:** Which concepts and properties of probabilities will be used?

  → Mostly a recap of things you're familiar with from school.

- **Independence** and **Basic Probabilistic Reasoning Methods:** What simple methods are there to avoid enumeration and to deduce probabilities from other probabilities?

  → A basic tool set we'll need. (Still familiar from school?)

- **Bayes' Rule:** What's that "Bayes"? How is it used and why is it important?

  → The basic insight about how to invert the "direction" of conditional probabilities.

- **Conditional Independence:** How to capture and exploit complex relations between random variables?

  → Explains the difficulties arising when using Bayes' rule on multiple evidences. Conditional independence is used to ameliorate these difficulties.

Unconditional Probabilities

Definition. *Given a random variable X, $P(X = x)$ denotes the unconditional probability, or prior probability, that X has value x in the absence of any other information.*

Example: $P(Cavity = true) = 0.2$, where $Cavity$ is a random variable whose value is true iff some given person has a cavity.

Notation/Terminology:

- We will refer to the fact $X = x$ as an event, or an outcome.
- The notation uppercase "X" for a variable, and lowercase "x" for one of its values will be used frequently. (Follows Russel/Norvig)

Random Variables

Note: In general, random variables can have arbitrary domains. Here, we consider finite-domain random variables only, and Boolean random variables most of the time.

Example:

$$P(Weather = sunny) = 0.7$$
$$P(Weather = rain) = 0.2$$
$$P(Weather = cloudy) = 0.08$$
$$P(Weather = snow) = 0.02$$
$$P(Headache = true) = 0.1$$

- By convention, we denote Boolean random variables with A, B, and more general finite-domain random variables with X, Y.
- For Boolean variable $Name$, we write $name$ for $Name = true$ and $\neg name$ for $Name = false$. (Follows Russel/Norvig)

Probability Distributions

Definition. *The probability distribution for a random variable X, written $\mathbf{P}(X)$, is the vector of probabilities for the (ordered) domain of X.*

Example:

$$\mathbf{P}(Headache) = \langle 0.1, 0.9 \rangle$$
$$\mathbf{P}(Weather) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$$

Terminology: Given a subset $\mathbf{Z} \subseteq \{X_1, \ldots, X_n\}$ of random variables, an event is an assignment of values to the variables in $\mathbf{Z}$. The joint probability distribution, written $\mathbf{P}(\mathbf{Z})$, lists the probabilities of all events.

Example: $\mathbf{P}(Headache, Weather) =$

	$Headache = true$	$Headache = false$
$Weather = sunny$	$P(W = sunny \wedge headache)$	$P(W = sunny \wedge \neg headache)$
$Weather = rain$		
$Weather = cloudy$		
$Weather = snow$		

The Full Joint Probability Distribution

Terminology:

- Given random variables $\{X_1, \ldots, X_n\}$, an atomic event is an assignment of values to all variables.
- Given random variables $\{X_1, \ldots, X_n\}$, the full joint probability distribution, denoted $\mathbf{P}(X_1, \ldots, X_n)$, lists the probabilities of all atomic events.

Example:

	$toothache$	$\neg toothache$
$cavity$	0.12	0.08
$\neg cavity$	0.08	0.72

$\rightarrow$ All atomic events are disjoint (their pairwise conjunctions all are $\perp$); the sum of all fields is 1 (corresponds to their disjunction $\top$).

Probabilities of Propositional Formulas

Definition. *Given random variables* $\{X_1, \ldots, X_n\}$*, a propositional formula, short proposition, is a propositional formula over the atoms* $X_i = x_i$ *where* x_i *is a value in the domain of* X_i*. A function* P *that maps propositions into* $[0, 1]$ *is a probability measure if (i)* $P(\top) = 1$ *and (ii) for all propositions* A*,* $P(A) = \sum_{\mathbf{e} \models A} P(\mathbf{e})$ *where* $\mathbf{e}$ *is an atomic event.*

$\rightarrow$ Propositions represent sets of atomic events: the interpretations satisfying the formula.

Example: $P(cavity \wedge toothache) = 0.12$ is the probability that some given person has both a cavity and a toothache. (Recall the use of $cavity$ for $Cavity = true$ and $toothache$ for $Toothache = true$.)

Notation:

- Instead of $P(a \wedge b)$, we often write $P(a, b)$.
- Propositions can be viewed as Boolean random variables; we will denote them with A, B as well.

Questionnaire

Theorem (Kolmogorov). *A function P that maps propositions into $[0,1]$ is a probability measure if and only if (i) $P(\top) = 1$ and (ii') for all propositions A, B: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$.*

$\rightarrow$ We can equivalently replace "(ii) for all propositions A, $P(A) = \sum_{I \models A} P(I)$" (cf. previous slide) with Kolmogorovs (ii').

Question!

Assume we have (iii) $P(\bot) = 0$. How to derive from (i), (ii'), and (iii) that, for all propositions A, $P(\neg a) = 1 - P(a)$?

$\rightarrow$ By (i), $P(\top) = 1$; as $a \vee \neg a \equiv \top$, we get $P(a \vee \neg a) = 1$. By (iii), $P(\bot) = 0$; as $a \wedge \neg a \equiv \bot$, we get $P(a \wedge \neg a) = 0$. Inserting this into (ii'), we get $P(a \vee \neg a) = 1 = P(a) + P(\neg a) - 0$.

Questionnaire, ctd.

Reminder 1: (i) $P(\top) = 1$; (ii') $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$.

Reminder 2: "Probabilities model our belief."

$\rightarrow$ If P represents an objectively observable probability, the axioms clearly make sense. But why should an agent respect these axioms, when modeling its subjective own belief?

Question!

Do you believe in Kolmogorov's axioms?

(A): Yes. (B): No.

$\rightarrow$ You're free to believe whatever you want, but note this (de Finetti, 1931): If an agent has a belief that violates Kolmogorov's axioms, then there exists a combination of "bets" on propositions so that the agent *always* loses money.

$\rightarrow$ If your beliefs are contradictory, then you will not be successful in the long run (and even the next minute if your opponent is clever).

Conditional Probabilities: Intuition

→ Do probabilities change as we gather new knowledge? Yes!
Probabilities model our *belief*, thus they depend on our knowledge.

Example: Your "probability of missing the connection train" increases when you are informed that your current train has 30 minutes delay. The "probability of cavity" increases when the doctor is informed that the patient has a toothache.

- In the presence of additional information, we can no longer use the unconditional (*prior!*) probabilities.
- Given propositions A and B, $P(a \mid b)$ denotes the conditional probability of a (i.e., $A = true$) given that all we know is b (i.e., $B = true$).

Example: $P(cavity) = 0.2$ vs. $P(cavity \mid toothache) = 0.6$. And $P(cavity \mid toothache \wedge \neg cavity) = 0$.

Conditional Probabilities: Definition

Definition. *Given propositions A and B where $P(b) \neq 0$, the conditional probability, or posterior probability, of a given b, written $P(a \mid b)$, is defined as:*

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$$

$\rightarrow$ The likelihood of having a and b, within the set of outcomes where we have b.

Example: $P(cavity \wedge toothache) = 0.12$ and $P(toothache) = 0.2$ yield $P(cavity \mid toothache) = 0.6$.

Conditional Probability Distributions

Definition. *Given random variables X and Y, the conditional probability distribution of X given Y, written $\mathbf{P}(X \mid Y)$, is the table of all conditional probabilities of values of X given values of Y.*

→ For sets of variables: $\mathbf{P}(X_1, \ldots, X_n \mid Y_1, \ldots, Y_m)$.

Example: $\mathbf{P}(Weather \mid Headache) =$

	$Headache = true$	$Headache = false$
$Weather = sunny$	$P(W = sunny \mid headache)$	$P(W = sunny \mid \neg headache)$
$Weather = rain$		
$Weather = cloudy$		
$Weather = snow$		

→ "The probability of sunshine given that I have a headache?" If you're susceptible to headaches depending on weather conditions, this makes sense. Otherwise, the two variables are independent (see next section).

Working with the Full Joint Probability Distribution

Example:

	$toothache$	$\neg toothache$
$cavity$	0.12	0.08
$\neg cavity$	0.08	0.72

→ How to compute $P(cavity)$? Sum across the row:

$$P(cavity \wedge toothache) + P(cavity \wedge \neg toothache)$$

→ How to compute $P(cavity \vee toothache)$? Sum across atomic events:

$$P(cavity \wedge toothache) + P(\neg cavity \wedge toothache) + P(cavity \wedge \neg toothache)$$

→ How to compute $P(cavity \mid toothache)$? $\frac{P(cavity \wedge toothache)}{P(toothache)}$

→ All relevant probabilities can be computed using the full joint probability distribution, by expressing propositions as disjunctions of atomic events.

Working with the Full Joint Probability Distribution??

→ So, is it a good idea to use the full joint probability distribution? No:

- Given n random variables with k values each, the joint probability distribution contains k^n probabilities.
- Computational cost of dealing with this size.
- *Practically impossible to assess all these probabilities.*

→ So, is there a compact way to represent the full joint probability distribution? Is there an efficient method to work with that representation?

→ Not in general, but it works in many cases. We can work directly with conditional probabilities, and exploit (conditional) independence.

→ Bayesian networks. (First, we do the simple case.)

Independence

Definition. *Events a and b are independent if* $P(a \wedge b) = P(a)P(b)$.

Proposition. *Given independent events a and b where $P(b) \neq 0$, we have* $P(a \mid b) = P(a)$.

Proof. By definition, $P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$, which by independence is equal to $\frac{P(a)P(b)}{P(b)} = P(a)$. ($\rightarrow$ Similarly, if $P(a) \neq 0$, we have $P(b \mid a) = P(b)$.)

Examples:

- $P(Dice1 = 6 \wedge Dice2 = 6) = 1/36$.
- $P(W = sunny \mid headache) = P(W = sunny)$ unless you're weather-sensitive (cf. slide 26).
- But $toothache$ and $cavity$ are NOT independent. The fraction of "$cavity$" is higher within "$toothache$" than within "$\neg toothache$". $P(toothache) = 0.2$ and $P(cavity) = 0.2$, but $P(toothache \wedge cavity) = 0.12 > 0.04$.

Definition. *Random variables X and Y are independent if* $\mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$. (System of equations!)

Illustration: Exploiting Independence

Example:

	$toothache$	$\neg toothache$
$cavity$	0.12	0.08
$\neg cavity$	0.08	0.72

Adding variable $Weather$ with values $sunny$, $rain$, $cloudy$, $snow$, the full joint probability distribution contains 16 probabilities.

→ But your teeth do not influence the weather, nor vice versa!

- $Weather$ is independent of each of $Cavity$ and $Toothache$: For all value combinations c, t of $Cavity$ and $Toothache$, and for all values w of $Weather$, we have $P(c \wedge t \wedge w) = P(c \wedge t)P(w)$.
- $\mathbf{P}(Cavity, Toothache, Weather)$ can be reconstructed from the separate tables $\mathbf{P}(Cavity, Toothache)$ and $\mathbf{P}(Weather)$. (8 probabilities)

→ Independence can be exploited to represent the full joint probability distribution more compactly.

→ Usually, random variables are independent only under particular conditions: conditional independence, see later.

The Product Rule

Proposition (Product Rule). *Given propositions A and B,* $P(a \wedge b) = P(a \mid b)P(b)$. (Direct from definition.)

Example: $P(cavity \wedge toothache) = P(toothache \mid cavity)P(cavity)$.

→ If we know the values of $P(a \mid b)$ and $P(b)$, then we can compute $P(a \wedge b)$.

→ Similarly, $P(a \wedge b) = P(b \mid a)P(a)$.

Notation: $\mathbf{P}(X, Y) = \mathbf{P}(X \mid Y)\mathbf{P}(Y)$ is a system of equations:

$$\begin{aligned}
P(W = sunny \wedge headache) &= P(W = sunny \mid headache)P(headache) \\
P(W = rain \wedge headache) &= P(W = rain \mid headache)P(headache) \\
\ldots &= \ldots \\
P(W = snow \wedge \neg headache) &= P(W = snow \mid \neg headache)P(\neg headache)
\end{aligned}$$

→ Similar for unconditional distributions, $\mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$.

The Chain Rule

Proposition (Chain Rule). *Given random variables $X_1, \ldots, X_n$, we have* $\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n \mid X_{n-1}, \ldots, X_1) * \mathbf{P}(X_{n-1} \mid X_{n-2}, \ldots, X_1) * \cdots * \mathbf{P}(X_2 \mid X_1) * \mathbf{P}(X_1)$.

Example: $P(\neg brush \wedge cavity \wedge toothache)$
$= P(toothache \mid cavity, \neg brush) P(cavity, \neg brush)$
$= P(toothache \mid cavity, \neg brush) P(cavity \mid \neg brush) P(\neg brush)$.

Proof. Iterated application of Product Rule. $\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n \mid X_{n-1}, \ldots, X_1) * \mathbf{P}(X_{n-1}, \ldots, X_1)$ by Product Rule. In turn, $\mathbf{P}(X_{n-1}, \ldots, X_1) = \mathbf{P}(X_{n-1} \mid X_{n-2}, \ldots, X_1) * \mathbf{P}(X_{n-2}, \ldots, X_1)$, etc.

Note: This works *for any ordering* of the variables.

$\rightarrow$ We can recover the probability of atomic events from sequenced conditional probabilities for any ordering of the variables.

$\rightarrow$ First of the four basic techniques in Bayesian networks.

Marginalization

→ Extracting a sub-distribution from a larger joint distribution:

Proposition (Marginalization). *Given sets* $\mathbf{X}$ *and* $\mathbf{Y}$ *of random variables, we have:*

$$\mathbf{P}(\mathbf{X}) = \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(\mathbf{X}, \mathbf{y})$$

where $\sum_{\mathbf{y} \in \mathbf{Y}}$ *sums over all possible value combinations of* $\mathbf{Y}$*.*

Example: (Note: Equation system!)

$$\mathbf{P}(Cavity) = \sum_{y \in Toothache} \mathbf{P}(Cavity, y)$$

$$P(cavity) = P(cavity, toothache) + P(cavity, \neg toothache)$$

$$P(\neg cavity) = P(\neg cavity, toothache) + P(\neg cavity, \neg toothache)$$

Questionnaire

Question!

Say $P(dog) = 0.4$, $\neg dog \leftrightarrow cat$, **and** $P(likeslasagna \mid cat) = 0.5$. **Then** $P(likeslasagna \wedge cat) =$

(A): 0.2 (B): 0.5

(C): 0.475 (D): 0.3

$\rightarrow$ We have $P(cat) = 0.6$ and $P(likeslasagna \mid cat) = 0.5$, hence (D) by the product rule.

Question!

Can we compute the value of $P(likeslasagna)$**, given the above informations?**

(A): Yes. (B): No.

$\rightarrow$ No. We don't know the probability that *dogs* like lasagna, i.e., $P(likeslasagna \mid dog)$.

Normalization: Idea

Problem: We know $P(cavity \wedge toothache)$ but don't know $P(toothache)$:

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.12}{P(toothache)}$$

Step 1: Case distinction over the values of $Cavity$:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.08}{P(toothache)}$$

Step 2: Assuming placeholder $\alpha := 1/P(toothache)$:

$$P(cavity \mid toothache) = \alpha P(cavity \wedge toothache) = \alpha 0.12$$
$$P(\neg cavity \mid toothache) = \alpha P(\neg cavity \wedge toothache) = \alpha 0.08$$

Step 3: Fixing $toothache$ to be true, view $P(cavity \wedge toothache)$ vs. $P(\neg cavity \wedge toothache)$ as the relative weights of $P(cavity)$ vs. $P(\neg cavity)$ within $toothache$. Then normalize their summed-up weight to 1:

$$1 = \alpha(0.12 + 0.08) \Rightarrow \alpha = 1/(0.12 + 0.08) = 1/0.2 = 5$$

$\rightarrow$ α is the normalization constant scaling the sum of relative weights to 1.

Normalization: Formal

Definition. *Given a vector* $\langle w_1, \ldots, w_k \rangle$ *of numbers in* $[0,1]$ *where* $\sum_{i=1}^{k} w_i \leq 1$, *the normalization constant* α *is* $\alpha\langle w_1, \ldots, w_k \rangle := 1/\sum_{i=1}^{k} w_i$.

Example: $\alpha\langle 0.12, 0.08 \rangle = 5\langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$.

Proposition (Normalization). *Given a random variable* X *and an event* $\mathbf{e}$, *we have* $\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e})$.

Proof. For each value x of X, $P(X = x \mid \mathbf{e}) = P(X = x \wedge \mathbf{e})/P(\mathbf{e})$. So all we need to prove is that $\alpha = 1/P(\mathbf{e})$. By definition, $\alpha = 1/\sum_x P(X = x \wedge \mathbf{e})$, so we need to prove $P(\mathbf{e}) = \sum_x P(X = x \wedge \mathbf{e})$ which holds by Marginalization.

Example: $\alpha\langle P(cavity \wedge toothache), P(\neg cavity \wedge toothache) \rangle = \alpha\langle 0.12, 0.08 \rangle$, so $P(cavity \mid toothache) = 0.6$, and $P(\neg cavity \mid toothache) = 0.4$.

Normalization+Marginalization: Given "query variable" X, "observed event" $\mathbf{e}$, and "hidden variables" set $\mathbf{Y}$: $\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$.

$\rightarrow$ Second of the four basic techniques in Bayesian networks.

Questionnaire

Question!

Say we know $P(likeschappi \wedge dog) = 0.32$ **and** $P(\neg likeschappi \wedge dog) = 0.08$**. Can we compute** $P(likeschappi \mid dog)$**?**

(A): Yes. (B): No.

$\rightarrow$ Yes, because we can compute $P(dog) = P(likeschappi \wedge dog) + P(\neg likeschappi \wedge dog)$, and thus $P(likeschappi \mid dog) = \frac{P(likeschappi \wedge dog)}{P(dog)}$.

$\rightarrow$ In other words, we can use Normalization: $\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e})$.

Inserting $LikesChappi$ for X and dog for $\mathbf{e}$, we get $\mathbf{P}(LikesChappi \mid dog) = \alpha \mathbf{P}(LikesChappi, dog) = \alpha \langle P(likeschappi \wedge dog), P(\neg likeschappi \wedge dog) \rangle = \alpha \langle 0.32, 0.08 \rangle$.

$\rightarrow$ So what is $P(likeschappi \mid dog)$? 0.8, because $\alpha = 1/P(dog) = 1/(0.32 + 0.08) = 2.5$.

Bayes' Rule

Proposition (Bayes' Rule). *Given propositions A and B where $P(a) \neq 0$ and $P(b) \neq 0$, we have:*

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$$

Proof. By definition, $P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$ which by product rule $P(a \wedge b) = P(b \mid a)P(a)$ is equal to the claim.

Notation: (System of equations)

$$\mathbf{P}(X \mid Y) = \frac{\mathbf{P}(Y \mid X)\mathbf{P}(X)}{\mathbf{P}(Y)}$$

Applying Bayes' Rule

Example: Say we know that $P(toothache \mid cavity) = 0.6$, $P(cavity) = 0.2$, and $P(toothache) = 0.2$.

$\rightarrow$ We can compute $P(cavity \mid toothache)$: By Bayes' rule,
$P(cavity \mid toothache) = \frac{P(toothache|cavity)P(cavity)}{P(toothache)} = \frac{0.6*0.2}{0.2} = 0.6.$

Ok, but: Why don't we simply assess $P(cavity \mid toothache)$ directly?

- $P(toothache \mid cavity)$ is causal, $P(cavity \mid toothache)$ is diagnostic.
- Causal dependencies are robust over frequency of the causes.
  $\rightarrow$ Example: If there is a cavity epidemic then $P(cavity \mid toothache)$ increases, but $P(toothache \mid cavity)$ remains the same.
- Also, causal dependencies are often easier to assess.

$\rightarrow$ Bayes' rule allows to perform diagnosis (observing a symptom, what is the cause?) based on prior probabilities and causal dependencies.

Questionnaire

Question!

Say $P(dog) = 0.4$**,** $P(likeschappi \mid dog) = 0.8$**, and** $P(likeschappi) = 0.5$**. What is** $P(dog \mid likeschappi)$**?**

(A): 0.8 (B): 0.64

(C): 0.9 (D): 0.32

→ By Bayes' rule,
$P(dog \mid likeschappi) = \frac{P(likeschappi|dog)P(dog)}{P(likeschappi)} = \frac{0.8*0.4}{0.5} = 0.64$ so (B).

→ Is $P(likeschappi \mid dog)$ causal or diagnostic? Causal; liking or not liking dog food may be caused by being or not being a dog.

→ Is $P(dog \mid likeschappi)$ causal or diagnostic? Diagnostic; liking Chappi does not cause anybody to be a dog.

Bayes' Rule with Multiple Evidence

Example: Say we know from medicinical studies that $P(cavity) = 0.2$, $P(toothache \mid cavity) = 0.6$, $P(toothache \mid \neg cavity) = 0.1$, $P(catch \mid cavity) = 0.9$, and $P(catch \mid \neg cavity) = 0.2$. Now, in case we did observe the symptoms toothache and catch (the dentist's probe catches in the aching tooth), what would be the likelihood of having a cavity? What is $P(cavity \mid catch, toothache)$?

By Bayes' rule we get:

$$P(cavity \mid catch, toothache) = \frac{P(catch, toothache \mid cavity)P(cavity)}{P(catch, toothache)}$$

Question!

So, is everything fine? Do we just need some more medicinical studies?

(A): Yes. (B): No.

$\rightarrow$ No! We would need $\mathbf{P}(toothache \wedge catch \mid Cavity)$, i.e., causal dependencies for all combinations of symptoms! ($\gg 2$, in general)

Bayes' Rule with Multiple Evidence, ctd.

Second attempt: First Normalization (slide 38), then Chain Rule (slide 34) using ordering $X_1 = Cavity, X_2 = Catch, X_3 = Toothache$:

$$\mathbf{P}(Cavity \mid catch, toothache) =$$
$$\alpha \mathbf{P}(Cavity, catch, toothache) =$$
$$\alpha \mathbf{P}(toothache \mid catch, Cavity)\mathbf{P}(catch \mid Cavity)\mathbf{P}(Cavity)$$

Close, but no Banana: Less red (i.e.unknown) probabilities, but still $\mathbf{P}(toothache \mid catch, Cavity)$.

But: Are $Toothache$ and $Catch$ independent?

$\rightarrow$ No. If a probe catches, we probably have a cavity which probably causes toothache.

But: They are independent given the presence or absence of cavity!

$\rightarrow$ See next slide.

Conditional Independence

Definition. *Given sets of random variables* $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}$, *we say that* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$ *are conditionally independent given* $\mathbf{Z}$ *if:*

$$\mathbf{P}(\mathbf{Z}_1, \mathbf{Z}_2 \mid \mathbf{Z}) = \mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z})\mathbf{P}(\mathbf{Z}_2 \mid \mathbf{Z})$$

We alternatively say that $\mathbf{Z}_1$ *is conditionally independent of* $\mathbf{Z}_2$ *given* $\mathbf{Z}$.

Example:

$$\mathbf{P}(Toothache, Catch \mid cavity) = \mathbf{P}(Toothache \mid cavity)\mathbf{P}(Catch \mid cavity)$$
$$\mathbf{P}(Toothache, Catch \mid \neg cavity) = \mathbf{P}(Toothache \mid \neg cavity)\mathbf{P}(Catch \mid \neg cavity)$$

$\rightarrow$ For $cavity$: this may cause both, but they don't influence each other.
For $\neg cavity$: $catch$ and/or $toothache$ would each be caused by something else.

Note: The definition is symmetric regarding the roles of $\mathbf{Z}_1$ and $\mathbf{Z}_2$: $Toothache$ is conditionally independent of $Catch$, *and vice versa*.

Conditional Independence, ctd.

Proposition. *If* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$ *are conditionally independent given* $\mathbf{Z}$*, then* $\mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z}_2, \mathbf{Z}) = \mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z})$.

Proof. By definition, $\mathbf{P}(\mathbf{Z}_1 \mid \mathbf{Z}_2, \mathbf{Z}) = \frac{\mathbf{P}(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z})}{\mathbf{P}(\mathbf{Z}_2,\mathbf{Z})}$ which by product rule is equal to $\frac{\mathbf{P}(\mathbf{Z}_1,\mathbf{Z}_2|\mathbf{Z})\mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbf{Z}_2,\mathbf{Z})}$ which by prerequisite is equal to $\frac{\mathbf{P}(\mathbf{Z}_1|\mathbf{Z})\mathbf{P}(\mathbf{Z}_2|\mathbf{Z})\mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbf{Z}_2,\mathbf{Z})}$. Since $\frac{\mathbf{P}(\mathbf{Z}_2|\mathbf{Z})\mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbf{Z}_2,\mathbf{Z})} = 1$ this proves the claim.

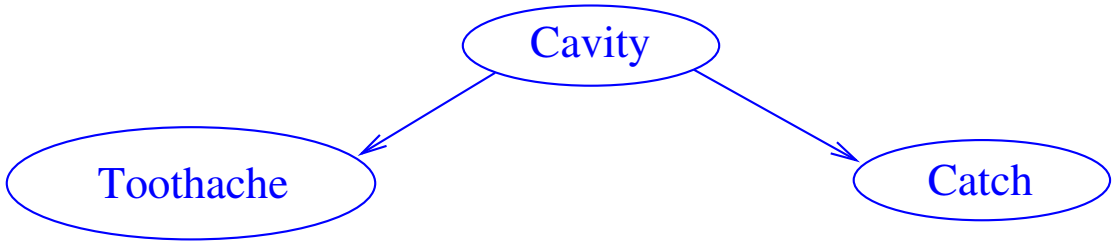
Example: Using $\{Toothache\}$ as $\mathbf{Z}_1$, $\{Catch\}$ as $\mathbf{Z}_2$, and $\{Cavity\}$ as $\mathbf{Z}$: $\mathbf{P}(Toothache \mid Catch, Cavity) = \mathbf{P}(Toothache \mid Cavity)$.

$\rightarrow$ In the presence of conditional independence, we can drop variables from the right-hand side of conditional probabilities.

$\rightarrow$ Third of the four basic techniques in Bayesian networks. Last missing technique: "Capture variable dependencies in a graph"; illustration see next slide, details see **Next Chapter**.

Exploiting Conditional Independence: Overview

1. **Graph captures variable dependencies:** (Variables $X_1, \ldots, X_n$)



$\rightarrow$ Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X \mid e)$. Remaining vars: $\mathbf{Y}$.

2. **Normalization+Marginalization:**
$\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e})$; if $\mathbf{Y} \neq \emptyset$ then $\mathbf{P}(X \mid \mathbf{e}) = \alpha \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$

$\rightarrow$ A sum over atomic events!

3. **Chain rule:** Order $X_1, \ldots, X_n$ consistently with dependency graph.

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n \mid X_{n-1}, \ldots, X_1) * \mathbf{P}(X_{n-1} \mid X_{n-2}, \ldots, X_1) * \cdots * \mathbf{P}(X_1)$$

4. **Exploit conditional independence:** Instead of $\mathbf{P}(X_i \mid X_{i-1}, \ldots, X_1)$, with previous slide we can use $\mathbf{P}(X_i \mid Parents(X_i))$.

$\rightarrow$ Bayesian networks!

Exploiting Conditional Independence: Example

1. **Graph captures variable dependencies:** (See previous slide.)
$\rightarrow$ Given $toothache$, $catch$, want $\mathbf{P}(Cavity \mid toothache, catch)$.
Remaining vars: $\emptyset$.

2. **Normalization+Marginalization:**
$$\mathbf{P}(Cavity \mid toothache, catch) = \alpha \mathbf{P}(Cavity, toothache, catch)$$

3. **Chain rule:** Order $X_1 = Cavity$, $X_2 = Toothache$, $X_3 = Catch$.
$$\mathbf{P}(Cavity, toothache, catch) =$$
$$\mathbf{P}(catch \mid toothache, Cavity)\mathbf{P}(toothache \mid Cavity)\mathbf{P}(Cavity)$$

4. **Exploit conditional independence:**
Instead of $\mathbf{P}(catch \mid toothache, Cavity)$ use $\mathbf{P}(catch \mid Cavity)$.

Thus: $\mathbf{P}(Cavity \mid toothache, catch) =$
$\alpha \mathbf{P}(catch \mid Cavity)\mathbf{P}(toothache \mid Cavity)\mathbf{P}(Cavity) =$
$\alpha\langle 0.9 * 0.6 * 0.2, 0.2 * 0.1 * 0.8\rangle = \alpha\langle 0.108, 0.016\rangle$. So $\alpha \approx 8.06$ and $\mathbf{P}(cavity \mid toothache \wedge catch) \approx 0.87$.

Questionnaire

Question!

Consider the random variables $X_1 = Animal$, $X_2 = LikesChappi$, and $X_3 = LoudNoise$; X_1 has values $\{dog, cat, other\}$, X_2 and X_3 are Boolean. Which statements are correct?

(A): $Animal$ is independent of $LikesChappi$.

(B): $LoudNoise$ is independent of $LikesChappi$.

(C): $Animal$ is conditionally independent of $LikesChappi$ given $LoudNoise$.

(D): $LikesChappi$ is conditionally independent of $LoudNoise$ given $Animal$.

→ (A) No: $likeschappi$ indicates dog.

→ (B) No: Not knowing what animal it is, $loudnoise$ is an indication for dog which indicates $likeschappi$.

→ (C) No: For example, even if we know $loudnoise$, knowing in addition that $likeschappi$ gives us a stronger indication of $Animal = dog$.

→ (D) Yes: If we already know what animal it is, also knowing $LoudNoise$ does not influence $LikesChappi$ nor vice versa.

Summary

- Uncertainty is unavoidable in many environments, namely whenever agents do not have perfect knowledge.
- Probabilities express the degree of belief of an agent, given its knowledge, into an event.
- Conditional probabilities express the likelihood of an event given observed evidence.
- Assessing a probability means to use statistics to approximate the likelihood of an event.
- Bayes' rule allows us to derive, from probabilities that are easy to assess, probabilities that aren't easy to assess.
- Given multiple evidence, we can exploit conditional independence.

  $\rightarrow$ Bayesian networks (up next) do this, in a comprehensive manner.

Reading

- *Chapter 13: Quantifying Uncertainty* [Russell and Norvig (2010)].

  Content: Sections 13.1 and 13.2 roughly correspond to my "Introduction" and "Probability Theory Concepts". Section 13.3 and 13.4 roughly correspond to my "Basic Probabilistic Inference". Section 13.5 roughly corresponds to my "Bayes' Rule" and "Multiple Evidence".

  In Section 13.6, RN go back to the Wumpus world and discuss some inferences in a probabilistic version thereof.

  Overall, the content is quite similar. I have added some examples, have tried to make a few subtle points more explicit, and I indicate already how these techniques will be used in Bayesian networks. RN gives many complementary explanations, nice as additional background reading.

References I

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (Third Edition)*. Prentice-Hall, Englewood Cliffs, NJ, 2010.