

UNCERTAINTY

(SLIDES ADAPTED FROM RUSSELL)

Outline

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

Uncertainty

Let action $A_t =$ leave for McGill t minutes before class
Will A_t get me there on time?

Problems:

- 1) partial observability (road state, other drivers' plans, etc.)
- 2) noisy sensors (traffic reports, weather forecast)
- 3) uncertainty in action outcomes (does STM bus stick to schedule, etc.)
- 4) immense complexity of modelling and predicting traffic

The Purely Logical Approach

Because of uncertainty, a purely logical approach either

- 1) risks falsehood: “ A_{25} will get me there on time”
- or 2) leads to conclusions that are too weak for decision making:
“ A_{25} will get me there on time if the STM hasn’t pulled a bus from service
and road construction doesn’t shut down Sherbrooke, etc.”

(A_{1440} might reasonably be said to get me there on time but I’d have to sleep at the university . . .)

Making decisions under uncertainty

Suppose I believe the following:

$$P(A_{15} \text{ gets me there on time} | \dots) = 0.04$$

$$P(A_{35} \text{ gets me there on time} | \dots) = 0.70$$

$$P(A_{1440} \text{ gets me there on time} | \dots) = 0.9999$$

Which action to choose depends on my **preferences** for waiting for bus, being late for class, etc.

Utility theory is used to represent and infer preferences

Decision theory = utility theory + probability theory

Maximum Expected Utility principle: An agent is rational iff it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action.

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

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

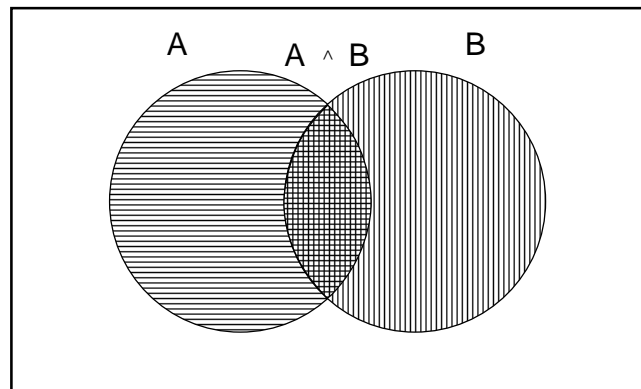
e.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Why use probability?

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

e.g., $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

True



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 is a proposition, also written *cavity*

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

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

Weather = rain is a proposition

Values must be exhaustive and mutually exclusive

Continuous random variables (*bounded* or *unbounded*)

e.g., *Temp = 21.6*; also allow, e.g., *Temp < 22.0*.

Arbitrary Boolean combinations of basic propositions

Prior probability

Prior or unconditional probabilities of propositions, e.g.,

$$P(\textit{Cavity} = \textit{true}) = 0.1$$

$$P(\textit{Weather} = \textit{sunny}) = 0.72$$

$$P(\textit{CoinToss} = \textit{heads}) = 0.5$$

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

Probability distribution

Probability distribution gives values for all possible assignments:

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

(normalized; 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:

<i>Weather =</i>	<i>sunny</i>	<i>rain</i>	<i>cloudy</i>	<i>snow</i>
<i>Cavity = true</i>	0.144	0.02	0.016	0.02
<i>Cavity = false</i>	0.576	0.08	0.064	0.08

Probability distribution

Probability distribution gives values for all possible assignments:

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

(normalized; 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:

<i>Weather =</i>	<i>sunny</i>	<i>rain</i>	<i>cloudy</i>	<i>snow</i>
<i>Cavity = true</i>	0.144	0.02	0.016	0.02
<i>Cavity = false</i>	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

Probability assessments are updated to reflect the new evidence available.

Conditional or posterior probabilities

e.g., $P(\text{cavity}|\text{toothache}) = 0.8$

i.e., **given that *toothache* is all I know**

NOT “if *toothache* then 80% chance of *cavity*”

(Notation for conditional distributions:

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

Knowing more

We may have

$$P(\text{cavity}|\text{toothache}) = 0.8$$

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

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

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

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

$$P(\text{cavity}|\text{toothache}, \text{busOnTime}) = P(\text{cavity}|\text{toothache}) = 0.8$$

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

Back to the coin toss

Posterior or conditional probability after we gather evidence.

$$P(\text{heads} | \text{a run of several tails}) = 0.5 - \epsilon$$

$$P(\text{heads} | \text{a run of many tails}) = \psi, \psi < 0.5$$

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.,

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

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

Chain rule is derived by successive application of product rule:

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

Inference by enumeration

Start with the joint distribution:

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

Can also compute conditional probabilities:

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

Normalization

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

Denominator can be viewed as a normalization constant α

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

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

Inference by enumeration, contd.

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:

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

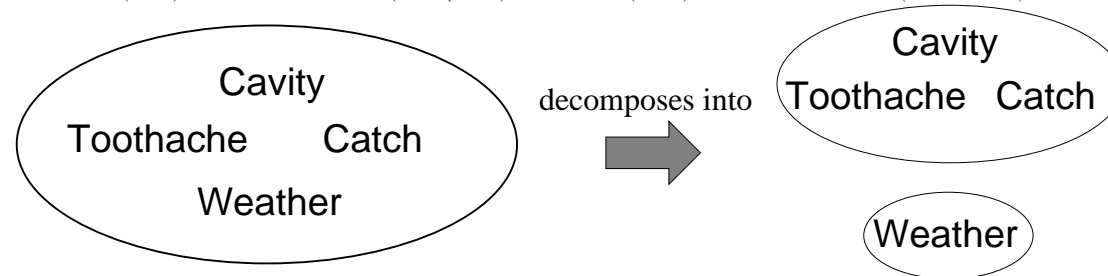
The terms in the summation are joint entries because \mathbf{Y} , \mathbf{E} , and \mathbf{H} together exhaust the set of random variables

Requires input table of size $O(2^n)$ for n Boolean variables!
How do we find the values for these $O(2^n)$ entries???

Independence

A and B are independent iff

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



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

32 entries reduced to 12; for n independent biased coins, $2^n \rightarrow n$

Absolute independence powerful but rare

Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

$\mathbf{P}(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$ 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:

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

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

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

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

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

Equivalent statements:

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

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

Conditional independence contd.

Write out full joint distribution using chain rule:

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

i.e., $2 + 2 + 1 = 5$ independent numbers (eq. 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.

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

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

Useful for assessing **diagnostic** probability from **causal** probability:

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

Bayes' Rule Example

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

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

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

This is an example of a **naive Bayes** model:

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



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