Lecture 2: Probabilistic reasoning

Frederik Mallmann-Trenn

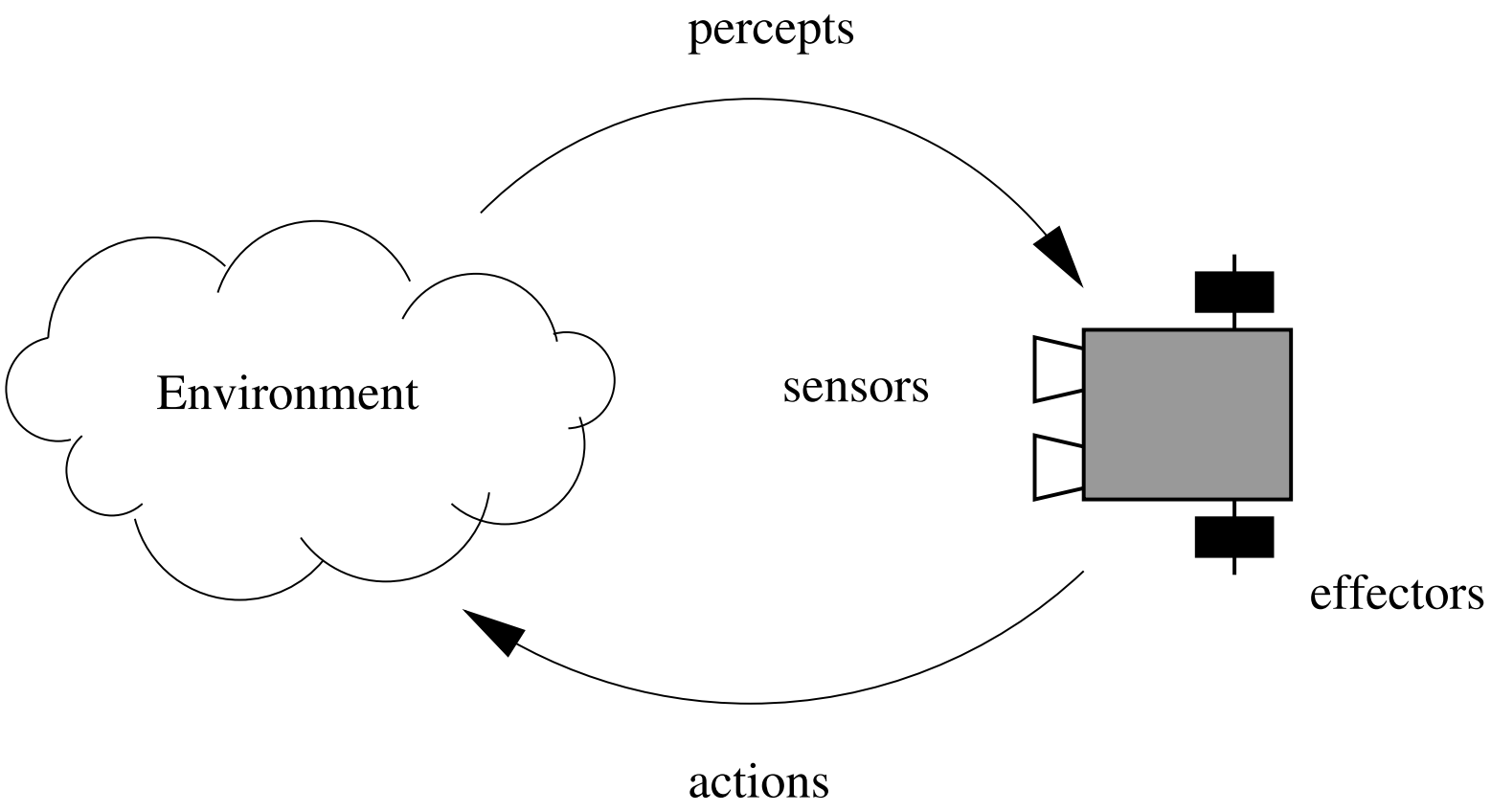
Department of Informatics
King's College London

(Version 1.3)

Today

- Introduction
- Probabilistic Reasoning I
- Probabilistic Reasoning II
- Sequential Decision Making
- Temporal Probabilistic Reasoning
- Game Theory
- Argumentation I
- Argumentation II
- (A peek at) Machine Learning
- AI & Ethics

Recap

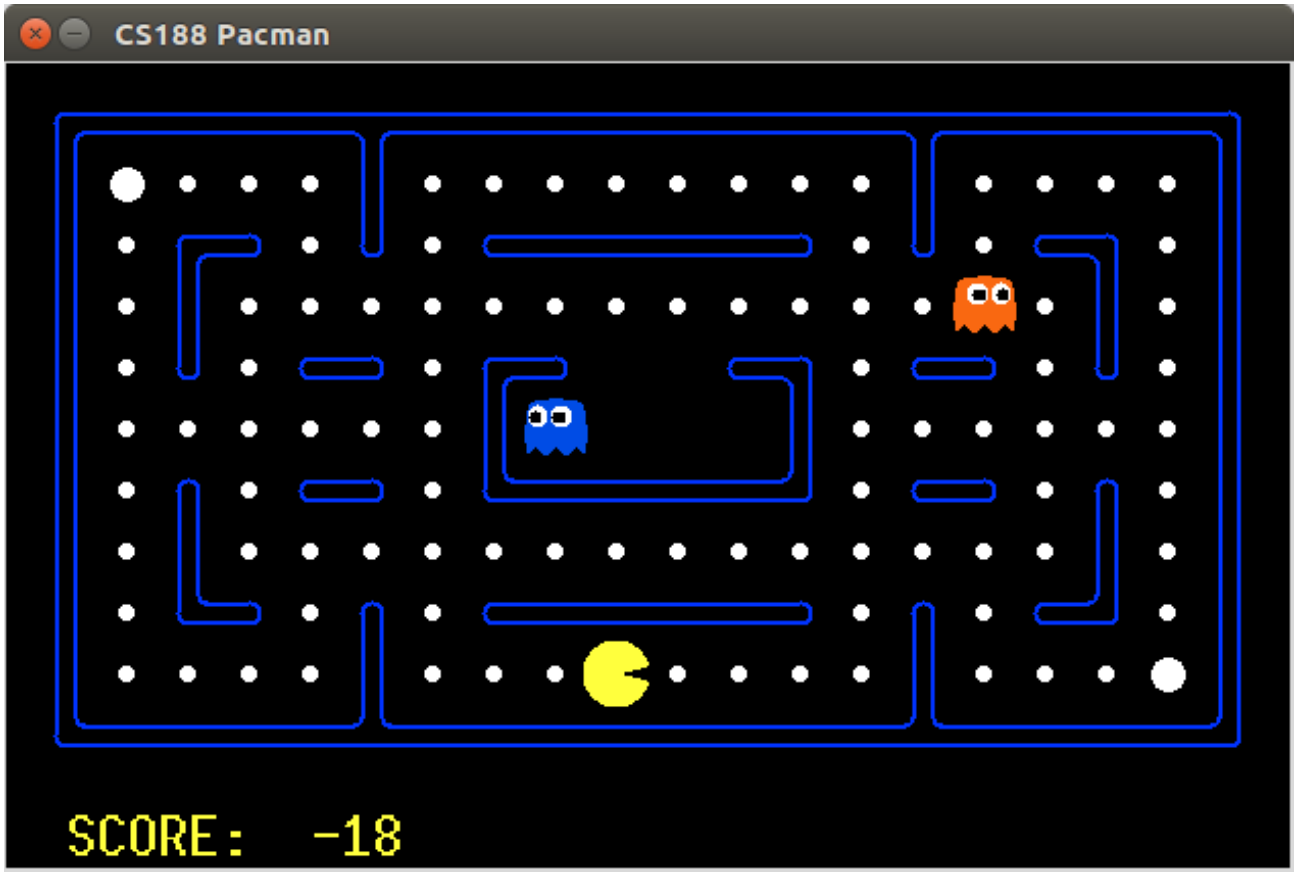


- AI Systems operate in environments

Recap

- Environments are:
  - Acessible or not.
  - Observable/ Fully observable / partially observable.
  - Deterministic / stochastic / non-deterministic
  - Static or Dynamic
  - Sequential or Episodic
  - Discrete or continuous

What about Pacman?



(ai.berkeley.edu)

- `https://pollev.com/frederikm106`

What about Pacman?

- It will get harder.
- Partial observability.
- Non-deterministic.

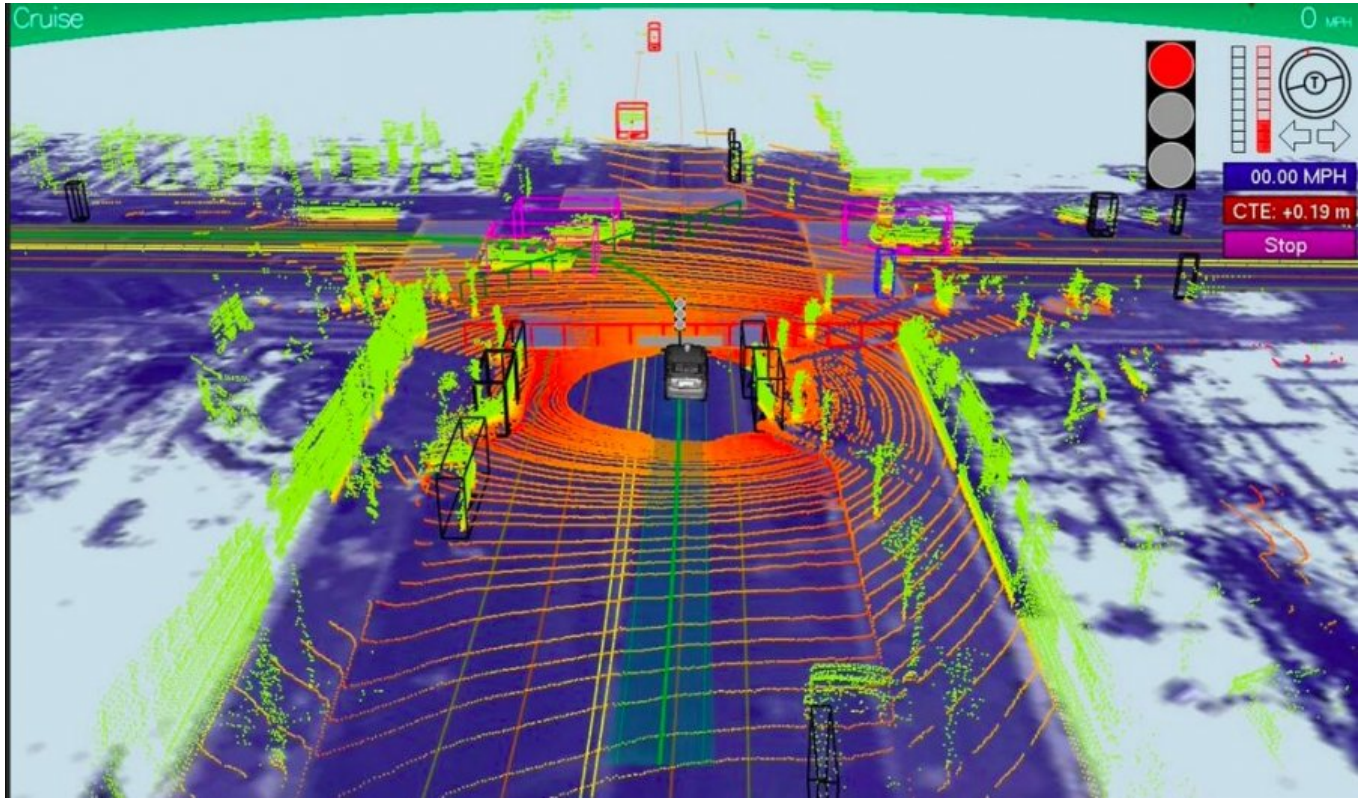
Partially observable Pacman

(mystorybook.com/books/42485)

Partial observability

- Full observability not true of the real world:
  - Is is raining on Strand?
  - Who is texting right now?
- Partial observability can arise for many reasons.
  - World structure *vs*. sensor ability.
  - Sensor noise.
  - Computational complexity.
- Partial observability → uncertainty

Partially observable world



(Sebastian Thrun & Chris Urmson/Google)

Uncertainty

- Say you have to pickup your family from the airport.
- When do you leave?
- Let action A_t = leave for airport t minutes before flight
- Will A_t get me there on time?
- Problems:
  1. partial observability (road state, other drivers' plans, etc.)
  2. noisy sensors (Google Maps' traffic updates)
  3. uncertainty in action outcomes (flat tire, etc.)
  4. immense complexity of modelling and predicting traffic

Uncertainty

- Making statements based on partial observations is tricky
  1. risks falsehood: "A_{90} will get me there on time"
  2. leads to conclusions that are too weak for decision making: "A_{90} will get me there on time if there's no accident on the Hammersmith flyover, and it doesn't rain and my tires remain intact etc etc."

(getwestlondon.co.uk)

(Getty Images)

Uncertainty

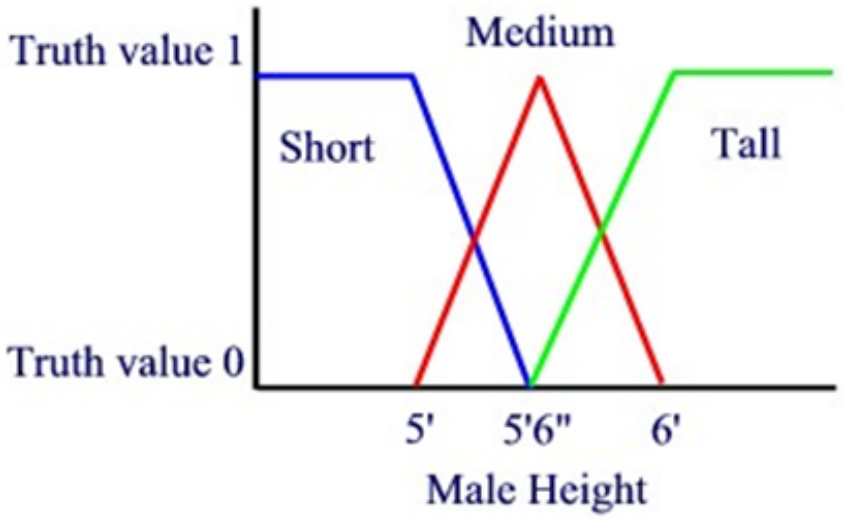
- Play safe: A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport . . .
- Can we do something better?

Methods for handling uncertainty

- Use probability
  - Given the available evidence,
    A_{25} will get me there on time with probability 0.04
- Probabilistic assertions summarize effects of
  - laziness: failure to enumerate exceptions, qualifications, etc.
  - ignorance: lack of relevant facts, initial conditions, etc.
- Even with probabilistic assumptions there are still issues: Computational complexity, obtaining values, semantics.
  - We will consider the computational issues in some detail.

Aside

- Fuzzy logic handles degree of truth NOT uncertainty
  - *WetGrass* is true to degree 0.2



(docs.docandys.com)

Probability

- Subjective or Bayesian probability:
  - Probabilities relate propositions to one's own state of knowledge
  - $P(A_{25}) = 0.04$. Knowing that no accident was reported, makes it more likely

$$P(A_{25}|\text{no reported accidents}) = 0.06$$

- Probabilities of propositions change with new evidence:

$$P(A_{25}|\text{no reported accidents, leaving at 5 a.m.}) = 0.15$$

$$P(A_{25}|\text{no reported accidents, leaving at 9 a.m.}) = 0.03$$

Making decisions under uncertainty

- Suppose I believe the following:

$$
\begin{aligned}
P(A_{25} \text{ gets me there on time} | \ldots) &= 0.04 \\
P(A_{90} \text{ gets me there on time} | \ldots) &= 0.70 \\
P(A_{120} \text{ gets me there on time} | \ldots) &= 0.95 \\
P(A_{1440} \text{ gets me there on time} | \ldots) &= 0.9999
\end{aligned}
$$

Which action to choose?

Methods for handling uncertainty

https://pollev.com/frederikm106

Methods for handling uncertainty

- Depends on our preferences for missing flight vs. airport cuisine, sleeping on a bench, and so on.
- Utility theory is used to represent and infer preferences
- Decision theory = utility theory + probability theory
- We will come back to decision theory in two weeks.

Probability basics

- Begin with a set Ω—the sample space.
- This is all the possible things that could happen.
  - 6 possible rolls of a die.
- $\omega \in \Omega$ is a sample point, possible world, atomic event.

(pngimg.com)

Probability basics

- A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ such that:

$$0 \leqslant P(\omega) \leqslant 1$$
$$\sum_{\omega} P(\omega) = 1$$

- For a typical die:

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

(pngimg.com)

(hartsport.com.au)

Probability basics

- An event A is any subset of Ω

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

- Again, for a regular die:

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

(pngimg.com)

(hartsport.com.au)

Random variables

- A random variable is a function from sample points to some range.
  - $raining(London) = true$.
  - $temperature(lecture\ room) = 27$
- P induces a probability distribution for any r.v. X:

$$P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

- Continuing our example:

$$\begin{aligned} P(Odd = true) &= P(1) + P(3) + P(5) \\ &= 1/6 + 1/6 + 1/6 \\ &= 1/2 \end{aligned}$$

Now it is your turn

- A pack of 52 playing cards.
- Four suites: Hearts, Diamonds, Clubs, Spades.
- Each suite contains:
  A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K.
- So the sample points are:
  Ace of Hearts, 2 of Hearts, . . . King of Spades
- What are the probabilities of. . .

Now it is your turn

Six of diamonds

https://pollev.com/frederikm106

Last year's success rate 0.89

A six

https://pollev.com/frederikm106

Last year's success rate 0.93

Diamond or club

https://pollev.com/frederikm106

Last year's success rate 0.87

Say you have a die

Your friend rolled a die and tells you the outcome is an even number. What's the probability that the die's outcome is a six?
`https://pollev.com/frederikm106`

Propositions

- We describe the world in terms of propositions
- Think of a proposition as the event (set of sample points) where the proposition is true
- Given Boolean random variables A and B:

  event a = set of sample points where $A(\omega) = \mathit{true}$
  event $\neg a$ = set of sample points where $A(\omega) = \mathit{false}$
  event $a \wedge b$ = points where $A(\omega) = \mathit{true}$ and $B(\omega) = \mathit{true}$

- A state can be defined by a set of Boolean variables.

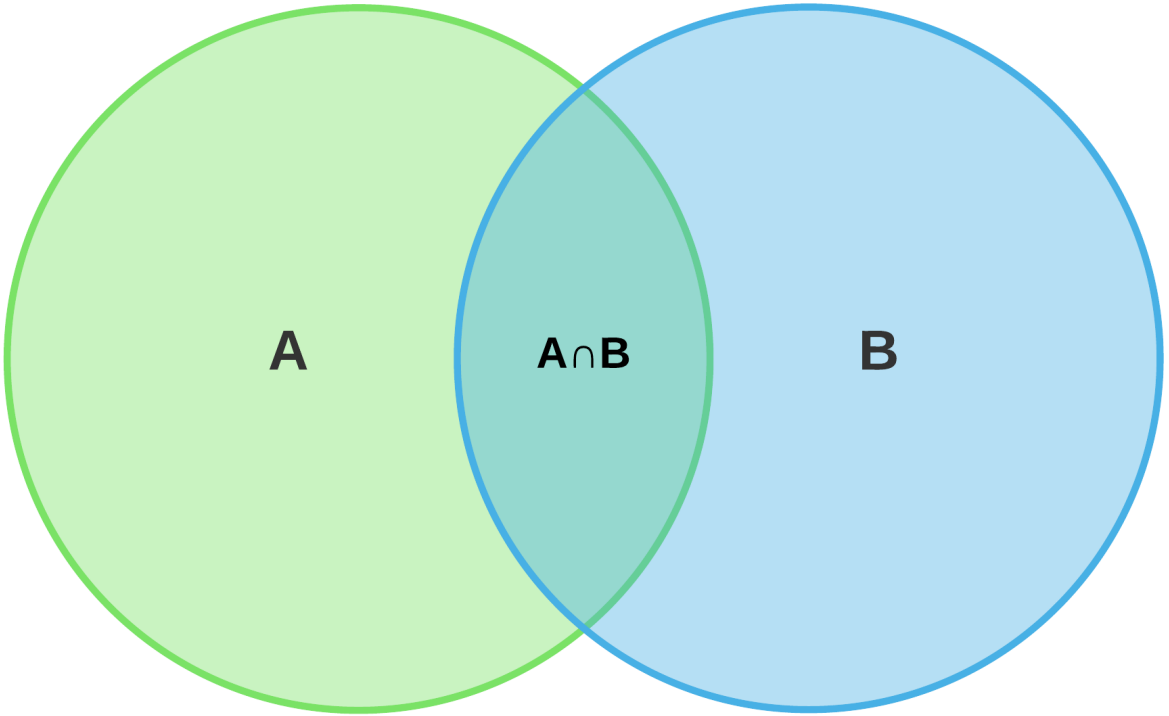
$$a \wedge b \wedge \neg c \qquad A = \mathit{true}, B = \mathit{true}, C = \mathit{false}$$

This is then just a sample point.

More complex expressions

- The definitions imply that certain logically related events must have related probabilities

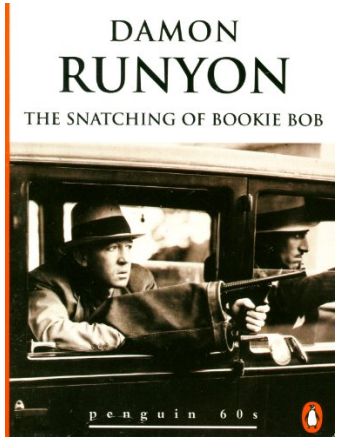
$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$



(lucidchart.com)

Dutch book argument

- We are going to take the subjective/Bayesian view (how likely do we think is something; we don't actually know the exact probabilities)
- Why is this a good way of modelling the world like that? Dutch book argument!
- 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.
- So we always have an incentive to make precise predictions
- Not everyone agrees with the argument



(Penguin)

(sfsu.edu)

Why Dutch?

(Henri-Pierre Danloux) (Thomas Whitcombe)

Syntax for propositions

- Propositional or Boolean random variables
  - *Cavity* (do I have a cavity?)
  - $Cavity = true$ is a proposition, also written *cavity*
  - $Cavity = false$ is a proposition, also written $\neg cavity$ or $\overline{cavity}$
- Discrete random variables (finite or infinite)
  - *Weather* is one of $\langle sunny, rain, cloudy, snow \rangle$
  - $Weather = rain$ is a proposition
- Continuous random variables (bounded or unbounded)
  - $Temp = 21.6$; also allow, e.g., $Temp < 22.0$.

Syntax for propositions

- We allow arbitrary combination of logical operators (AND, OR, NOT) and comparison operators $(<, \leqslant, =, \neq, \dots)$.
- E.g., $\mathit{Temp} < 22.0$ AND $\mathit{Cavity} = \mathit{true}$
- Sometimes we will write $\wedge$, $\vee$ and $\neg$ for (AND, OR, NOT).

Prior probability

- Prior or unconditional probabilities of propositions

  $$P(Cavity = true) = 0.1 \text{ and } P(Weather = sunny) = 0.72$$

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

Prior probability

- Probability distribution gives values for all possible assignments (assumes a fixed ordering):

$$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle \tag{1}$$

  means
  - $P(Weather = sunny) = 0.72$,
  - $P(Weather = rain) = 0.1$,
  - $P(Weather = cloudy) = 0.08$ and
  - $P(Weather = snow) = 0.1$
- Values must be exhaustive (everything covered) and mutually exclusive (no overlap)
- Values have to sums to 1

Joint probability distribution

- 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 2×4 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

- Every question about a domain can be answered by the joint distribution because every event is a sum of sample points.
- E.g., $P(Cavity = true \ AND \ Weather \neq cloudy)$

Inference by enumeration

(Paramount Pictures)

Inference by enumeration

- Example of a joint distribution with three variables (catch is something the doctor can test):

	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 it is true:

$$P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega)$$

Inference by enumeration

	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 it is true:

$$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$$

$$\begin{aligned} P(toothache) &= 0.108 + 0.012 + 0.016 + 0.064 \\ &= 0.2 \end{aligned}$$

Inference by enumeration

	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 it is true:

$$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$$

$$\begin{aligned} P(cavity \vee toothache) &= 0.108 + 0.012 + 0.072 \\ &\quad + 0.008 + 0.016 + 0.064 \\ &= 0.28 \end{aligned}$$

Your turn

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

- Calculate $Pr(toothache \wedge cavity)$.
- `https://pollev.com/frederikm106`

Your turn

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

- Calculate $Pr(cavity \mid toothache)$.
- `https://pollev.com/frederikm106`

Conditional probability

- Conditional or posterior probabilities

  $$P(cavity|toothache) = 0.6$$

  given that *toothache* is all I know
- Recall, for convenience we write *cavity* for $Cavity = true$ and *toothache* for $Toothache = true$.
- Notation for conditional distributions:

  $$\mathbf{P}(Cavity|Toothache)$$

  The outcome is a 2-dimensional vector (and so is Cavity and Toothache)

Conditional probability

- If we know more, e.g., *cavity* is also given, then we have

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

  Note: the less specific belief (toothache) remains valid after more evidence arrives, but is not always useful
- New evidence may be irrelevant, allowing simplification

  $$P(cavity|toothache, soxWin) = P(cavity|toothache) = 0.6$$
- This kind of inference, sanctioned by domain knowledge, is crucial.

Conditional probability

- Definition of conditional probability:

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

- Product rule gives an alternative formulation:

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

- A general version holds for joint distributions,

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

(View as a 4×2 set of equations, not matrix multiplication)

Chain rule

- Chain rule is derived by successive application of product rule. Concrete example:

$$\begin{aligned} P(a,b,c) &= P(a,b)P(c|b,a) \\ &= P(a)P(b|a)P(c|b,a) \end{aligned}$$

- In general,

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

Inference by enumeration

- Can also compute conditional probabilities:

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

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

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

Normalization

$$
\begin{aligned}
\mathbf{P}(Cavity|toothache) &= \alpha\,\mathbf{P}(Cavity, toothache) \\
&= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
&= \alpha\,\langle 0.12, 0.08\rangle \\
&= \langle 0.6, 0.4\rangle
\end{aligned}
$$

Denominator can be viewed as a normalization constant α

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

- Green boxes show step in calculation, not the desired outcome

General version

- Let $\mathbf{X}$ be all the variables.
- 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 $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$
- In our example, $\mathbf{H} = \{Catch\}$, $\mathbf{E} = \{Toothache\}$, $\mathbf{Y} = \{Cavity\}$.

General version

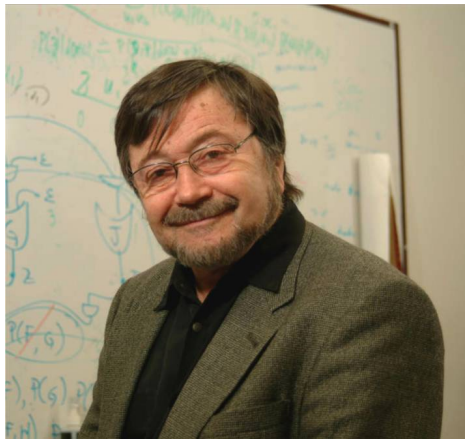
- Then the required summation of joint entries is done by summing out the hidden variables:

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

- The terms in the summation are joint entries because **Y**, **E**, and **H** together exhaust the set of random variables

Inference by enumeration

- 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???
- These problems effectively stopped the use of probability in AI until the mid 80s



Judea Pearl

Daphne Koller

Computational efficiency

- To get efficient probabilistic computations, we need two things.
  1. (Conditional) independence.
  2. Bayes rule.
- Will cover these now, in order.

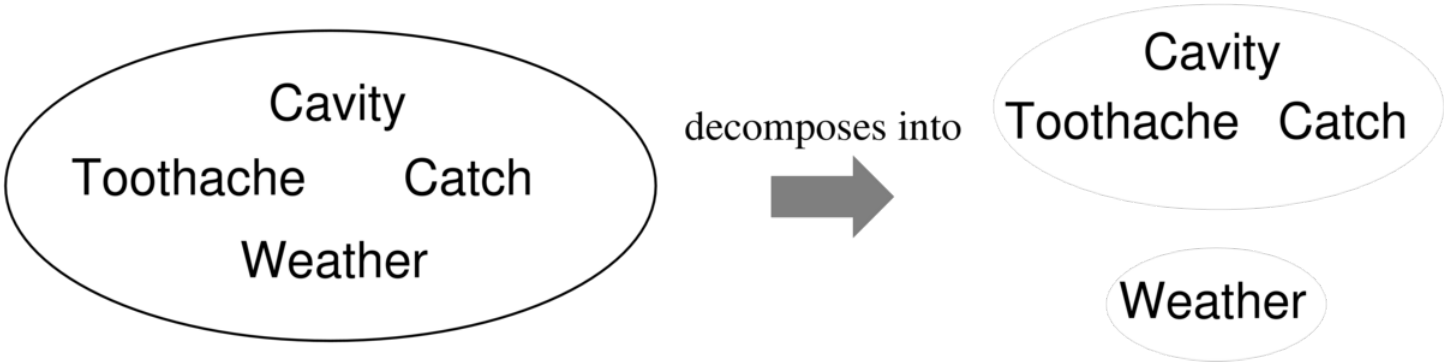
Independence

- A and B are independent iff

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

- Why is this interesting?
- Can help with the size of the problem.

Independence



- $\mathbf{P}(Toothache, Catch, Cavity, Weather)$
  $= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$
- 31 values reduced to 10.
- For n independent biased coins, $2^n \rightarrow n$

Independence

- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent.
  What to do?
- Conditional independence

Conditional independence

- $\mathbf{P}(Toothache, Cavity, Catch)$ has:
- Three binary variables.
- Thus 2^3 entries in the joint probability table.
- But these sum to 1.
- So $2^3 - 1$ independent entries
- That's 7 independent entries

Conditional independence

- But, wait! There's more!
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

$$P(catch|toothache, cavity) = P(catch|cavity) \tag{2}$$

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

$$P(catch|toothache, \neg cavity) = P(catch|\neg cavity) \tag{3}$$

- *Catch* is conditionally independent of *Toothache* given *Cavity*

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

Conditional independence

- Equivalent statements:

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

Conditional independence

- Write out full joint distribution using chain rule:

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

- 2 + 2 + 1 = 5 independent numbers
- Equations 1 and 2 remove 2.

Conditional independence

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to (close to) linear in n.
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.
- Can often make conditional independence statements when little else is known.

Greg Provan

Bayes' Rule

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

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

or in distribution form

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

(John Bayes)

Bayes' Rule

- Probability that some hypothesis h is true, given that some event e has occured.

$$\mathbf{P}(h|e) = \frac{\mathbf{P}(e|h)\mathbf{P}(h)}{\mathbf{P}(e)}$$

- Now, since:

$$P(e) = P(e|h)P(h) + P(e|\neg h)P(\neg h)$$

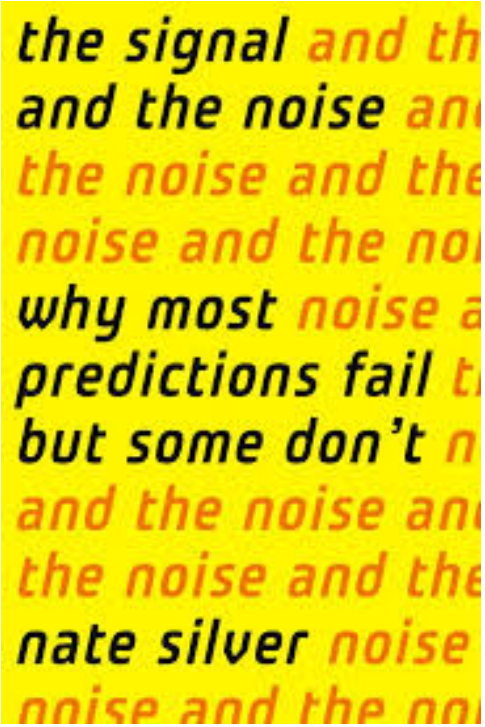
we can rewrite this in a form in which it is commonly applied:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)(1 - P(h))}$$

where variables are binary valued.

Some examples

- From:



Nate Silver

The wrong underwear

You live with your partner.
You come home from a business trip to discover a strange pair of underwear in your drawer.
What should you think?

In popular culture

(Metro-Goldwyn-Mayer)

- Major plot device in the 2003 romantic comedy.

The wrong underwear

- What do you think?

What do you think?

https://pollev.com/frederikm106

The wrong underwear

- Probability that some hypothesis is true, given that some event has occured.

$$P(h|e) = \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)(1 - P(h))}$$

- event = underwear
- hypothesis = partner cheating
- To apply Bayes rule we need:
  - $P(e|h)$
  - $P(e|\neg h)$
  - $P(h)$

The wrong underwear

- $P(e|h)$
- If they cheated, how likely is the underwear?
- Well, if they cheated, then maybe it is likely the underwear would appear.
- But also, wouldn't they be more careful?
- Say 50% chance.

The wrong underwear

- $P(e|\neg h)$
- Is there an innocent explanation?
- A friend stayed over? Something was left in the dryer?
- Say 5% chance

The wrong underwear

- $P(h)$?
- Prior probability they cheated — before there was any evidence.
- Hard to quantify.
- Studies suggest about 4% of married partners cheat in a given year.
- Say 4%.

The wrong underwear

- The verdict?

$$
\begin{aligned}
P(h|e) &= \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)(1 - P(h))} \\
&= \frac{0.5 \times 0.04}{0.5 \times 0.04 + 0.05 \times (1 - 0.04)} \\
&= 0.294
\end{aligned}
$$

- Low because of the low prior.

The wrong underwear

- Now it happens again.

The wrong underwear

- What do you think?

What do you think?

https://pollev.com/frederikm106

The wrong underwear

- Well now the prior is 0.294

$$
\begin{aligned}
P(h|e) &= \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)(1 - P(h))} \\
&= \frac{0.5 \times 0.294}{0.5 \times 0.294 + 0.05 \times (1 - 0.294)} \\
&= 0.806
\end{aligned}
$$

- and the verdict is pretty clear.

A medical example

- Let M be meningitis, S be stiff neck:

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

- Posterior probability of meningitis still very small, again because of low prior.
- Common pattern in medical test results.
- Note that here we used a slightly different formulation of Bayes Rule.

Bayes rule & knowledge representation

- Useful for assessing diagnostic probability from causal probability:

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

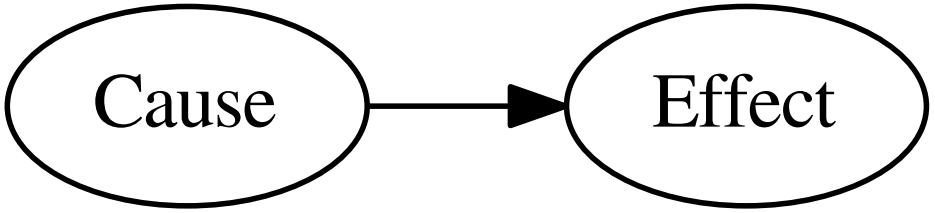
- We saw this in both examples.

Bayes rule & knowledge representation

- Often easier to assess causal probabilities.

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

- Can visualise this as:

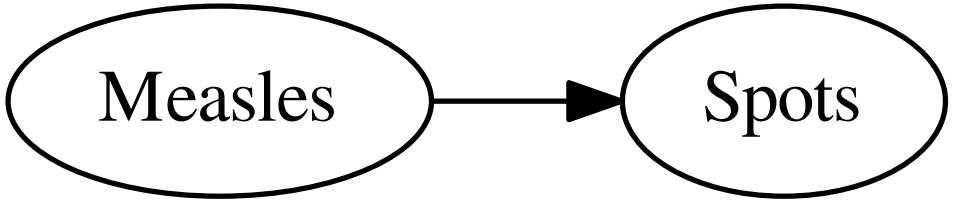


Bayes rule & knowledge representation

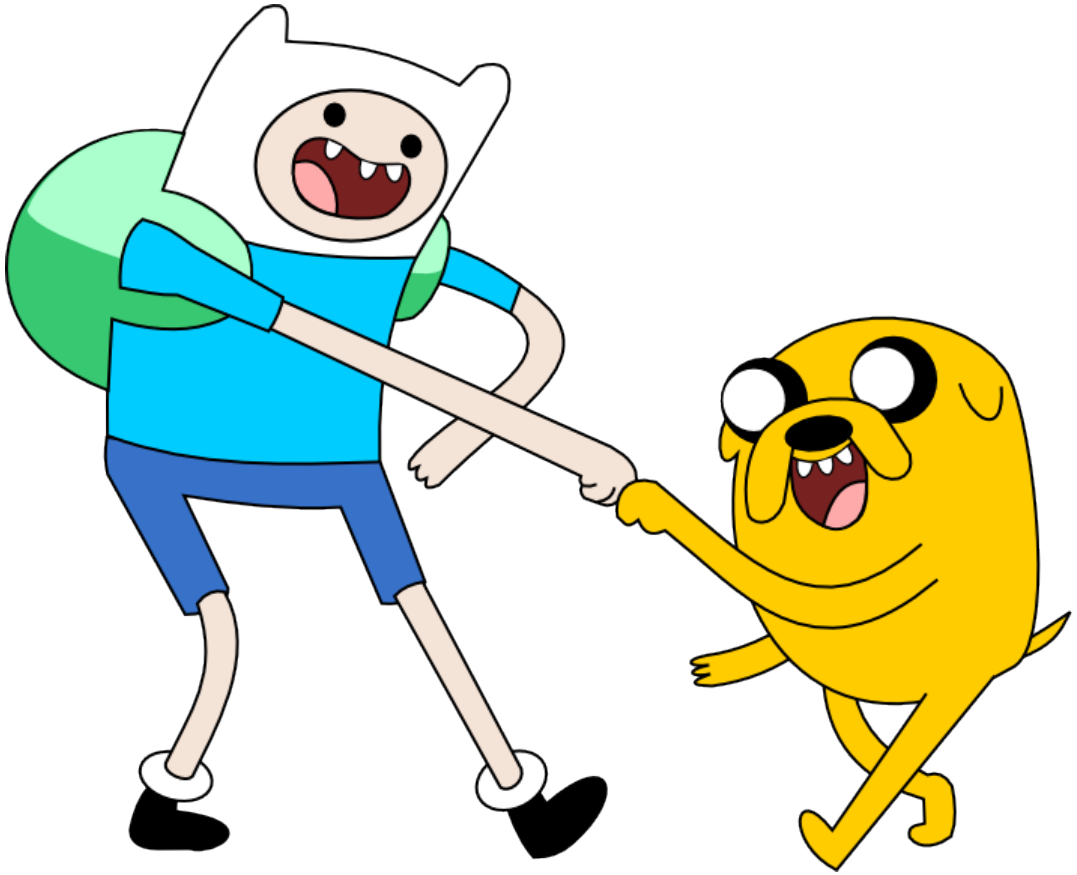
- Often easier to assess causal probabilities.

$$P(Measles|Spots) = \frac{P(Spots|Measles)P(Measles)}{P(Spots)}$$

- Can visualise this as:



Mathematical!



(Pendleton Ward/Cartoon Network)

What we learned so far

- 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 for efficient computation along with Bayes' rule.
- Next week we'll look at how they are used.