Uncertainty (Chapter 13)

July 25, 2021

Outline

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

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 (radio/GPS traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Purely logical approach either

1. risks falsehood: "A_{25} will get me there on time" ot
2. leads to conclusions that are too weak for decision making: "A_{25} will get me there on time if there's no accident on the 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 ...)

Methods for handling uncertainty

Default or 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*??
Probability

- Given the available evidence, A_{25} will get me there on time with probability 0.04 (Mahaviracarya (9th C.), Cardamo (1565) theory of gambling)

Methods for handling uncertainty

- Fuzzy logic handles *degree of truth* NOT probabilistic uncertainty e.g., *WetGrass* is true to degree 0.2

- Fuzzy logic handles uncertainty due to lack of definition, rather than to lack of data
  - Most human defined / natural concepts are fuzzy, in that their boundaries are not precise
  - Example: tall/short/medium height; young/old/middle age; round; cloudy; intelligent; beautiful

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 e.g., $P(A_{25}|\text{no reported accidents}) = 0.06$
- These are *not* claims of a "probabilistic tendency" in the current situation (but might be learned from past experience of similar situations)
- Probabilities of propositions change with new evidence: e.g., $P(A_{25}|\text{no reported accidents, 5 a.m.}) = 0.15$ (Analogous to logical entailment status $KB \models \phi$, 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/tollerance for missing flight vs. airport cuisine, etc.
- Utility theory is used to represent and infer preferences
- Decision theory = utility theory + probability theory

Probability basics

- Begin with 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 $\omega \in \Omega$ s.t.

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

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

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

Random variables

- A random variable is a function from sample points to some range, e.g., the reals or Booleans – e.g., $Odd(1) = true$.
- P induces a probability distribution for any r.v. X:

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

e.g.,

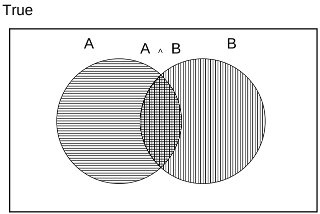
$$P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$$

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 $A(\omega) = \mathit{true}$
- event $\neg a$ = set of sample points where $A(\omega) = \mathit{false}$
- event $a \wedge b$ = points where $A(\omega) = \mathit{true}$ and $B(\omega) = \mathit{true}$
- Often in AI applications, the sample points are *defined* by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables
- With Boolean variables, sample point = propositional logic model – e.g., $A = \mathit{true}$, $B = \mathit{false}$, or $a \wedge \neg b$.
- Proposition = disjunction of atomic events in which it is true e.g., $(a \vee b) = (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$ $\implies P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Why use probability?

The definitions imply that certain logically related events must have related probabilities
E.g., $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



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.

Syntax for propositions

- **Propositional** or **Boolean** random variables.
  e.g., Cavity (do I have a cavity?):

  $$Cavity = true \text{ is a proposition, also written } cavity$$

- **Discrete** random variables (finite or infinite)

  $$\text{e.g., } Weather \text{ is one of } < sunny, rain, cloudy, snow >$$

  $Weather = rain$ is a proposition. Values must be exhaustive and **mutually exclusive.**
- **Continuous** random variables (**bounded** or **unbounded**)

  $$\text{e.g., } Temp = 21.6\text{; also allow, e.g., } Temp < 22.0$$

  Arbitrary Boolean combinations of basic propositions

Prior probability

- *Prior* or *unconditional probabilities* of propositions

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

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

  $(\text{Weather}) = < 0.72, 0.1, 0.08, 0.1 >$ (**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)

  $$(\text{Weather}, \text{Cavity}) = \text{a } 4 \times 2 \text{ matrix of values:}$$

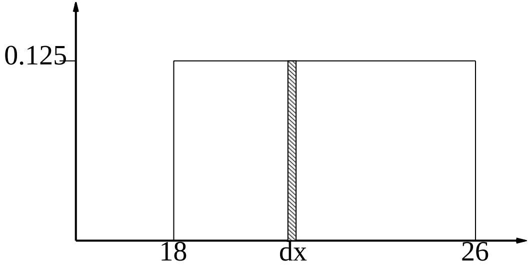
Weather =	*sunny*	*rain*	*cloudy*	*snow*
Cavity = *true*	0.144	0.02	0.016	0.02
Cavity = *false*	0.576	0.08	0.064	0.08

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

Probability for continuous variables

Express distribution as a parameterized function of value:

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

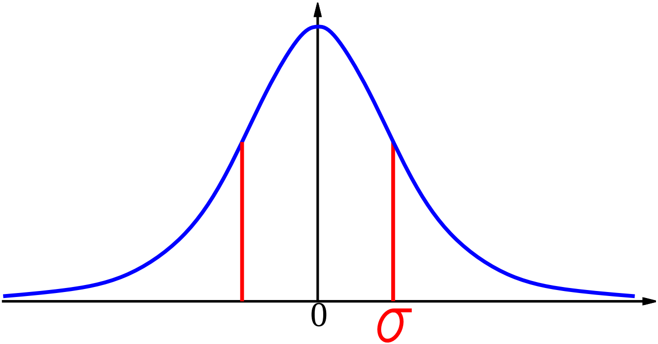


Here P is a density; integrates to 1.
$P(X = 20.5) = 0.125$ really means

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

Gaussian density

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



Conditional probability

- Conditional or posterior probabilities, e.g.,

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

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

$$(\text{Cavity }|\text{Toothache}) = 2 \times 2 \text{ matrix}$$

Cavity/Toothache	T	F
T	w	x
F	y	z

  such that $w + y = 1$ and $x + z = 1$
- If we know more, for example, *cavity* is also given, then we have

$$\text{P(cavity } | \text{ toothache,cavity)} = 1$$

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

Conditional probability

- New evidence may be irrelevant, allowing simplification, in which case adding it does not change the belief.
  - For example, we add that Cincinnati Reds won the last game - CinciRedsWin, which is (or ought to be) irrelevant to the probability of cavity, or toothache:

$$P(\text{cavity} \mid \text{toothache}, \text{CinciRedsWin}) = P(\text{cavity} \mid \text{toothache}) = 0.8$$

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 whole distributions, e.g.,

$$P(Weather, Cavity) = P(Weather|Cavity)P(Cavity)$$

(View as a 4×2 set of equations, *not* matrix mult.)

- Chain rule is derived by successive application of product rule:

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

Inference by enumeration – 1

- 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 it is true:

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

Inference by enumeration – 2

- 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:

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

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

Inference by enumeration – 3

- 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:

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

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

Inference by enumeration – 4

- 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**

- 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** k

$$P(Cavity|toothache) = kP(Cavity, toothache)$$

$$= k[P(Cavity, toothache, catch) + P(Cavity, toothache, \neg catch)]$$

$$= k[(0.108, 0.016) + (0.012, 0.064)]$$

$$= k(0.12, 0.08) = (0.6, 0.4)$$

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

Inference by enumeration – 5

- 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

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

Note: $\setminus$ denotes the set difference operation: for two sets A, B, $A \setminus B = \{x \mid x \in A \& x \notin B\} = A \cap \overline{B}$.

This is different from $A - B = \{x \mid x = a - b, \text{ with } a \in A \& b \in B\}$.

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

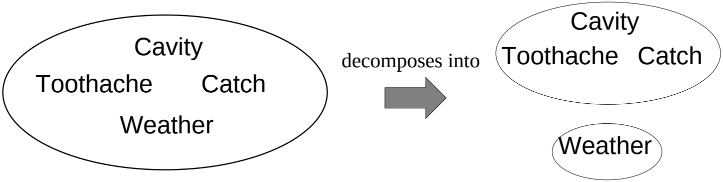
$$P(\mathbf{Y}|\mathbf{E} = \mathbf{e}) = kP(\mathbf{Y}, \mathbf{E} = \mathbf{e})$$
$$= k \sum_h P(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$$

  The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ 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???

Independence

- A and B are independent $\iff$

$$P(A|B) = P(A) \quad or \quad P(B|A) = P(B) \quad or \quad P(A,B) = P(A)P(B)$$



P(Toothache,Catch,Cavity,Weather)=P(Toothache,Catch,Cavity)P(Weather)

- 32 entries reduced to 12; for n independent biased coins, $2^n \to n$
- Absolute independence powerful but rare (e.g., Dentistry is a large field with hundreds of variables, none of which are independent.) What to do?

Conditional independence – 1

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

1. If cavity, the probability that the probe catches in it doesn't depend on toothache:

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

2. The same independence holds if NO cavity:

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

We say that *Catch* is conditionally independent of *Toothache* given *Cavity*:

$$P(Catch|Toothache, Cavity) = P(Catch|Cavity)$$

Equivalent statements:

$$P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$$

$$P(Toothache, Catch|Cavity) = P(Toothache|Cavity)P(Catch|Cavity)$$

Conditional independence – 2

Write out full joint distribution using chain rule:

$$P(Toothache, Catch, Cavity)$$

$$= P(Toothache|Catch, Cavity)P(Catch, Cavity)$$

$$= P(Toothache|Catch, Cavity)P(Catch|Cavity)P(Cavity)$$

$$= P(Toothache|Cavity)P(Catch|Cavity)P(Cavity)$$

i.e., $2 + 2 + 1 = 5$ 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.

Claim: *Conditional independence is the most basic and robust form of knowledge about uncertain environments.*

Bayes' Rule

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

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

or in distribution form

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

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!, although 8 times bigger than prior probability of meningitis.

Bayes' Rule and conditional independence

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

This is an example of a naive Bayes model:

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

Total number of parameters is *linear* in n

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