GHENT UNIVERSITY

FACULTY OF ENGINEERING AND ARCHITECTURE

E016350 - Artificial Intelligence

Lecture 9

Reasoning under Uncertainty & Bayesian ML

Intro to Probabilistic Reasoning

Aleksandra Pizurica

Ghent University

Spring 2025

Overview

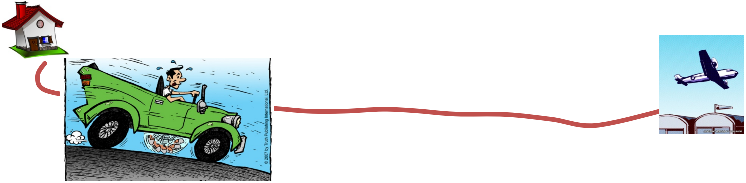
- Uncertainty
- Probability
- Marginalization
- Independence and Bayes' Rule
- Inference

[R&N], Chapter 12

This presentation is based on: S. Russel and P. Norvig: *Artificial Intelligence: A Modern Approach*, (Fourth Ed.), denoted as [R&N] and the course Artificial Intelligence UC Berkeley

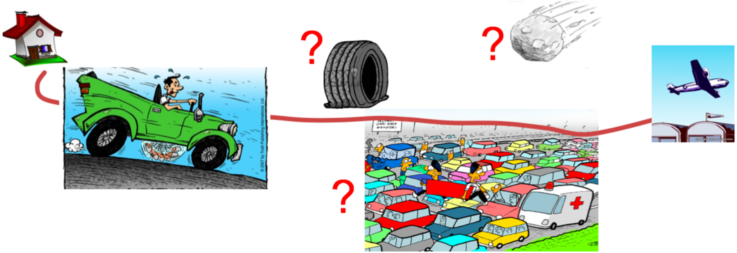
Why do we need reasoning under uncertainty?

Let A_t denote the action "leave for airport t minutes before flight"
Will A_t get me there on time?



Why do we need reasoning under uncertainty?

Let A_t denote the action "leave for airport t minutes before flight"
Will A_t get me there on time?



Purely logical approach

- risks falsehood, e.g.,: "A_{90} gets me on time"
- or leads to weak conclusions, e.g.: "A_{90} gets me on time if no accidents on the way and it doesn't rain and I don't get a flat tire and no meteorite hits the car, etc." (success of the plan cannot be inferred)

Why do we need reasoning under uncertainty?

Consider making a diagnosis for a patient with headache. Many reasons are possible: sinus problems, eye vision, tense muscles ... A logical rule that attempts to express this:

$$Headache \implies Sinusitis \vee EyeSight \vee StiffNeck \vee Flu \vee Cancer \vee \ldots$$

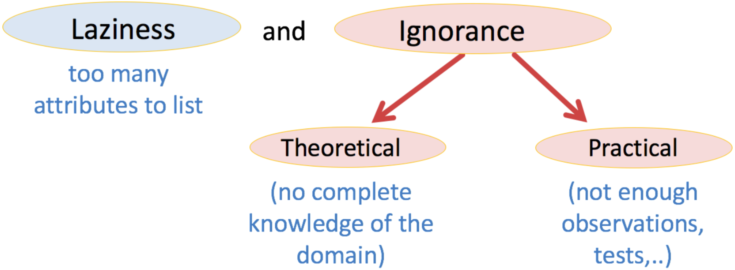
Doesn't work because the list of possible causes is practically unlimited.
The causal rule like $StiffNeck \implies Headache$ doesn't work either (stiff neck doesn't always cause headache).

Trying to use logic in these domains fails because

- there is too much work to list all the attributes
- no complete theory or knowledge
- not all the necessary tests can be or have been run

Probabilistic reasoning as a remedy when logic fails

Logic often fails due to inability to list all the attributes, for different reasons that can be grouped as follows



In this view, we can say that
probabilistic assertions summarize the effects of “laziness” and “ignorance”

Probabilistic reasoning

A consistent framework for dealing with degrees of belief

We don't know for sure the cause to a given manifestation but we know that there is a certain chance (or probability) of a given cause

(e.g. 80% of patients with toothache have a cavity

$\rightarrow$ the patient with a toothache has a cavity with probability 0.8)

Probabilities relate propositions to one's own state of knowledge e.g.,

$P(A_{120}|\text{no reported accidents}) = 0.6$

Probabilities of propositions change with new evidence, e.g.,

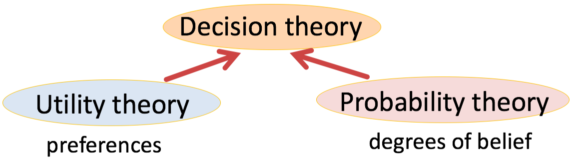
$P(A_{120}|\text{no reported accidents} \ \wedge \ 4am) = 0.8$

Uncertainty and rational decisions

Let A_t denote the action "leave for airport t minutes before flight".
Suppose I believe the following:

$$P(A_{30} \text{ gets me there on time}|\ldots) = 0.05$$
$$P(A_{120} \text{ gets me there on time}|\ldots) = 0.75$$
$$P(A_{180} \text{ gets me there on time}|\ldots) = 0.95$$
$$P(A_{1440} \text{ gets me there on time}|\ldots) = 0.9999$$

Which action to choose?
Depends on my preferences for missing flight vs. waiting at the airport.



Digression: Maximum expected utility (MEU) principle

Fundamental idea of decision theory: **choose an action that yields MEU**

Definition (Rational agent and the MEU principle)

An agent is **rational** if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action. This is called the **principle of maximum expected utility (MEU)**.

Let $U(s)$ denote the utility of state s. Expected utility of action a under evidence $\mathbf{e}$ is

$$EU(a|\mathbf{e}) = \sum_{s} P(\textsc{Result}(a) = s|a, \mathbf{e})U(s)$$

MEU and rational action under evidence $\mathbf{e}$:

$$MEU(\mathbf{e}) = \max_{a} EU(a|\mathbf{e}); \quad action = \arg\max_{a} EU(a|\mathbf{e})$$

This is basis for **reinforcement learning** (Part 2 of the 6-credit version of the course)

Probability basics and notation

A set Ω – the sample space

e.g., 6 possible rolls of a die.

$\omega \in \Omega$ is a sample point/possible world/atomic event

A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ subject to

$0 \leq P(\omega) \leq 1$

$\sum_\omega P(\omega) = 1$

e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event A is any subset of Ω

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

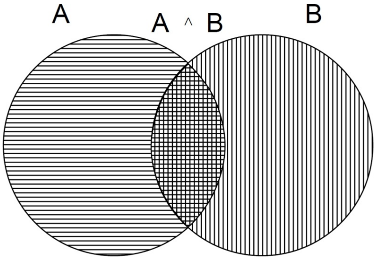
Probability axioms

$0 \leq P(\omega) \leq 1$ for every ω and $\sum_{\omega \in \Omega} P(\omega) = 1$

For any proposition ϕ, $P(\phi) = \sum_{\omega \in \phi} P(\omega)$

$\Rightarrow P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Basic axioms of probability



Probability axioms

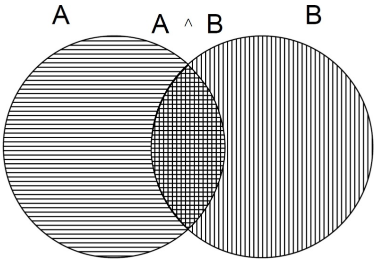
$0 \leq P(\omega) \leq 1$ for every ω and $\sum_{\omega \in \Omega} P(\omega) = 1$

For any proposition ϕ, $P(\phi) = \sum_{\omega \in \phi} P(\omega)$

$\Rightarrow P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Kolmogoroff axioms

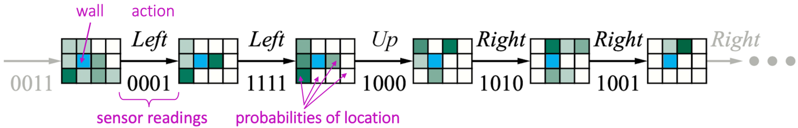
Basic axioms of probability



de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

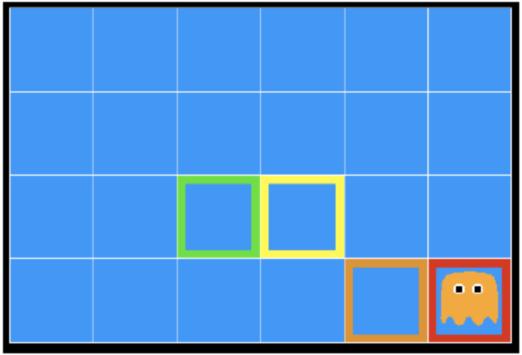
Evidence and query variables

- Observed (**evidence**): Agent knows something about the state of the world
  (e.g., sensor readings or symptoms)
- Unobserved variables: Need to reason about other aspects
  - **query** variables: What the agent is interested in
    (e.g. where an object is or what disease is present)
  - **hidden** variables: Other relevant variables in the problem description
    (may be useful to answer query)
- Model: Agent knows something about
  - how the known variables relate to the unknown variables
    likelihood model, e.g., sensor model
  - the unknown variables (for a particular type of problem)
    prior model, e.g., transition model



Evidence, query and model: Example

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - on the ghost: red
  - 1 or 2 away: orange
  - 3 or 4 away: yellow
  - 5+ away: green
- Sensor readings are noisy!



- We know the sensor model: $P(\underbrace{Color}_{evidence} \mid \underbrace{Distance}_{query})$

$P(red \mid 3)$	$P(orange \mid 3)$	$P(yellow \mid 3)$	$P(green \mid 3)$
0.05	0.15	0.5	0.3

Adapted from D. Klein & P. Abbeel: Artificial Intelligence (UC Berkeley)

Random variables, events and propositions in AI - practically

- Random variable: a variable whose value is affected by some random phenomenon
  E.g.,
  - D = How long will it take to drive to airport?
  - A = Are there reported accidents?
  - R = Is it raining?
  - L = In which square is the ghost?
- Categorization
  - discrete
    - **Boolean** (propositional): take only two values $\{true, false\}$, i.e., $\{1, 0\}$ (E.g., A, R)
    - General discrete – countable number of distinct values (E.g., L)
  - continuous (E.g., D)
- Assignment of a realization to a random variable is an event.
  - E.g. $R = true$. What is the probability of the event "it rains"?
- In AI, event = proposition
  For any proposition ϕ, $P(\phi) = \sum_{\omega \in \phi} P(\omega)$

Probability distributions

Temperature:

T	P
hot	0.5
cold	0.5

Weather:

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Illustration credit: D. Klein & P. Abbeel: Artificial Intelligence (UC Berkeley)

Probability distributions

Unobserved random variables have distributions

- A distribution: table of probabilities of values

$\mathbf{P}(T)$

T	P
hot	0.5
$cold$	0.5

$\mathbf{P}(W)$

W	P
sun	0.6
$rain$	0.1
fog	0.3
$meteor$	0

- A probability is a single number
  $P(W = rain) = 0.1$

- It holds:
  $\forall x \quad P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Shorthand notation:

$$P(hot) = P(T = hot)$$

$$P(cold) = P(T = cold)$$

$$P(rain) = P(W = rain)$$

$$\dots$$

OK if domain entries unique

Joint distributions

- A joint distribution over a set of random variables $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (outcome):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

or shorter: $P(x_1, x_2, \ldots x_n)$

- Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$\mathbf{P}(T, W)$

T	W	P
hot	*sun*	0.4
hot	*rain*	0.1
cold	*sun*	0.2
cold	*rain*	0.3

- Size of distribution if n variables with domain sizes d?
  - For all but the smallest distributions, impractical to write out!

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables
- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called outcomes or realizations
  - joint distributions: say whether outcomes are likely
  - Normalized: sum to 1.0
  - Ideally: only certain variables directly interact

Distribution over T, W

T	W	P
hot	sun	0.4
hot	$rain$	0.1
$cold$	sun	0.2
$cold$	$rain$	0.3

Events

- An event is a set E of outcomes

$$P(E) = \sum_{(x_1,x_2,\dots x_n)\in E} P(x_1, x_2, \dots x_n)$$

- From a joint distribution, we can calculate the probability of any event connected to (some or all of) the involved variables, e.g.,
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?

$\mathbf{P}(T, W)$

T	W	P
hot	sun	0.4
hot	$rain$	0.1
$cold$	sun	0.2
$cold$	$rain$	0.3

- Typically, the events we care about are partial assignments, like $P(T = hot)$, i.e., probabilistic assertions are usually not about particular atomic events but about sets of them.

Quiz: Events

- $P(x, y) = \; \ldots\ldots$
- $P(x) = \ldots\ldots$
- $P(x \vee \neg y) = \ldots\ldots$

$\mathbf{P}(X, Y)$

X	Y	P
x	y	0.2
x	$\neg y$	0.3
$\neg x$	y	0.4
$\neg x$	$\neg y$	0.1

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$\mathbf{P}(T, W)$

T	W	P
hot	sun	0.4
hot	$rain$	0.1
$cold$	sun	0.2
$cold$	$rain$	0.3

$$P(t) = \sum_w P(t, w)$$

$\mathbf{P}(T)$

T	P
hot	0.5
$cold$	0.5

$$P(w) = \sum_t P(t, w)$$

$\mathbf{P}(W)$

W	P
sun	0.6
$rain$	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Quiz: Marginal Distributions

$\mathbf{P}(X, Y)$

X	Y	P
x	y	0.2
x	$\neg y$	0.3
$\neg x$	y	0.4
$\neg x$	$\neg y$	0.1

$$P(x) = \sum_y P(x, y)$$

$\mathbf{P}(X)$

X	P
x	...
$\neg x$	...

$$P(y) = \sum_x P(x, y)$$

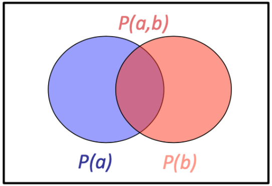
$\mathbf{P}(Y)$

Y	P
y	...
$\neg y$	...

Conditional probabilities

Definition (Conditional probability)

$$P(a|b) = \frac{P(a,b)}{P(b)} \text{ if } P(b) \neq 0$$



Product rule gives an alternative formulation:

$$P(a,b) = P(a|b)P(b) = P(b|a)P(a)$$

A general version holds for whole distributions, e.g.,

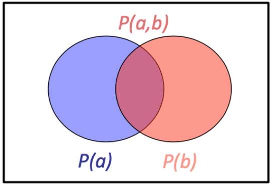
$$\mathbf{P}(T,W) = \mathbf{P}(T|W)\mathbf{P}(W)$$

(View as a 2×2 set of equations, **not** matrix multiplication)

Conditional probabilities

Definition (Conditional probability)

$$P(a|b) = \frac{P(a,b)}{P(b)} \text{ if } P(b) \neq 0$$



$\mathbf{P}(T, W)$

T	W	P
hot	*sun*	0.4
hot	*rain*	0.1
cold	*sun*	0.2
cold	*rain*	0.3

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W = s, T = c) + P(W = r, T = c)$$

$$= 0.2 + 0.3 = 0.5$$

Quiz: Conditional probabilities

$\mathbf{P}(X, Y)$

X	Y	P
x	y	0.2
x	$\neg y$	0.3
$\neg x$	y	0.4
$\neg x$	$\neg y$	0.1

- $P(x \mid y) = \;\ldots\ldots$
- $P(\neg x \mid y) = \ldots\ldots$
- $P(\neg y \mid x) = \ldots\ldots$

Conditional distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Joint distribution

$\mathbf{P}(T, W)$

T	W	P
hot	sun	0.4
hot	$rain$	0.1
$cold$	sun	0.2
$cold$	$rain$	0.3

Conditional distributions

$\mathbf{P}(W|T = hot)$

W	P
sun	0.8
$rain$	0.2

$\mathbf{P}(W|T = cold)$

W	P
sun	0.4
$rain$	0.6

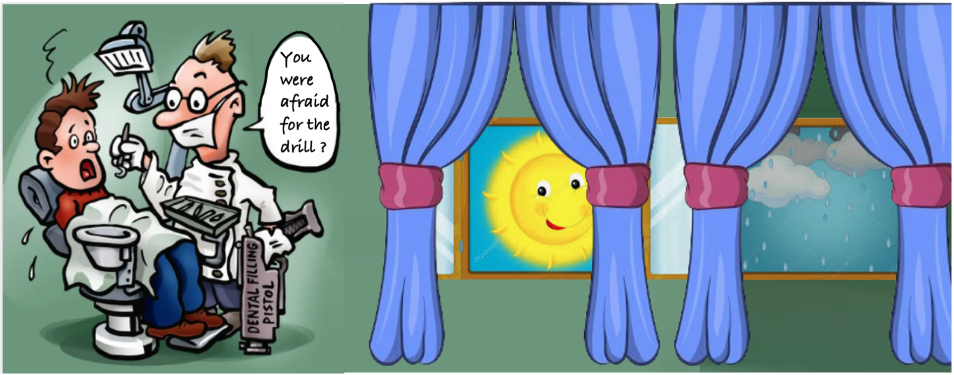
$\mathbf{P}(W|T)$

Chain rule

Chain rule is derived by successive application of product rule:

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

Dentist use case



What can a dentist conclude when the steel probe catches in the aching tooth?

Model this by 3 Boolean r.v.s: $Catch$, $Cavity$, $Toothache$

Perceive also $Weather$ as a discrete r.v. (sunny, rainy, cloudy or snow)

Syntax for propositions

Propositional or Boolean random variables

e.g., $Cavity$ (do I have a cavity?)

$Cavity = true$ is a proposition, also written $cavity$

Discrete random variables (finite or infinite)

e.g., $Weather$ is one of $\langle sunny, rain, cloudy, snow \rangle$

$Weather = rain$ is a proposition

Values must be exhaustive and mutually exclusive

Continuous random variables (bounded or unbounded)

e.g., $WaitingTime = 383.4$; also allow, e.g., $WaitingTime < 60.0$.

Arbitrary Boolean combinations of basic propositions

Prior probability

Prior or unconditional probabilities of propositions

e.g., $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$

correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)

$\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

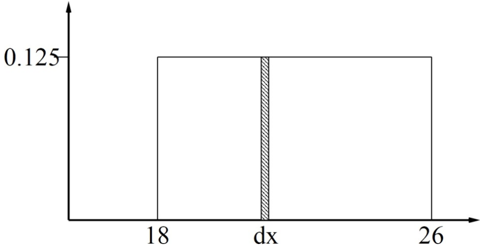
$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

Note: every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Probability for continuous variables

For continuous variables, we define the probability that a random variable takes some value x as a parametrized function of x, e.g.,

$P(X=x) = U[18,26](x) =$ uniform density between 18 and 26



Here P is a probability density function (pdf) or just density; it integrates to 1.

$P(X=20.5) = 0.125$ really means

$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

Conditional probabilities exemplified for the Dentist case

Conditional probabilities express belief given some evidence

e.g., $P(cavity|toothache) = 0.8$

means 80% chance of $cavity$ **given that** $toothache$ **is all I know**

NOT "if $toothache$ then 80% chance of $cavity$"

If we know more, e.g., that there is no gum disease, we might get

$P(cavity|toothache, \neg gumdisease) = 0.93$

Note: the less specific belief **remains valid** after more evidence arrives, but is not always **useful**

New evidence may be irrelevant, allowing simplification, e.g.,

$P(cavity|toothache, kaagentwins) = P(cavity|toothache) = 0.8$

This kind of inference, sanctioned by domain knowledge, is crucial

Notation for conditional distributions:

$\mathbf{P}(Cavity|Toothache) =$ 2-element vector of 2-element vectors.

We call it conditional probability table (CPT)

Inference by enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where the proposition is true:

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where the proposition is true:

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where the proposition is true:

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where the proposition is true:

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Can also compute conditional probabilities:

$$\begin{aligned} P(\neg cavity | toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 \end{aligned}$$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Denominator can be viewed as a normalization constant α

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

$$= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$

$$= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]$$

$$= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$$

General idea: compute distribution on query variable
by fixing evidence variables and summing over hidden variables

Inference by enumeration: Summary

Let $\mathbf{X}$ denote all the variables in a given problem formulation. Typically, we want:
the posterior joint distribution of the query variables $\mathbf{Y}$
given specific values $\mathbf{e}$ for the evidence variables $\mathbf{E}$

Let the hidden variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$ (neither query nor evidence vars)

Then we obtain the desired posterior by summing out the hidden variables:

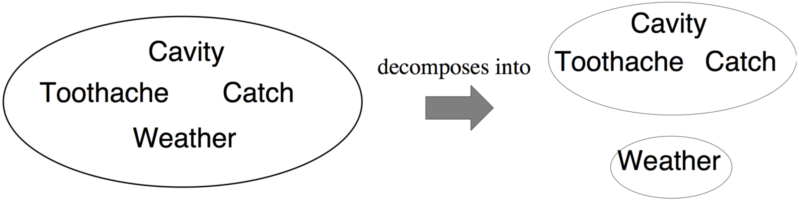
$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E}=\mathbf{e}) = \alpha \sum_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E}=\mathbf{e}, \mathbf{H}=\mathbf{h})$$

Obvious problems:
1) Worst-case time complexity $O(d^n)$ where d is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Independence

A and B are independent iff

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



$$\mathbf{P}(Toothache, Catch, Cavity, Weather) \\ = \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

- 32 entries reduced to 12
- for n independent biased coins $\mathbf{P}(C_1, ..., C_n) = \prod_i P(C_i)$, so reduction $2^n \to n$
- Absolute independence powerful but rare in practice. What to do?

Conditional independence

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is conditionally independent of $Toothache$ given $Cavity$:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Equivalent statements:

$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$

$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$

Conditional independence contd.

Write out full joint distribution using chain rule:

$$\begin{aligned}
&\mathbf{P}(Toothache, Catch, Cavity) \\
&\quad = \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity) \\
&\quad = \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity) \\
&\quad = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

I.e., reduced from $2^3 - 1 = 7$ to 2 + 2 + 1 = 5 independent numbers

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

The decomposition of large probabilistic domains into weakly connected subsets through conditional independence is crucial in AI.

Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\implies \text{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

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

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis, S be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

Bayes' Rule and conditional independence

Remember the dentist problem:

$$\begin{aligned}
&\mathbf{P}(Cavity|toothache \wedge catch) \\
&= \alpha\,\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity) \\
&= \alpha\,\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

This is an example of a naïve Bayes model:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i|Cause)$$

Total number of parameters is **linear** in n.

Example of probabilistic inference: Wumpus world

Each square other than [1,1] contains a pit with probability 0.2

Pits cause breezes in neighbouring squares; $\mathbf{B}$: breeze felt; $\mathbf{OK}$: safe location

$P_{ij} = true$ iff $[i, j]$ contains a pit. The agent dies when entering a square with a pit.

$B_{ij} = true$ iff $[i, j]$ is breezy

Goal: infer where is it safest to move on outside of the explored $\mathbf{OK}$ locations.

Specifying the probability model

The full joint distribution is $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$

Apply product rule: $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \,|\, P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$
(Do it this way to get $P(Effect|Cause)$.)

First term: 1 if pits are adjacent to breezes, 0 otherwise
Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Observations and query

We know the following facts:

$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$

$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

Query is $\mathbf{P}(P_{1,3}|known, b)$

Define $Unknown = P_{ij}$s other than $P_{1,3}$ and $Known$

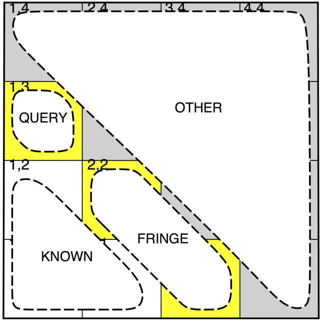
For inference by enumeration, we have

$$\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Grows exponentially with number of squares!

Using conditional independence

Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares



Define $Unknown = Fringe \cup Other$

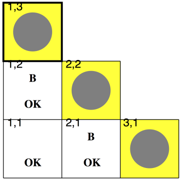
$\mathbf{P}(b|P_{1,3}, Known, Unknown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$

Manipulate query into a form where we can use this!

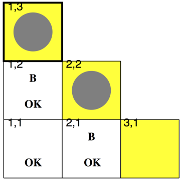
Using conditional independence contd.

$$\begin{aligned}
\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other) \\
&= \alpha \, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other)
\end{aligned}$$

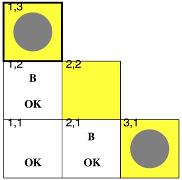
Using conditional independence contd.



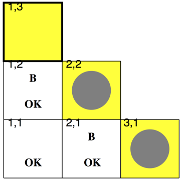
0.2 x 0.2 = 0.04



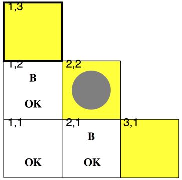
0.2 x 0.8 = 0.16



0.8 x 0.2 = 0.16



0.2 x 0.2 = 0.04



0.2 x 0.8 = 0.16

$$
\begin{aligned}
\mathbf{P}(P_{1,3}|known,b) &= \alpha' \, \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) P(fringe) \\
&= \alpha' \, \langle 0.2(0.04+0.16+0.16), \;\; 0.8(0.04+0.16) \rangle \\
&\approx \langle 0.31, 0.69 \rangle
\end{aligned}
$$

$$
\mathbf{P}(P_{2,2}|known,b) \approx \langle 0.86, 0.14 \rangle \quad \text{(derived equivalently)}
$$

Obviously, the agent should avoid [2,2].

Summary

- Probability is a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
- Independence and conditional independence provide the tools
- **Next time**: Bayesian networks