Uncertainty

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 (KCBS 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:
  - $A_{25} \rightarrow_{0.3}$ AtAirportOnTime
  - Sprinkler $\rightarrow_{0.99}$ WetGrass
  - WetGrass $\rightarrow_{0.7}$ Rain
- Issues: Problems with combination, e.g., Sprinkler causes Rain??

- Probability
  - Given the available evidence, A_{25} will get me there on time with probability 0.04

- (Fuzzy logic handles degree of truth NOT uncertainty e.g., WetGrass is true to degree 0.2)

Probability

- Probabilistic assertions summarize effects of
  - Laziness: failure to enumerate exceptions, qualifications, etc.
  - Ignorance: lack of relevant facts, initial conditions, etc.
- Subjective or Bayesian probability:
- Probabilities relate propositions to one's own state of knowledge
  - e.g., $P(A_{25} | \text{no reported accidents}) = 0.06$
- These are not claims of a “probabilistic tendency” in the current situation (but might be learned from past experience of similar situations)
- Probabilities of propositions change with new evidence:
  - e.g., $P(A_{25} | \text{no reported accidents, 5 a.m.}) = 0.15$

Making Decisions Under Uncertainty

- Suppose I believe the following:
  - P(A_{25} gets me there on time | …) = 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. airport cuisine, etc.
- Utility theory is used to represent and infer preferences
- Decision theory = utility theory + probability theory

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 $\omega \in \Omega$ such that
  - $0 \leq P(\omega) \leq 1$
  - $\Sigma_{\omega} P(\omega) = 1$
- e.g., $P(1)=P(2)=P(3)=P(4)=P(5)=P(6)=1/6$
- An event A is any subset of Ω
  - $P(A) = \Sigma_{\{\omega \in A\}} P(\omega)$
- E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Random Variables

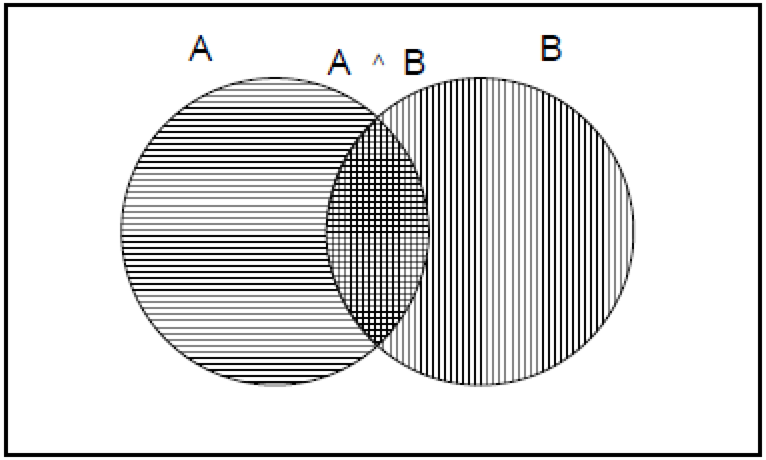
- A random variable is a function from sample points to some range, e.g., the reals or Booleans
  - e.g., Odd(1)=true.
- P induces a probability distribution for any random variable X:
  - $P(X = x_i) = \Sigma \{\omega:X(\omega)=x_i\} P(\omega)$
  - e.g., P(Odd=true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2

Propositions

- Think of a proposition as the event (set of sample points) where the proposition is true
- Given Boolean random variables A and B:
  - event a = set of sample points where $A(\omega)$=true
  - event ¬a = set of sample points where $A(\omega)$=false
  - event $a \land b$ = points where $A(\omega)$=true and $B(\omega)$=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
- With Boolean variables, sample point = propositional logic model
  - e.g., A=true, B =false, or $a \land \neg b$.
- Proposition = disjunction of atomic events in which it is true
  - e.g., $(a \lor b) \equiv (\neg a \land b) \lor (a \land \neg b) \lor (a \land b) \Rightarrow$ $P(a \lor b) = P(\neg a \land b) + P(a \land \neg b) + P(a \land b)$

Why use Probability?

- The definitions imply that certain logically related events must have related probabilities
- E.g., $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
- 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 <sunny, rain, cloudy, snow>
  - 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

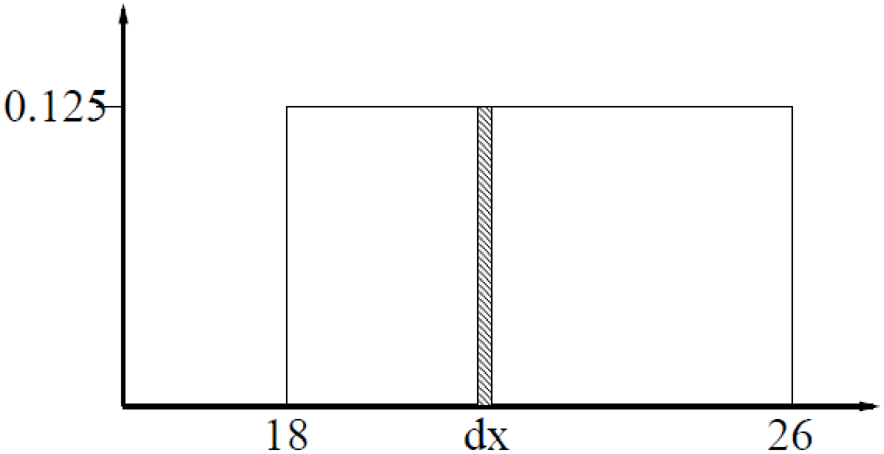
- Prior or unconditional probabilities of propositions
  - e.g., P(Cavity =true) = 0.1 and P(Weather =sunny) = 0.72
    correspond to belief prior to arrival of any (new) evidence
- Probability distribution gives values for all possible assignments:
  - P(Weather) = <0.72, 0.1, 0.08, 0.1>
    (normalized, i.e., sums to 1)
- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables (i.e., every sample point)
  - P(Weather,Cavity) = a 4 x 2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	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

Probability for Continuous Variables

- Express distribution as a parameterized function of value:
  - P(X =x) = U[18, 26](x) = uniform density between 18 and 26

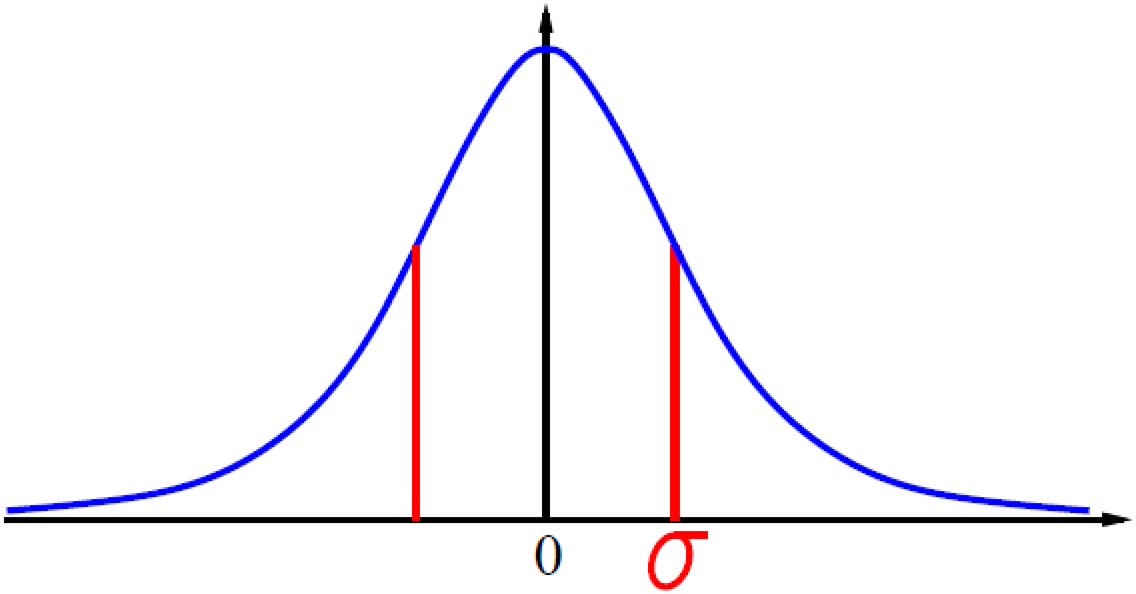


- Here P is a density; integrates to 1.
- P(X =20.5) = 0.125 really means

$$\lim_{dx \to 0} P(20.5 \le X \le 20.5 + dx)/dx = 0.125$$

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Gaussian Density



Conditional Probability

- Conditional or posterior probabilities
  - e.g., P(cavity|toothache) = 0.8
  - i.e., given that toothache is all I know
  - NOT “if toothache then 80% chance of cavity”
- 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, 49ersWin) = P(cavity|toothache) = 0.8
- This kind of inference, sanctioned by domain knowledge, is crucial

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.,
- P(Weather,Cavity) = P(Weather|Cavity)P(Cavity)
- (View as a 4 x 2 set of equations, not matrix multiplication)
- Chain rule is derived by successive application of product rule:

$$\begin{aligned}\mathbf{P}(X_1,\ldots,X_n) &= \mathbf{P}(X_1,\ldots,X_{n-1})\,\mathbf{P}(X_n|X_1,\ldots,X_{n-1})\\ &= \mathbf{P}(X_1,\ldots,X_{n-2})\,\mathbf{P}(X_{n_1}|X_1,\ldots,X_{n-2})\,\mathbf{P}(X_n|X_1,\ldots,X_{n-1})\\ &= \ldots\\ &= \prod_{i=1}^{n}\mathbf{P}(X_i|X_1,\ldots,X_{i-1})\end{aligned}$$

Inference by Enumeration

- Start with the joint 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:
- $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

Inference by Enumeration

- Start with the joint 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:
- $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

- Start with the joint distribution:

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

Inference by Enumeration

- For any proposition ϕ, sum the atomic events where it is true:
- $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

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

Inference by Enumeration

- Start with the joint distribution:

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

- Can also compute conditional probabilities

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

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

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

Normalization

- Denominator can be viewed as a normalization constant

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

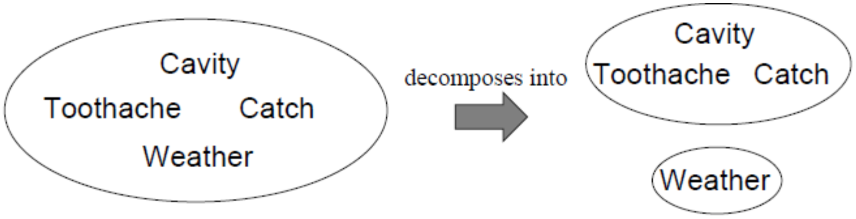
- General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Inference by Enumeration

- Let X be all the variables. Typically, we want 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:
- $P(Y|E=e) = \alpha P(Y,E=e) = \alpha \Sigma_h P(Y,E=e,H=h)$
- 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
- P(A|B)=P(A) or P(B|A)=P(B) or P(A,B)=P(A)P(B)
  - P(Toothache,Catch,Cavity,Weather) = P(Toothache,Catch,Cavity)P(Weather)
- 32 entries reduced to 12; for n independent biased coins, $2^n \longrightarrow n$
- 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:
  - 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 (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 $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$\Rightarrow$ 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:

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

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

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

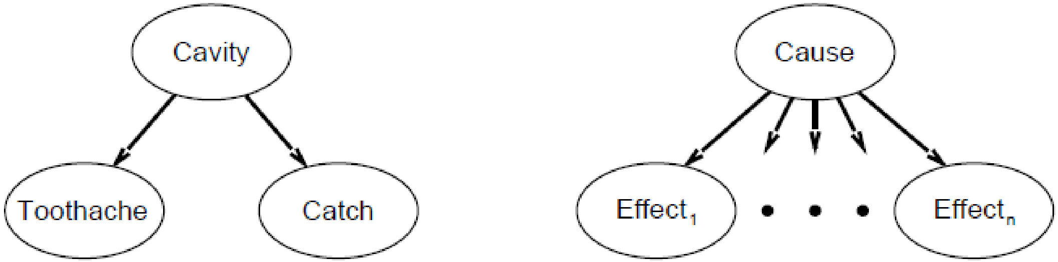
Note: posterior probability of meningitis still very small!

Bayes' Rule and Conditional Independence

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

This is an example of a naive Bayes model:

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



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

An Example

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Estimated Probabilities for Weather Data

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/ 14	5/ 14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

Bayes's Rule

- Probability of event *H* given evidence *E:*

$$Pr[H \mid E] = \frac{Pr[E \mid H]Pr[H]}{Pr[E]}$$

- *A priori* probability of *H* : $Pr[H]$
  - Probability of event *before* evidence is seen
- *A posteriori* probability of *H* : $Pr[H \mid E]$
  - Probability of event *after* evidence is seen

Naïve Bayes for Classification

- Classification learning: what's the probability of the class given an instance?
  - Evidence E = instance
  - Event H = class value for instance
- Naïve assumption: evidence splits into parts (i.e. attributes) that are *independent*

$$Pr[H \mid E] = \frac{Pr[E_1 \mid H]Pr[E_2 \mid H] * \cdots * Pr[E_n \mid H]Pr[H]}{Pr[E]}$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Evidence E

$$Pr[yes \mid E] = Pr[Outlook = Sunny \mid yes]$$

$$\times Pr[Temperature = Cool \mid yes]$$

$$\times Pr[Humidity = High \mid yes]$$

$$\times Pr[Windy = True \mid yes]$$

$$\times \frac{Pr[yes]}{Pr[E]} \qquad = \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{Pr[E]}$$

Probability of class "yes"

The “Zero-Frequency Problem”

- What if an attribute value doesn’t occur with every class value?
(e.g. “Outlook = overcast” for class “no”)
  - Probability will be zero!
  - *A posteriori* probability will also be zero!
(No matter how likely the other values are!)
- Remedy: add a small value to the count for every attribute value-class combination (*Laplace estimator)*
- Result: probabilities will never be zero!
(also: stabilizes probability estimates)

$$Pr[Humidity = High \mid yes] = 0$$

$$Pr[yes \mid E] = 0$$

Modified Probability Estimates

- Example: attribute *outlook* for class *yes*

$$\frac{2+\mu/3}{9+\mu} \qquad \frac{4+\mu/3}{9+\mu} \qquad \frac{3+\mu/3}{9+\mu}$$

Sunny ***Overcast*** ***Rainy***

- Weights don't need to be equal (but they must sum to 1)

$$\frac{2+\mu p_1}{9+\mu} \qquad \frac{4+\mu p_2}{9+\mu} \qquad \frac{3+\mu p_3}{9+\mu}$$

Missing Values

- Training: instance is not included in frequency count for attribute value-class combination
- Classification: attribute will be omitted from calculation
- Example:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

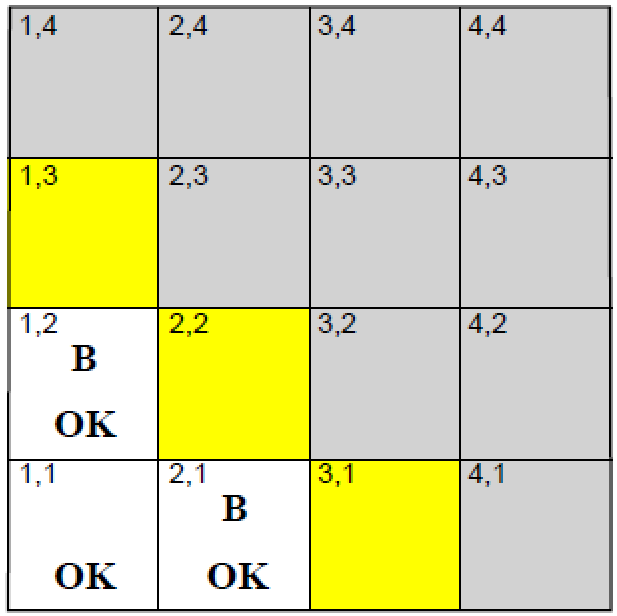
Likelihood of "yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$

Likelihood of "no" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$

P("yes") = $0.0238 / (0.0238 + 0.0343) = 41\%$

P("no") = $0.0343 / (0.0238 + 0.0343) = 59\%$

Wumpus World



$P_{ij} = true$ iff $[i, j]$ contains a pit

$B_{ij} = true$ iff $[i, j]$ is breezy
Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Specifying the Probability Model

First term: 1 if pits are adjacent to breezes, 0 otherwise

Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \Pi_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Observations and Query

We know the following facts:

$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$

$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

Query is $\mathbf{P}(P_{1,3}|known, b)$

Define $Unknown = P_{ij}$s other than $P_{1,3}$ and $Known$

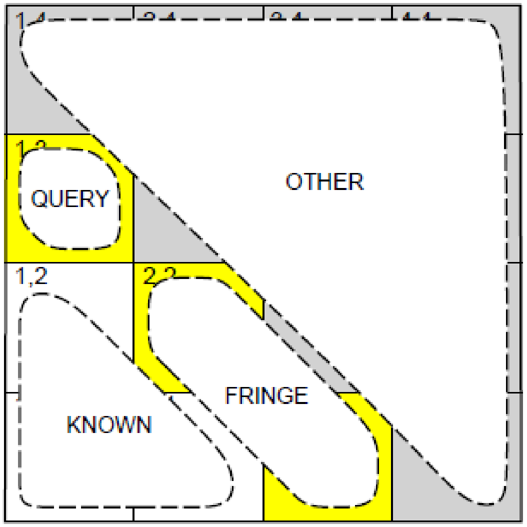
For inference by enumeration, we have

$$\mathbf{P}(P_{1,3}|known, b) = \alpha \Sigma_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Grows exponentially with number of squares!

Using Conditional Independence

Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares

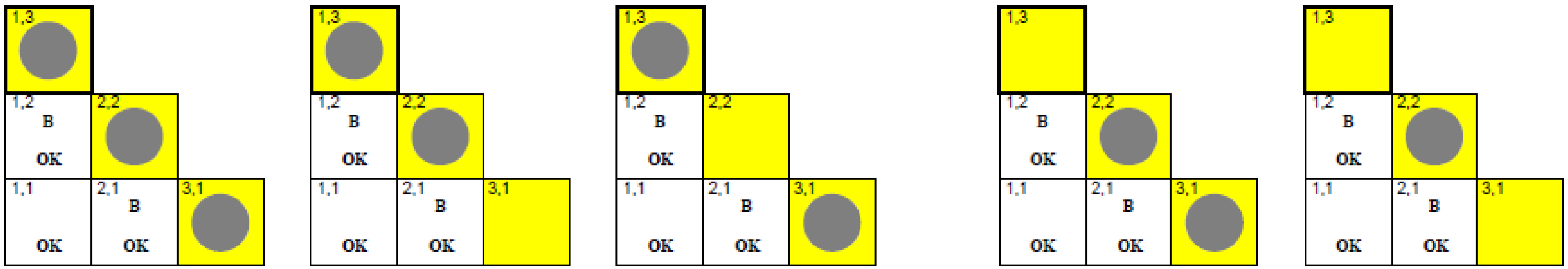


Define $Unknown = Fringe \cup Other$

$\mathbf{P}(b|P_{1,3}, Known, Unknown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$

Manipulate query into a form where we can use this!

Using Conditional Independence



0.2 x 0.2 = 0.04   0.2 x 0.8 = 0.16   0.8 x 0.2 = 0.16   0.2 x 0.2 = 0.04   0.2 x 0.8 = 0.16

$$\mathbf{P}(P_{1,3}|known, b) = \alpha' \langle 0.2(0.04+0.16+0.16),\ 0.8(0.04+0.16) \rangle \approx \langle 0.31, 0.69 \rangle$$

$$\mathbf{P}(P_{2,2}|known, b) \approx \langle 0.86, 0.14 \rangle$$

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