Uncertainty

Chapter 13

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, noisy sensors)
2. uncertainty in action outcomes (flat tire, etc.)
3. 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 …)

Probability to the Rescue

- Probability
  - Model agent's degree of belief, given the available evidence.
  - A_{25} will get me there on time with probability 0.04

Probability in AI models our ignorance, not the true state of the world.

The statement “With probability 0.7 I have a cavity” means:
I either have a cavity or not, but I don’t have all the necessary information to know this for sure.

Probability

Subjective probability:

- Probabilities relate propositions to agent's own state of knowledge
  e.g., $P(A_{25} \mid \text{no reported accidents at 3 a.m.}) = 0.06$

- Probabilities of propositions change with new evidence:
  e.g., $P(A_{25} \mid \text{no reported accidents at 5 a.m.}) = 0.15$

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

Syntax

Capital letter: random variable
lower case: single value

- Basic element: random variable

- 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*>

- 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 (i.e. a full assignment of values to all variables in the universe, a unique single world).

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

  $Cavity = false \wedge Toothache = false$
  $Cavity = false \wedge Toothache = true$
  $Cavity = true \wedge Toothache = false$
  $Cavity = true \wedge Toothache = true$

- Atomic events are mutually exclusive and exhaustive

  if some atomic event is true, then all other other atomic events are false.

  There is always some atomic event true.

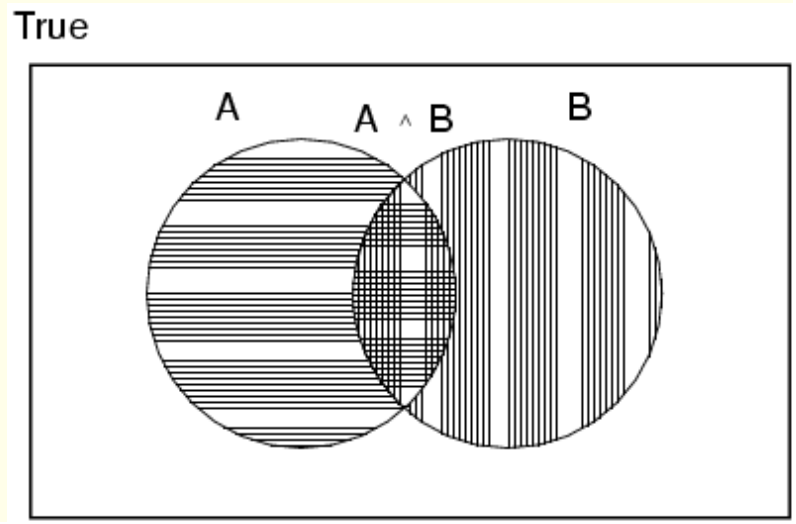
  Hence, there is exactly 1 atomic event true.

Axioms of probability

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

true in all worlds e.g. P(a OR NOT(a))

false in all worlds: P(a AND NOT(a))



Think of P(A) as the number of worlds in which A is true divided by the total number of possible worlds.

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)

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

- Conditional or posterior probabilities
  e.g., P(*cavity* | *toothache*) = 0.8 i.e., given that *Toothache*=*true* is all I know.

- Note that **P**(Cavity|Toothache) is a 2x2 array, normalized over columns.

- 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

Conditional probability

- Definition of conditional probability:
  $P(a \mid b) = P(a \wedge b) / P(b) \quad \text{if} \quad P(b) > 0$
- Product rule gives an alternative formulation:
  $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$
- Bayes Rule: $P(a|b) = P(b|a) P(a) / P(b)$
- 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 multiplication)
- Chain rule is derived by successive application of product rule:

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

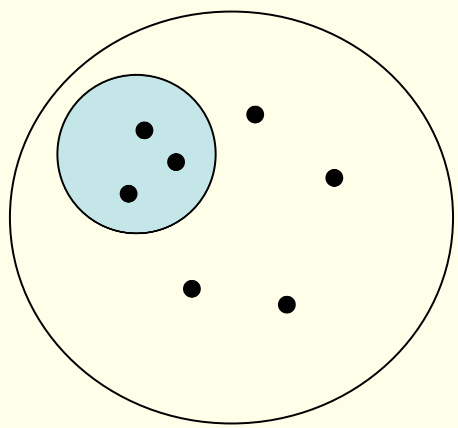
Inference by enumeration

- 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 a, sum the atomic events where it is true: $P(a) = \Sigma_{\omega \text{ s.t. } a=\text{true}} P(\omega)$

P(a)=1/7 + 1/7 + 1/7 = 3/7



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 a, sum the atomic events where it is true: $P(a) = \Sigma_{\omega:\omega \text{ s.t. } a=\text{true}} P(\omega)$

- $P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

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

- Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

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 x **P**(*Cavity,toothache*)
= α x [**P**(*Cavity,toothache,catch*) + **P**(*Cavity,toothache*,¬ *catch*)]
= α x [<0.108,0.016> + <0.012,0.064>]
= α x <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

Typically, we are interested in
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 $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

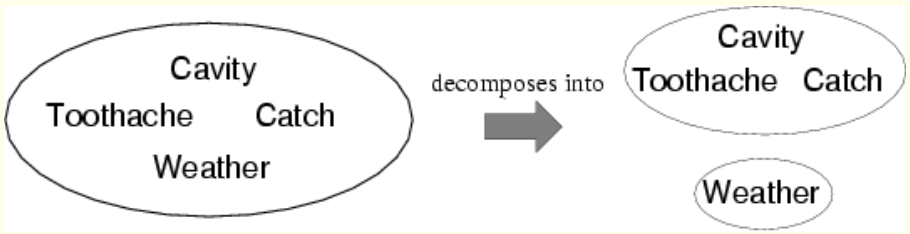
Then the required summation of joint entries is done by summing out the hidden variables:

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

- The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$ 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
  $\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(\mathrm{A, B}) = \mathbf{P}(A)\,\mathbf{P}(B)$



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

- 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*)

Note: catch and toothache are *not independent*, they are *conditionally independent* given that I know cavity.

Conditional independence cont.

- 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

- Product rule $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
  $\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a)\, P(a) / P(b)$
- or in distribution form
  $$\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\, \mathbf{P}(Y) / \mathbf{P}(X) = \alpha \mathbf{P}(X|Y)\, \mathbf{P}(Y)$$
- 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:
    $P(m|s) = P(s|m)\, P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$
  - Note: even though the probability of having a stiff neck given meningitis is very large (0.8), the posterior probability of meningitis given a stiff neck is still very small (why?).
  - $P(s|m)$ is more ‘robust’ than $P(m|s)$. Imagine a new disease appeared which would also cause a stiff neck, then $P(m|s)$ changes but $P(s|m)$ not.

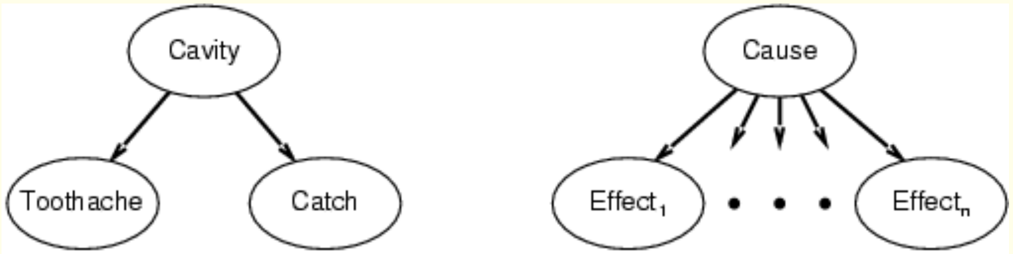
Bayes' Rule and conditional independence

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

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

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

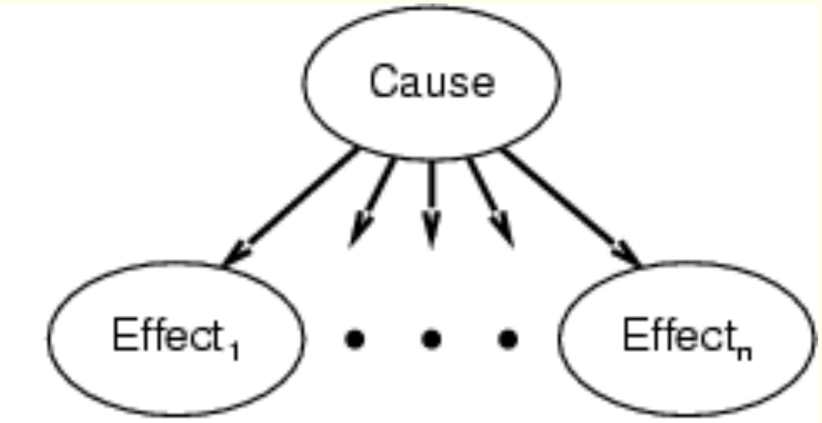
- This is an example of a naïve Bayes model:
  $\mathbf{P}(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause})\ \pi_i \mathbf{P}(\text{Effect}_i \mid \text{Cause})$



- Total number of parameters is linear in *n*

- A naive Bayes classifier computes: P(cause|effect1, effect2...)

The Naive Bayes Classifier



Imagine we have access to the probabilities of

1. P(disease)
2. P(symptoms|disease)=P(headache|disease)P(backache|disease)....

Then, the probability of a disease is computed using Bayes rule:

P(disease|symptoms) = constant x P(symptoms|disease) x P(disease)

Learning a Naive Bayes Classifier

What to do if we only have observations from a doctors office?

For instance:
flu1→ headache, fever, muscle ache
lungcancer1 → short breath, breast pain
flu2 → headache, fever, cough
....

In general {(x1,y1), (x2,y2), (x3,y3),....}

symptoms (attributes)

disease (label)

$$P(\text{disease} = y) = \frac{\#\text{ people with disease } y}{\text{total }\#\text{ of people in dataset}} = \text{fraction of people with disease } y$$

$$P(\text{symptom\_A}=\text{x\_A}|\text{disease} = y) = \frac{\#\text{ people with disease } y \text{ that have symptom A}}{\text{total }\#\text{ people with disease } y}$$

4.(12 pts) **Uncertainty**

Joe needs to go to the doctor to check if he has "monkey pox" (MP). The doctor asks him 2 questions: 1) "Do you have red bumps (RB) on your body" and 2) "Do you have a fever (FR)". Joe's answers are "yes, I have red bumps" (RB=T), and "yes, I have a fever" (FR=T). The doctor (now worried) has the following joint probability table at his disposal:

	MP=T	MP=T	MP=F	MP=F
	RB=T	RB=F	RB=T	RB=F
FR = T	7.2E-7	1.8E-7	0.71999928	0.17999982
FR = F	8E-8	2E-8	0.07999992	0.01999998

a.(4 pts) Compute the prior probability of getting monkey pox $P(MP = T)$ from the joint table.

a) answer: $P(MP = T) = 7.2e - 7 + 1.8e - 7 + 8e - 8 + 2e - 8 = 1e - 6.$

b.(4 pts) Compute the conditional probability $P(FR = T, RB = T|MP = T)$.

b) answer: $P(FR = T, RB = T|MP = T) = P(FR = T, RB = T, MP = T)/P(MP = T) = 0.72$

c.(4 pts) Use Bayes' rule to compute what the doctor needs to know: $P(MP = T|FR = T, RB = T)$. Explain why this probability is actually very small, even though all the symptoms for Monkey Pox are present.

b) answer: $P(MP = T|FR = T, RB = T) = P(FR = T, RB = T, MP = T)/P(FR = T, RB = T) = 7.2e - 7/(7.2e - 7 + 0.71999928) = 1e - 6$ This is small because the prior probability on MP is very small, and the symptoms FR and RB did not add any information to the prior probability.

5.(20 pts) **Probability**

John likes recognizing cars. He classifies cars into one of 3 classes: Car=[Ferrari,RollsRoyce,Other]. John observes 3 features: Color=[red,other], Speed=[fast,slow] and Weight=[heavy,light]. We will assume that the features Color, Speed and Weight are all conditionally independent given Car. Furthermore, it is given that:

P(Color=red|Car=Ferrari)=0.5 ,
P(Speed=high|Car=Ferrari)=0.5,
P(Weight=light|Car=Ferrari)=0.9,
P(Color=red|Car=RollsRoyce)=0,
P(Speed=high|Car=RollsRoyce)=0.1,
P(Weight=light|Car=RollsRoyce)=0,
P(Color=red|Car=other)=0.1,
P(Speed=high|Car=other)=0.4,
P(Weight=light|Car=other)=0.5,
P(Car=Ferrari)=0.01 (John lives in Newport Beach),
P(Car=RollsRoyce)=0.01.

a.(4 pts) Use conditional independence to express P(Color,Speed,Weight,Car) as function of P(Color|Car), P(Speed|Car), P(Weight|Car) and P(Car).

a)answer: P(Color,Speed,Weight,Car)=P(Color|Car)P(Speed|Car)P(Weight|Car)P(Car).

b.(4 pts) How many entries does the joint probability table have for P(Color,Speed,Weight,Car)?

b)answer: 24 entries.

c.(4pts) Using the available information, compute the probability of:
P(Color=red,Weight=light,Speed=high,Car=Ferrari) and of
P(Color=other,Weight=heavy,Speed=low,Car=RollsRoyce).

c)answer: P(Color=red,Weight=light,Speed=high,Car=Ferrari)=0.5x0.9x0.5x0.01=0.00225
P(Color=other,Weight=heavy,Speed=low,Car=RollsRoyce)=1*1*0.9*0.01=0.009

d.(4 pts) Use Bayes rule to express P(Car|Color,Speed,Weight) in terms of the joint probability table. Note: this expression may involve terms where you need to sum over all possible values of certain variables.

d)answer: P(Car|Color,Speed,Weight)=P(Color,Speed,Weight,Car)/P(Color,Speed,Weight). The denominator can be be expressed a sum over all values for Car of the joint probability table.

e.(4 pts) John sees a car and observes: Color=red, Speed=high, Weight=light. Compute the probability that the car is a Ferrari.

e) answer: Applying the equation in c: $0.0025/(0.0025 + 0 + 0.0196) = 0.113$

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