UNCERTAINTY

Introduction

- So far we have considered mainly accessible/observable environments.
  - Or pretended that environments were accessible/observable.
- Clearly not true of the real world:
  - Is is raining in Manhattan?
- Partial observability can arise for many reasons.
  - World structure *vs.* sensor ability.
  - Sensor noise.
  - Computational complexity.

Types of problem

- Who is outside in the corridor?
  - Uncertainty
- The radio says it is raining in Manhattan, but when I phone my wife she says it isn't raining.
  - Ambiguity
  - Contradiction.
- Is it true that "Simon is tall"
  - Vagueness
- Who will be in next year's World Series?
  - Ignorance

Uncertainty

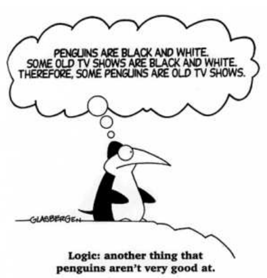
- Let action A_t = leave for airport t minutes before flight
  - Will A_t get me there on time?
- Problems:
  1. partial observability (road state, other drivers' plans, etc.)
  2. noisy sensors (1010 WINS traffic reports)
  3. uncertainty in action outcomes (flat tire, etc.)
  4. immense complexity of modelling and predicting traffic

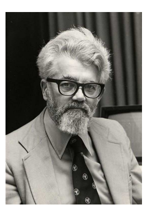
- Hence a purely logical approach either:
  1. risks falsehood: "A_{90} will get me there on time"
  2. leads to conclusions that are too weak for decision making: "A_{90} will get me there on time if there's no accident on the Williamsburg bridge, and it doesn't rain and my tires remain intact etc etc."
- (A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport . . .)

- How could an agent cope with this?

Methods for handling uncertainty

- *Nonmonotonic* logic
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?



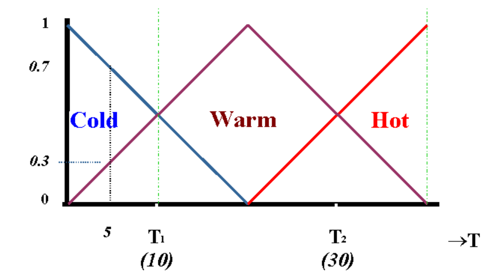


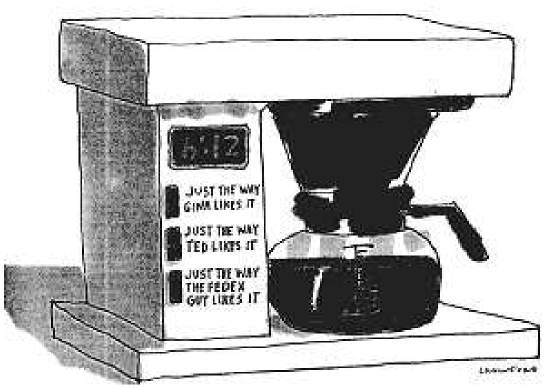
- *Rules with fudge factors*:
  - $A_{25} \mapsto_{0.3}$ *AtAirportOnTime*
  - *Sprinkler* $\mapsto_{0.99}$ *WetGrass*
  - *WetGrass* $\mapsto_{0.7}$ *Rain*
- Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*??
- Semantics?

- *Probability*
  - Given the available evidence, A_{25} will get me there on time with probability 0.04
- 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





Probability

- Probabilistic assertions *summarize* effects of
  - laziness: failure to enumerate exceptions, qualifications, etc.
  - ignorance: lack of relevant facts, initial conditions, etc.
- *Subjective* or *Bayesian* probability:
  - Probabilities relate propositions to one's own state of knowledge

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

- Probabilities of propositions change with new evidence:

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

(Analogous to logical entailment status $KB \models \alpha$, not truth.)

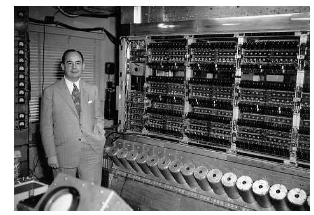
Making decisions under uncertainty

- 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?

- Depends on my *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 with a vengence next time.



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*.

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

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

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

- An *event* A is any subset of Ω

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

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

Random variables

- A *random variable* is a function from sample points to some range.
  - $raining(Brooklyn) = true$.
  - $temperature(234NE) = 73$
- P induces a *probability distribution* for any r.v. X:

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

$P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

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 $A(\omega) = true$
  event $\neg a$ = set of sample points where $A(\omega) = false$
  event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$
- Often in AI applications, the sample points are *defined* by the values of a set of random variables.
- A state can be defined by a set of Boolean variables.

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

This is then just a sample point.

- Thus, with Boolean variables, sample point = propositional logic model

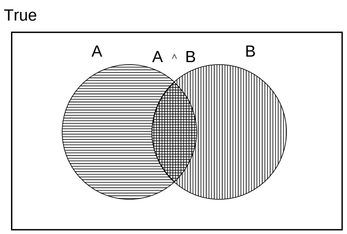
  $A = true, B = false \qquad a \wedge \neg b$
- Proposition = disjunction of atomic events in which it is true

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

Why use probability?

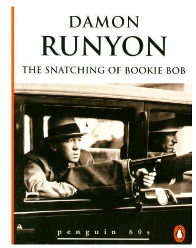
- The definitions imply that certain logically related events must have related probabilities

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



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.



Why Dutch?

Syntax for propositions

- *Propositional* or *Boolean* random variables
  - *Cavity* (do I have a cavity?)
  - $Cavity = true$ is a proposition, also written *cavity*
- *Discrete* random variables (finite or infinite)
  - *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)
  - $Temp = 21.6$; also allow, e.g., $Temp < 22.0$.
- Arbitrary Boolean combinations of basic propositions

Prior probability

- *Prior* or *unconditional probabilities* of propositions

  $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$

  correspond to belief before (prior) to arrival of any (new) evidence.
- *Probability distribution* gives values for all possible assignments:

  $\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$
- Distribution is 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

- *Every question about a domain can be answered by the joint distribution because every event is a sum of sample points*

Conditional probability

- *Conditional* or *posterior probabilities*

  $P(cavity|toothache) = 0.8$

  given that toothache is all I know NOT "if *toothache* then 80% chance of *cavity*"
- Notation for conditional distributions:

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

  A 2-element vector of 2-element vectors.

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

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

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

  $P(cavity|toothache, jetsWin) = P(cavity|toothache) = 0.8$
- This kind of inference, sanctioned by domain knowledge, is crucial

- 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 whole 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* is derived by successive application of product rule:

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

- Or, in terms of a more concrete example:

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

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 it is true:

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

	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:

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

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

	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 it is true:

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

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

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

- Can also compute conditional probabilities:

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

Normalization

	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

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

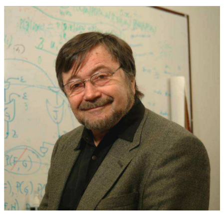
Inference by enumeration

- 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 $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$
- 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

- 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???
- This problem effectively stopped the use of probability in AI until the mid 80s

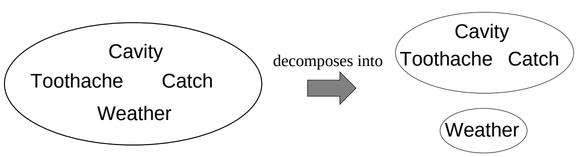


Independence

- A and B are *independent* iff

$$\begin{aligned} \mathbf{P}(A|B) &= \mathbf{P}(A), \text{ or} \\ \mathbf{P}(B|A) &= \mathbf{P}(B), \text{ or} \\ \mathbf{P}(A,B) &= \mathbf{P}(A)\mathbf{P}(B) \end{aligned}$$

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



- $\mathbf{P}(Toothache, Catch, Cavity, Weather)$
  $= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$
- 32 entries reduced to 12; for n independent biased coins, $2^n \to n$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. 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:

$$\mathbf{P}(catch|toothache, cavity) = \mathbf{P}(catch|cavity) \quad (1)$$

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

$$\mathbf{P}(catch|toothache, \neg cavity) = \mathbf{P}(catch|\neg cavity) \quad (2)$$

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

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

- Equivalent statements:

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

- 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)
- 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.
- 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.

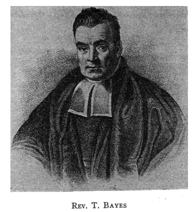
Bayes' Rule

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

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

  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:

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

- Let M be meningitis, S be stiff neck:

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

- Note: posterior probability of meningitis still very small!

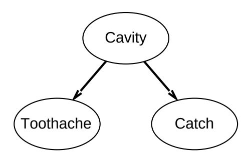
Bayes' Rule and conditional independence

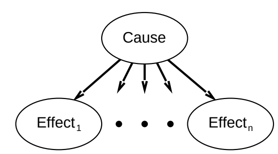
- So, in our running example

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

- This is an example of a *naive 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

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
- Next lecture we'll look at how this is used.