Uncertainty

Chapter 13

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 (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 …)

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:
  - $A_{25} \mid\!\!\rightarrow_{0.3}$ get there on time
  - *Sprinkler* $\mid\!\!\rightarrow_{0.99}$ *WetGrass*
  - *WetGrass* $\mid\!\!\rightarrow_{0.7}$ *Rain*
- Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*??
- Probability
  - Model agent's degree of belief
  - Given the available evidence,
  - A_{25} will get me there on time with probability 0.04

Probability

Probabilistic assertions summarize effects of
- laziness: failure to enumerate exceptions, qualifications, etc.
- ignorance: lack of relevant facts, initial conditions, etc.

Subjective probability:
- Probabilities relate propositions to agent's own state of knowledge

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

These are not assertions about the world

Probabilities of propositions change with new evidence:

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

Making decisions under uncertainty

Suppose I believe the following:

$P(A_{25}$ gets me there on time $\mid \ldots) = 0.04$
$P(A_{90}$ gets me there on time $\mid \ldots) = 0.70$
$P(A_{120}$ gets me there on time $\mid \ldots) = 0.95$
$P(A_{1440}$ gets me there on time $\mid \ldots) = 0.9999$

- 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

Need way of reasoning

- Update belief state based on an action and percept – calculate new probabilities – choose the action with highest expected utility.
- Formal Language for reasoning under uncertainty – probability theory
  - Nature of sentences to which degrees of belief are assigned (**propositions**)
  - Dependence of the degree of belief on the agent's experience

Syntax of language for reasoning about uncertainty

- Basic element: random variable – refers to a part of the world whose "status" is initially unknown
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.
- Boolean random variables
  e.g., *Cavity* (do I have a cavity?)
- Discrete random variables
  e.g., *Weather* is one of <*sunny,rainy,cloudy,snow*>
- Domain values must be exhaustive and mutually exclusive
- Elementary proposition constructed by assignment of a value to a random variable: e.g., *Weather = sunny*, *Cavity = false* (abbreviated as ¬*cavity*)
- Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather = sunny ^ Cavity = false*

Syntax

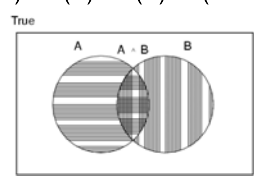
- Atomic event: A complete specification of the state of the world about which the agent is uncertain

  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  Cavity = false ^Toothache = false
  Cavity = false ^ Toothache = true
  Cavity = true ^ Toothache = false
  Cavity = true ^ Toothache = true
- Atomic events are mutually exclusive and exhaustive

Axioms of probability

- For any propositions *A, B*
  - $0 \leq P(A) \leq 1$
  - $P(true) = 1$ and $P(false) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Prior probability

- Prior or unconditional probabilities of propositions

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

  The above corresponds to P(Weather=sunny)=0.72, P(Weather=rain)=0.1, P(Weather=cloudy)=0.08, P(Weather=snow)=0.1

Prior probability

Sometimes it is useful to think about a complete set of variables all at once..

- 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

Conditional probability

- Once the agent has some evidence concerning the previously unknown random variables, prior probabilities are no longer applicable.
- 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*) = gives the values P(Cavity=xi | Toothache=yj) for all values of xi and yj – here 4 elements)

Conditional probability

- If we know more, e.g., *cavity* is also given, then we have
  P(*cavity* | *toothache,cavity*) = 1
- 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 | b) = P(a ^ b) / P(b) if P(b) > 0
- Product rule gives an alternative formulation:
  P(a ^ b) = P(a | b) P(b) = P(b | a) P(a)

Conditional probability

- A general version holds for whole distributions, e.g.,
  P(*Weather,Cavity*) = **P**(*Weather* | *Cavity*) **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 using Full Joint Distributions

- Probabilistic inference – the computation from observed evidence of posterior probabilities for query propositions.
- We can use the full joint distribution as a Knowledge Base from which we can derive answers to all questions about the probabilities of the probabilities of the values for the random variables.

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

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: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 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: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
- P(cavity V toothache) =
  0.108 + 0.012 + 0.016 + 0.064 + 0.072 + 0.008

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:

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

Inference by enumeration

- Just to check, let's try opposite value for cavity:

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

- Can also compute conditional probabilities:

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.108+0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

Normalization

- Notice in the above two calculations the term 1/P(toothache) remains constant, no matter what value of Cavity we calculate.
- It can be viewed as a normalization constant for the distribution ensuring that it sums to 1. The book uses α to denote that constant.
- We can calculete the above two values at once

P(Cavity | toothache)

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, [<0.108,0.016> + <0.012,0.064>]
 = α, <0.12,0.08> = <0.6,0.4>

General Inference Procedure: compute distribution on query variable (Cavity) by fixing evidence variables (Toothache) and summing over hidden variables (Catch)

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**

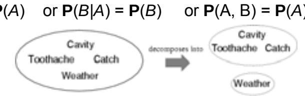
Then the required summation of joint entries is done by summing out the hidden variables:

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

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



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

Bayes' Rule

- Product rule P(a^b) = P(a | b) P(b) = P(b | a) P(a)

  ⇒ Bayes' rule: P(a | b) = P(b | a) P(a) / P(b)

This rule underlies all modern AI systems for probabilistic inference.

- or in distribution form

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

- Or conditionalized on some background evidence e

  $\mathbf{P}(Y|X,e) = \mathbf{P}(X|Y,e)\,\mathbf{P}(Y,e) / \mathbf{P}(X,e)$

Bayes' Rule

⇒ Bayes' rule: P(a | b) = P(b | a) P(a) / P(b)

What's the win?

- Useful for assessing diagnostic probability from causal probability:
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)
  - E.g., let *M* be meningitis, *S* be stiff neck:
    P(m|s) = P(s|m) P(m) / P(s) = 0.8 × 0.0001 / 0.1 = 0.0008
  - Note: doctor may know p(m|s) through observations (i.e., may have quantitative information in the diagnostic direction from symptoms to causes).
  - But, diagnostic knowledge is often more fragile than causal knowledge. E.g., P(m) would go up in an epidemic – as would P(m|s) – so observations no longer correct. P(s|m) would not change, however.

Bayes' Rule: Combining Evidence

How can we apply Bayes' rule when we have more evidence? E.g.,

P(*Cavity* | *toothache* ^ *catch*) ?

With the full joint distribution this is easy – just read off the values:
P(*Cavity* | *toothache* ^ *catch*) = α<0.108, 0.016> ≈ <0.871, 0.129>

Problem: this approach won't scale!

How about using Bayes' rule?

P(*Cavity* | *toothache* ^ *catch*)
= α**P**(*toothache* ^ *catch* | *Cavity*) **P**(*Cavity*)

Still have a problem with scaling up as we have more evidence. Might as well use the full joint distribution!

Conditional independence

P(*Cavity* | *toothache* ^ *catch*)
= α**P**(*toothache* ^ *catch* | *Cavity*) **P**(*Cavity*)

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries
- It would be nice if Toothache and Catch were independent – but they are not. However…
- 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 contd.

- 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*.
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayes' Rule and conditional independence

P(*Cavity* | *toothache* ^ *catch*)
= α**P**(*toothache* ^ *catch* | *Cavity*) **P**(*Cavity*)
= α**P**(*toothache* | *Cavity*) **P**(*catch* | *Cavity*) **P**(*Cavity*)

- This is an example of a naïve Bayes model:
  $\mathbf{P}(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause}) \pi_i \mathbf{P}(\text{Effect}_i | \text{Cause})$

- 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