COSC1125/1127 Artificial Intelligence

Week 9: Probability

[RN2] Part V – Chapters 13 & 14

[RN3] Part IV – Chapters 13 & 14

A decision making scenario

How long does it take to get to the airport?
How long before the flight should you leave?

30 minutes? "Might be okay"
90 minutes? "Should be fine"
150 minutes? "Yes, that will work"
12 hours? "What! Way too early!!"

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

Hence 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 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 …)

Uncertainty…

A simple diagnosis example

- A possible rule using first-order logic

  $\forall p\ Symptom(p, Toothache) \Rightarrow Disease(p, Cavity)$

- The above rule could be wrong…, so we change it to

  $\forall p\ Symptom(p, Toothache) \Rightarrow Disease(p, Cavity) \lor Disease(p, GumDisease)$
  $\lor\ Disease(p, ImpactedWisdom) \lor \ldots\ldots$

- The problem is that we could have an unlimited list. How about

  $\forall p\ Disease(p, Cavity) \Rightarrow Symptom(p, Toothache)$

- The above rule is still incorrect, since not all cavities cause pain.

How come predicates calculus does not work here? Because “Cavity causes toothache” is not a certain rule, but a probable rule.

Probability

Probabilistic assertions *summarize* effects of

- laziness: failure to enumerate exceptions, qualifications, etc.

  (e.g. enumerate all causes of toothache: cavity, injury …)

- ignorance: lack of relevant facts, initial conditions, etc.

  (e.g. toothache might have unknown causes…)

Subjective or **Bayesian probability**:

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

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

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

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

[more explanation later]

Probability

Probabilities are not claims of some probabilistic tendency in the current situation. They might be learned from past experience of similar situations.

Probabilities of propositions might change when new information arrives.

Probability measures degree of belief.

A probability of 0.8 does not mean "80% true" but rather an 80% degree of belief (a fairly strong expectation). E.g. "Probability of rain is 0.8" does not mean "it is 80% raining", but "the chance of raining is 80%".

This is different from "degree of truth", which is the subject of fuzzy logic.

Making decisions under uncertainty

Suppose I believe the following:

$P(A_{25}$ gets me there on time | …) = 0.04

$P(A_{90}$ gets me there on time | …) = 0.70

$P(A_{120}$ gets me there on time | …) = 0.95

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

Which action to choose?

- Depends on my preferences for missing flight vs. time spent waiting, etc.

Utility theory is used to represent and infer preferences

Decision theory = probability theory + utility theory

Probability Basics

Begin with a set Ω – the sample space

E.g. 6 possible rolls of a die.

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

E.g. rolled number is 1.

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

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

$$\sum_w P(w) = 1$$

Use die as example: $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$$

Probability Basics…

An **event** *A* is any subset of Ω

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

E.g., $P\,(\,Die\ Roll < 4\,) = P\,(1) + P\,(2) + P\,(3) = 1/6 + 1/6 + 1/6 = ½$

A **random variable** refers to a "part" of the world whose "status" is initially unknown, e.g. die roll could be an odd or even number. Each random variable has a domain of values that it can take on.

P induces a *probability distribution* for any random variable *X*:

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

E.g., $P\,(\,Odd(die) = true\,) = P\,(1) + P\,(3) + P\,(5) = 1/6 + 1/6 + 1/6 = ½$

Propositions

Think of a proposition as the event (set of sample points) where the proposition is true.

Given Boolean random variable A and B:

– event a $=$ set of sample points where A *is true*

– event $\neg$ a $=$ set of sample points where A *is false*

– event a $\wedge$ b $=$ points where both A and B *are 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.

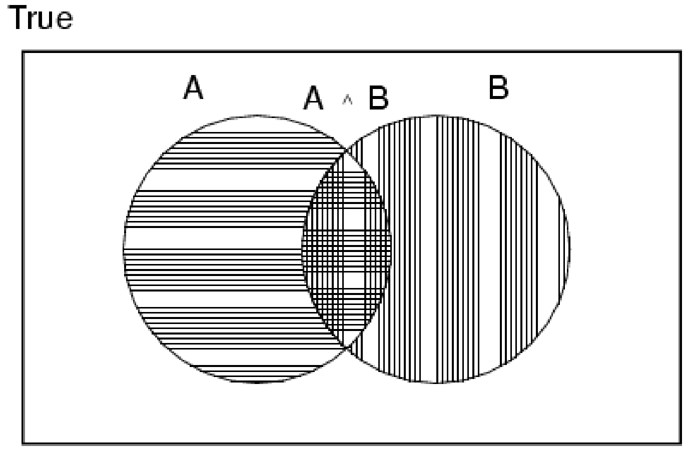
Propositions…

With Boolean variables, sample point = propositional logic model

E.g., $A = true, B = false, a \wedge \neg b$

Proposition = disjunction of atomic events in which it is true. E.g.,

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



Certain logically related events must have related probabilities, e.g,

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

Syntax

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

- Discrete random variable (finite or infinite)
  e.g., *Weather* 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)
  e.g., *Temperature* = *21.6*
  Temperature < *22.0* is also allowed

- Arbitrary Boolean combination of basic propositions.
  e.g., $(Temperature < 22.0) \wedge (Weather = rain)$

Prior Probability

- **Prior** or **unconditional probabilities** of propositions

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

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

- **Probability distribution** gives values for all possible assignments:

  $P(Weather) = <0.72,0.1,0.08,0.1>$ (normalized, i.e., sums to 1)

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables

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

Weather =	sunny	rainy	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**

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

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

– Notation for conditional distributions:

P(Cavity | Toothache) = 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, e.g.,

P(cavity | toothache, sunny) = P(cavity | toothache) = 0.8

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

Conditional Probability

– Definition of conditional probability:

$$P(a \mid b) = P(a \wedge b) \ / \ P(b) \text{ if } P(b) > 0$$

– **Product rule** gives an alternative formulation:

$$P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$$

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

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

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

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

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

Inference by Enumeration

– Start with the joint probability 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:

$$P(\Phi) = \Sigma_{\omega:\omega \models \Phi} P(\omega)$$

Inference by Enumeration

– Start with the joint probability 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:

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

$$P(\mathit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = \underline{0.2}$$

Inference by Enumeration

– Start with the joint probability 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:

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

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

Inference by Enumeration

– Start with the joint probability distribution:

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

– Can also compute conditional probabilities:

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

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

P(*Cavity* | *toothache*) = α **P**(*Cavity,toothache*)

= α [**P**(*Cavity,toothache,catch*) + **P**(*Cavity,toothache,¬ catch*)]

= α [<**P** *(cavity,toothache,catch*), P (*¬ cavity,toothache,catch*)> + <**P**(*cavity,toothache,¬ catch*), **P**(*¬ cavity,toothache,¬ catch*) >]

= α [<0.108,0.016> + <0.012,0.064>]

= α <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, contd.

Typically, we are interested in

the posterior joint distribution of the **query variables Y**

given specific values **e** for the **evidence variables E**

Let the **hidden variables** be **H** = **X** – **Y** – **E**

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

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

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

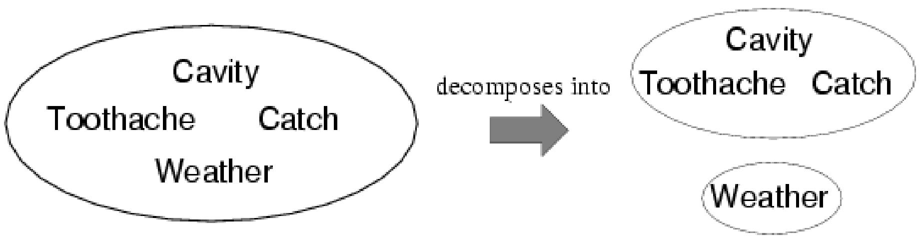
Obvious problems:

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?

Independence

A and B are independent iff

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



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

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

Absolute independence powerful but rare.

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

Conditional Independence

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*,¬*cavity*) = **P**(*catch* | ¬*cavity*)

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...

Write out full joint distribution using chain rule:

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

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

Bayes' Rule

Product rule $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

$\Rightarrow$ **Bayes' rule**: $$P(a \mid b) = \frac{P(b \mid a)\ P(a)}{P(b)}$$

or in distribution form

$$\mathrm{P}(Y \mid X) = \frac{\mathrm{P}(X \mid Y)\ \mathrm{P}(Y)}{\mathrm{P}(X)} = \alpha\, \mathrm{P}(X \mid Y)\mathrm{P}(Y)$$

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

$$\mathrm{P}(Cause \mid Effect) = \frac{\mathrm{P}(Effect \mid Cause)\ \mathrm{P}(Cause)}{\mathrm{P}(Effect)}$$

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

$$P(m|s) = P(s|m)\, P(m) \,/\, P(s) = 0.8 \times 0.0001 \,/\, 0.1 = 0.0008$$

Note: posterior probability of meningitis still very small!

Bayes' Rule and Conditional Independence

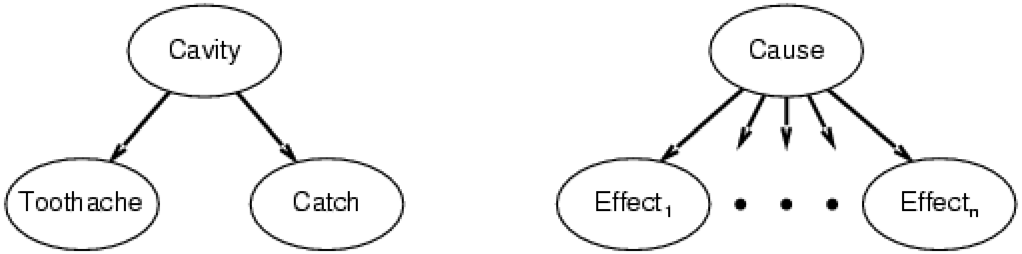
$\mathbf{P}(\mathit{Cavity} \mid \mathit{toothache} \wedge \mathit{catch})$

$= \alpha\, \mathbf{P}(\mathit{toothache} \wedge \mathit{catch} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{Cavity})$

$= \alpha\, \mathbf{P}(\mathit{toothache} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{catch} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{Cavity})$

This is an example of a **naïve Bayes** model

$\mathbf{P}(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause})\, \pi_i \mathbf{P}(\text{Effect}_i \mid \text{Cause})$



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

Summary

- Uncertainty arises because of both laziness and ignorance.
- Probabilities express the agent's inability to reach a definite decision regarding the truth of a sentence. Probabilities summarize the agent's belief.
- Prior probabilities and conditional probabilities.
- Full joint probability distribution.
- Bayes' rule computes unknown probabilities from known conditional probabilities, usually in the causal direction.
- Conditional independence.
- Naïve Bayes model assumes the conditional independence of all effect variables.

Acknowledgement: the slides were developed based on notes from Russell & Norvig's text, and several RMIT computer science staff members over the years.