22c:145 Artificial Intelligence

Uncertainty

- Reading: Ch 13. Russell & Norvig

Problem of Logic Agents

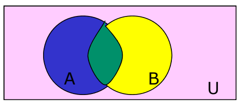
- Logic-agents almost never have access to the whole truth about their environments.
- A rational agent is one that makes rational decisions in order to maximize its performance measure.
- Logic-agents may have to either risk falsehood or make weak decisions in uncertain situation
- A rational agent's decision depends on relative importance of goals, likelihood of achieving them.
- Probability theory provides a quantitative way of encoding likelihood

Foundations of Probability

- Probability Theory makes the same ontological commitments as FOL
- Every sentence S is either true or false.
- The *degree of belief, or probability,* that S is true is a *number* P between 0 and 1.
- P(S) = 1 iff S is certainly true
- P(S) = 0 iff S is certainly false
- P(S) = 0.4 iff S is true with a 40% chance
- P(not A) = probability that A is false
- P(A and B) = probability that both A and B are true
- P(A or B) = probability that either A or B (or both) are true

Axioms of Probability

- All probabilities are between 0 and 1
- Valid propositions have probability 1. Unsatisfiable propositions have probability 0. That is,
  - P(A v : A) = P(true) = 1
  - P(A Æ : A) = P(false) = 0
  - P(: A) = 1 – P(A)
- The probability of disjunction is defined as follows.
  - P(A v B) = P(A) + P(B) – P(A Æ B)
  - P(A Æ B) = P(A) + P(B) – P(A v B)



Exercise Problem I

Prove that

- P(A v B v C) =
  P(A) + P(B) + P(C) –
  P(A Æ B) – P(A Æ C) – P(B Æ C) +
  P(A Æ B Æ C)

How to Decide Values of Probability

P(the sun comes up tomorrow) = 0.999

- Frequentist
  - Probability is inherent in the process
  - Probability is estimated from measurements

Probs can be wrong!

A Question

Jane is from Berkeley. She was active in anti-war protests in the 60’s. She lives in a commune.

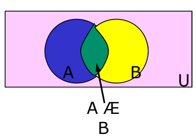
- Which is more probable?
  1. Jane is a bank teller
  2. Jane is a feminist bank teller

A Question

Jane is from Berkeley. She was active in anti-war protests in the 60’s. She lives in a commune.

- Which is more probable?
  1. Jane is a bank teller
  2. Jane is a feminist bank teller

1. A
2. A Æ B



Conditional Probability

- P(A) is the unconditional (or prior) probability
- An agent can use unconditional probability of A to reason about A only in the absence of no further information.
- If some further evidence B becomes available, the agent must use the conditional (or posterior) probability:

  P(A|B)

  the probability of A given that the agent already knew that B is true.
- P(A) can be thought as the conditional probability of A with respect to the empty evidence:

  P(A) = P(A|).

Conditional Probability

1. P(Blonde) =
2. P(Blonde | Swedish) =
3. P(Blonde | Kenian) =
4. P(Blonde | Kenian Æ : EuroDescent) =

- If we know nothing about a person, the probability that he/she is blonde equals a certain value, say 0.1.
- If we know that a person is Swedish the probability that s/he is blonde is much higher, say 0.9.
- If we know that the person is Kenyan, the probability s/he is blonde much lower, say 0.000003.
- If we know that the person is Kenyan and not of European descent, the probability s/he is blonde is basically 0.
- **Computation:** P(A | B) = P(A Æ B)/P(B)

Random Variables

Variable	Domain
Age	{ 1, 2, …, 120 }
Weather	{ sunny, dry, cloudy, raining }
Size	{ small, medium, large }
Raining	{ true, false }

- The probability that a random variable X has value val is written as P(X=val)
- P: domain ! [0, 1]
  - Sums to 1 over the domain:
    - P(Raining = true) = P(Raining) = 0.2
    - P(Raining = false) = P(: Raining) = 0.8

Probability Distribution

- If X is a random variable, we use the **bold case P(X)** to denote a vector of values for the probabilites of each individual element that **X** can take.
- Example:
  - P(Weather = sunny) = 0.6
  - P(Weather = rain) = 0.2
  - P(Weather = cloudy) = 0.18
  - P(Weather = snow) = 0.02
- Then **P(**Weather**)** = <0.6, 0.2, 0.18, 0.02> (the value order of "sunny", "rain", "cloudy", "snow" is assumed).
- **P(**Weather**)** is called a probability distribution for the random variable Weather.

- Joint distribution: **P**(X1, X2, …, Xn)
  - Probability assignment to all combinations of values of random variables

Joint Distribution Example

	Toothache	:Toothache
Cavity	0.04	0.06
: Cavity	0.01	0.89

- The sum of the entries in this table has to be 1

Joint Distribution Example

	Toothache	:Toothache
Cavity	0.04	0.06
: Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]

Joint Distribution Example

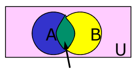
	Toothache	:Toothache
Cavity	0.04	0.06
: Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]
- P(A | B) = P(A Æ B)/P(B) [prob of A when U is limited to B]

Joint Distribution Example

	Toothache	:Toothache
Cavity	0.04	0.06
: Cavity	0.01	0.89

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
- P(cavity) = 0.1 [add elements of cavity row]
- P(toothache) = 0.05 [add elements of toothache column]
- P(A | B) = P(A Æ B)/P(B) [prob of A when U is limited to B]
- P(cavity | toothache) = 0.04/0.05 = 0.8



A Æ B

Joint Probability Distribution (JPD)

- A joint probability distribution **P**(X1, X2 …, Xn) provides complete information about the probabilities of its random variables.
- However, JPD's are often hard to create (again because of incomplete knowledge of the domain).
- Even when available, JPD tables are very expensive, or impossible, to store because of their size.
- A JPD table for n random variables, each ranging over k distinct values, has k^n entries!
- A better approach is to come up with conditional probabilities as needed and compute the others from them.

Bayes' Rule

- Bayes' Rule
  - P(A | B) = P(B | A) P(A) / P(B)
- What is the probability that a patient has meningitis (M) given that he has a stiff neck (S)?
  - P(M|S) = P(S|M) P(M)/P(S)

P(S|M) is easier to estimate than P(M|S) because it refers to causal knowledge:

- meningitis typically causes stiff neck.
- P(S|M) can be estimated from past medical cases and the knowledge about how meningitis works.
- Similarly, P(M), P(S) can be estimated from statistical information.

Bayes' Rule

- Bayes' Rule: P(A | B) = P(B | A) P(A) / P(B)
- The Bayes rule is helpful even in absence of (immediate) causal relationships.
- What is the probability that a blonde (B) is Swedish (S)?
- P(S|B) = P(B|S) P(S)/P(B)
- All P(B|S), P(S), P(B) are easily estimated from statistical information.
  - P(B|S) = (# of blonde Swedish)/(Swedish population) = 9/10
  - P(S) = Swedish population/world population = …
  - P(B) = # of blondes/world population = …

Lecture 14 • 19

Conditional Independence

- Conditioning
  - P(A) = P(A | B) P(B) + P(A | :B) P(:B)
    = P(A Æ B) + P(A Æ :B)
- In terms of exponential explosion, conditional probabilities do not seem any better than JPD's for computing the probability of a fact, given n>1 pieces of evidence.
  - P(Meningitis | StiffNeck Æ Nausea Æ … Æ DoubleVision)
- However, certain facts do not always depend on all the evidence.
  - P(Meningitis | StiffNeck Æ Astigmatic) = P(Meningitis | StiffNeck)
- Meningitis and Astigmatic are conditionally independent, given StiffNeck.

Lecture 14 • 20

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) ¢ P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)

Lecture 14 • 21

Independence

- A and B are independent iff
  - P(A Æ B) = P(A) ¢ P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)
- Independence is essential for efficient probabilistic reasoning

- A and B are conditionally independent given C iff
  - P(A | B, C) = P(A | C)
  - P(B | A, C) = P(B | C)
  - P(A Æ B | C) = P(A | C) ¢ P(B | C)

Lecture 14 • 22

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)

Lecture 14 • 23

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

Lecture 14 • 24

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
- None of these propositions are independent of one other
- T and X are conditionally independent given C

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)
- None of these propositions are independent of one another
- R and S are conditionally independent given B

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} $|\rightarrow_{0.3}$ get there on time
  - *Sprinkler* $|\rightarrow_{0.99}$ *WetGrass*
  - *WetGrass* $|\rightarrow_{0.7}$ *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

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* ⋁ *cavity*) = 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:

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

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 **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>

where α = 1/ **P**(*toothache*)

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 **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:

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

where $\alpha = 1/\ \mathbf{P}(\mathbf{E} = \mathbf{e})$

- 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?**

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:
  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: posterior probability of meningitis still very small!

Lecture 14 • 37

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

- Total number of parameters is linear in *n*

Lecture 14 • 38

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

Lecture 14 • 39

Bayesian Networks

- To do probabilistic reasoning, you need to know the joint probability distribution
- But, in a domain with N propositional variables, one needs 2^N numbers to specify the joint probability distribution
- We want to exploit independences in the domain
- Two components: structure and numerical parameters

Bayesian networks

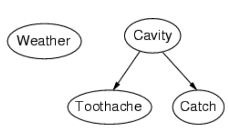
- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- Syntax:
  - a set of nodes, one per variable
  -
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:

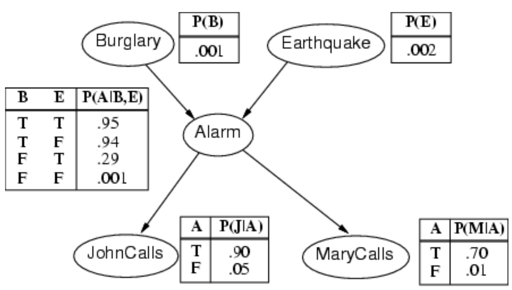


- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



P(B)
.001

P(E)
.002

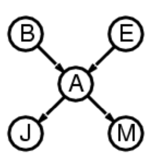
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Compactness

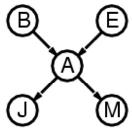
- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values
- Each row requires one number p for X_i = *true* (the number for X_i = *false* is just $1-p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, 1 + 1 + 4 + 2 + 2 = 10 numbers (vs. 2^5-1 = 31)



Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$



e.g., $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$= \mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}(\neg e)$

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For i = 1 to n
  - add X_i to the network
  -
  - select parents from $X_1, \ldots, X_{i-1}$ such that
    $\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$

This choice of parents guarantees:

$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$
(chain rule)

$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$
(by construction)

Example

- Suppose we choose the ordering *M, J, A, B, E*

$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

Example

- Suppose we choose the ordering *M, J, A, B, E*

P*(J | M) =* ***P****(J)?*
No
P*(A | J, M) =* ***P****(A | J)*? ***P****(A | J, M) =* ***P****(A)*?

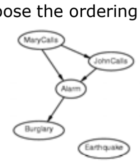
Example

- Suppose we choose the ordering *M, J, A, B, E*

P*(J | M) =* ***P****(J)?*
No
P*(A | J, M) =* ***P****(A | J)*? ***P****(A | J, M) =* ***P****(A)*? **No**
P*(B | A, J, M) =* ***P****(B | A)*?
P*(B | A, J, M) =* ***P****(B)*?

Example

- Suppose we choose the ordering M, J, A, B, E



P*(J | M) =* ***P****(J)?*
No
P*(A | J, M) =* ***P****(A | J)*? ***P****(A | J, M) =* ***P****(A)*? **No**
P*(B | A, J, M) =* ***P****(B | A)*? **Yes**
P*(B | A, J, M) =* ***P****(B)*? **No**
P*(E | B, A ,J, M) =* ***P****(E | A)*?
P*(E | B, A, J, M) =* ***P****(E | A, B)*?

Example

- Suppose we choose the ordering M, J, A, B, E

P*(J | M) =* ***P****(J)?*
No
P*(A | J, M) =* ***P****(A | J)*? ***P****(A | J, M) =* ***P****(A)*? **No**
P*(B | A, J, M) =* ***P****(B | A)*? **Yes**
P*(B | A, J, M) =* ***P****(B)*? **No**
P*(E | B, A ,J, M) =* ***P****(E | A)*? **No**
P*(E | B, A, J, M) =* ***P****(E | A, B)*? **Yes**

Example contd.

- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct