Introduction to Artificial Intelligence

Instructor: Wei-Min Shen

Week 6.1 and 6.2
Knowledge Representation

• Readings for this week
  – Read Chapter 12 (Wed)
  – Read Chapter 13 (Friday)

• Methods of KR
  – Logic (PL & FOL)
  – Probability

Learn AI and ML
Learn how to learn by yourself
Today’s Lecture: Probability

• Where is probability? (in the world or in your mind)
• Notation: variable $X$, value $x_i$, $P(X=x_i)$, $P(X)$ denotes for all values of $X$, $P(X,Y)$, $P(X,y)$
• Remember two axioms
  – Sum axiom: $P(A|B)+P(\sim A|B)=1$
  – Product axiom: $P(AB|C)=P(A|C)P(B|AC)=P(B|C)P(A|BC)$
• They can do much more than logics
  – Deductive reasoning:
    • If $A\rightarrow B$ and $A$, then $B$
    • If $A\rightarrow B$ and $\sim B$, then $\sim A$
    • If $A\rightarrow B$ and $B$, then “$A$ become more plausible”
  – Inductive reasoning:
    • If $A\rightarrow B$ and $\sim A$, then “$B$ become less plausible”
    • If $A\rightarrow”B$ becomes more plausible” and $B$, then “$A$ become more plausible”
  – HM: work out the math why the above is true (ALFE, p102)
Outline

• Uncertainty
• Probability
• Syntax
• Semantics
• Inference rules
Example: Guessing a Pit

A = Agent
B = Breeze
S = Smell
P = Pit
W = Wumpus
OK = Safe
V = Visited
G = Glitter

Which one has a PIT?
In the world Wumpus

- Logic can only guess randomly which of [1,3], [2,2],[3,1] has a pit
- Using probability you can calculate which one is more likely have a pit than others
Uncertainty

Let action $A_t = \text{leave for airport } t \text{ minutes before flight}$
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 modelling 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:**

$A_{25} \mapsto_{0.3} \text{get there on time}$

$Sprinkler \mapsto_{0.99} WetGrass$

$WetGrass \mapsto_{0.7} Rain$

**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

Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

(Fuzzy logic handles *degree of truth* NOT uncertainty e.g.,

$WetGrass$ is true to degree 0.2)
Logics and Probabilities

• Not all knowledge are certain
  – Earlier stories of expert systems (medical, legal)
• Two big challenges for logic-like approaches
  – Common sense (vague): “Water flows down”
  – Uncertainty
• Probability =?= A Logic of Science
Examples of Probability

- N people (e.g., N=14) with different age $j$
  - Probability $P(j) = \frac{N(j)}{N}$
  - Most probable age? Media age=? Average $<j>$?
Example of Standard Deviation

- Deviation shows how spread the distribution
  - E.g., same median, average, most probable
- Variance $\sigma^2 = <(\Delta j)^2> = <j^2> - <j>^2$
  - where $\Delta j = j - <j>$
Probability Density $\rho(x)$

$$P_{ab} = \int_a^b \rho(x) \, dx,$$

$$\int_{-\infty}^{+\infty} \rho(x) \, dx = 1,$$

$$\langle x \rangle = \int_{-\infty}^{+\infty} x \rho(x) \, dx,$$

$$\langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x) \rho(x) \, dx,$$

$$\sigma^2 \equiv \langle (\Delta x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2.$$
Probability

Where is Probability? In the world or in your mind? E.g., when you are throwing a dice, ...

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

**Subjective** or **Bayesian** probability: (vs. Objective View)
Probabilities relate propositions to one’s own state of knowledge
  e.g., $P(A_{25}|\text{no reported accidents}) = 0.06$

These are **not** assertions about the world

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

(Analogous to logical entailment status $KB \models \alpha$, not truth.)
Example: Cavity-Toothache-Catch

Other facts and interactions may exist, but they are either insignificant, unknown or irrelevant. We leave them out.
### Table of “Uncertainties”

<table>
<thead>
<tr>
<th>Cavity</th>
<th>Catch</th>
<th>Toothache</th>
<th>Logic Truth</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>{0,1}</td>
<td>0.576</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>{0,1}</td>
<td>0.064</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>{0,1}</td>
<td>0.144</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>{0,1}</td>
<td>0.016</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>{0,1}</td>
<td>0.008</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>{0,1}</td>
<td>0.012</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>{0,1}</td>
<td>0.072</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>{0,1}</td>
<td>0.108</td>
</tr>
</tbody>
</table>

**Sum=1.000**

<table>
<thead>
<tr>
<th></th>
<th>Toothache</th>
<th>~Toothache</th>
<th></th>
<th>~Catch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catch</td>
<td>0.108</td>
<td>0.012</td>
<td>Catch</td>
<td>0.072</td>
</tr>
<tr>
<td>~Catch</td>
<td>0.016</td>
<td>0.064</td>
<td>~Catch</td>
<td>0.008</td>
</tr>
<tr>
<td>Cavity</td>
<td>0.016</td>
<td>0.064</td>
<td>0.144</td>
<td>0.576</td>
</tr>
<tr>
<td>~Cavity</td>
<td>0.016</td>
<td>0.064</td>
<td>0.144</td>
<td>0.576</td>
</tr>
</tbody>
</table>
(Fully) Joint Probability Distribution

P(X,Y,Z)

X: Cavity; Y: Catch; Z: Toothache
There is SO MUCH in the Distribution!

- How likely you have cavity?
- How likely you don’t have cavity?
- How likely you have toothache and cavity?
- How likely you have all three?
- How likely you have none?
- 
- 
- 
- ........
Making decisions under uncertainty

Suppose I believe the following:

\[ P(A_{25} \text{ gets me there on time}|\ldots) = 0.04 \]
\[ P(A_{90} \text{ gets me there on time}|\ldots) = 0.70 \]
\[ P(A_{120} \text{ gets me there on time}|\ldots) = 0.95 \]
\[ P(A_{1440} \text{ gets me there on time}|\ldots) = 0.99999 \]

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
Axioms of Probability

For any propositions $A$, $B$

1. $0 \leq P(A) \leq 1$
2. $P(\text{True}) = 1$ and $P(\text{False}) = 0$
3. $P(A \lor B) = P(A) + P(B) - P(A \land B)$

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.
Probability Density $\rho(x)$

$$P_{ab} = \int_a^b \rho(x) \, dx,$$

$$\int_{-\infty}^{+\infty} \rho(x) \, dx = 1,$$

$$\langle x \rangle = \int_{-\infty}^{+\infty} x \rho(x) \, dx,$$

$$\langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x) \rho(x) \, dx,$$

$$\sigma^2 \equiv \langle (\Delta x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2.$$
Similar to propositional logic: possible worlds defined by assignment of values to random variables.

Propositional or Boolean random variables
  e.g., \( Cavity \) (do I have a cavity?)
Include propositional logic expressions
  e.g., \( \neg Burglary \lor Earthquake \)

Multivalued random variables
  e.g., \( Weather \) is one of \( \{ \text{sunny, rain, cloudy, snow} \} \)
Values must be exhaustive and mutually exclusive

Proposition constructed by assignment of a value:
  e.g., \( Weather = \text{sunny} \); also \( Cavity = \text{true} \) for clarity
Syntax

Prior or unconditional probabilities of propositions

e.g., \( P(Cavity) = 0.1 \) and \( P(Weather = \text{sunny}) = 0.72 \)
correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:
\[
\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle \quad \text{(normalized, i.e., sums to 1)}
\]

Joint probability distribution for a set of variables gives
values for each possible assignment to all the variables
\[
\mathbf{P}(Weather, Cavity) = \text{a 4 \times 2 matrix of values:}
\]

<table>
<thead>
<tr>
<th>Weather = sunny</th>
<th>rain</th>
<th>cloudy</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = true</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cavity = false</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Syntax

Conditional or posterior probabilities
  e.g., $P(\text{Cavity}|\text{Toothache}) = 0.8$
  i.e., given that $\text{Toothache}$ is all I know

Notation for conditional distributions:
  $P(\text{Weather}|\text{Earthquake}) = 2$-element vector of $4$-element vectors

If we know more, e.g., $\text{Cavity}$ is also given, then we have
  $P(\text{Cavity}|\text{Toothache}, \text{Cavity}) = 1$

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.,
  $P(\text{Cavity}|\text{Toothache}, \text{49ers Win}) = P(\text{Cavity}|\text{Toothache}) = 0.8$

This kind of inference, sanctioned by domain knowledge, is crucial
### Syntax Examples

<table>
<thead>
<tr>
<th>Cavity</th>
<th>Catch</th>
<th>Toothache</th>
<th>Logic Truth</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>{0,1}</td>
<td>0.576</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>{0,1}</td>
<td>0.064</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>{0,1}</td>
<td>0.144</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>{0,1}</td>
<td>0.016</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>{0,1}</td>
<td>0.008</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>{0,1}</td>
<td>0.012</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>{0,1}</td>
<td>0.072</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>{0,1}</td>
<td>0.108</td>
</tr>
</tbody>
</table>

**Exercises:**

- $P(\text{Cavity}=1)=?$, $P(\text{Cavity})=?$, $P(\text{Catch}=0)=?$, $P(\text{Catch})=?$, $P(\text{Toothache}=1)=?$, ...
- $P(\text{Cavity}=1, \text{Catch}=1, \text{Toothache}=1)=?$, $P(\text{Cavity, Catch, Toothache})=?$
- $P(\text{Cavity}=0, \text{Catch}=0)=?$, $P(\text{Cavity, Catch}=0)=?$
Two Key Elements in Probability

Probability Distribution Model
- Variables, Value assignments (possible worlds)
- Represented as a table or a graph

- Inferences can be made from the model
  - Sum rule
  - Product rule
  - Conditional
  - Marginalization
  - Normalization
Probability Distributions (discrete vs continuous)

$P(X,Y,Z)$

$\begin{array}{cccccc}
000 & 001 & 010 & 011 & 100 & 101 \\
110 & 111
\end{array}$

$X$: Cavity; $Y$: Catch; $Z$: Toothache

$\rho(x) = \frac{1}{2\sqrt{\pi x}}$

$\frac{1}{2h}$

$h$
A complete probability model specifies every entry in the joint distribution for all the variables $X = X_1, \ldots, X_n$. I.e., a probability for each possible world $X_1 = x_1, \ldots, X_n = x_n$.

(Cf. complete theories in logic.)

E.g., suppose *Toothache* and *Cavity* are the random variables:

<table>
<thead>
<tr>
<th>Cavity = true</th>
<th>Toothache = true</th>
<th>Toothache = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = true</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Cavity = false</td>
<td>0.01</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Possible worlds are mutually exclusive $\Rightarrow P(w_1 \land w_2) = 0$

Possible worlds are exhaustive $\Rightarrow w_1 \lor \cdots \lor w_n$ is True

hence $\sum_i P(w_i) = 1$
How many possible worlds in a Discrete Distribution?

• How many possible worlds?
  – For $P(X,Y)$, where $X,Y$ are binary variables, it is 2x2
  – For $P(X,Y=y)$, it is 2x1

• In general, the number of possible worlds
  – Is the product of the size of each variable
How many possible worlds in a discrete probability distribution?

P(X,Y,Z)  

X: Cavity;  Y: Catch;  Z: Toothache

“Mutually Exclusive” & “Exhaustive”

“Possible worlds” “value assignments”

2 x 2 x 2 = 8
Possible worlds in these distributions?

<table>
<thead>
<tr>
<th></th>
<th>Toothache</th>
<th>~Toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catch</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>~Catch</td>
<td>0.016</td>
<td>0.064</td>
</tr>
</tbody>
</table>

P(Toothache | Cavity) = 0.108

P(Toothache) = 0.108

# possible worlds = 2x2

# possible worlds = 2
How many possible worlds in a continuous probability distribution?

There are infinite real numbers in \([0,h]\)
Two Key Elements in Probability

- Probability Distribution Model
  - Variables, Value assignments (possible worlds)
  - Represented as a table or a graph
- Inferences can be made from the model
  - Sum rule
  - Product rule
  - Conditional
  - Marginalization
  - Normalization
Inference Rules for $P(A^\wedge B)$

• Rules: You only need to remember two 😊
  – Sum axiom:
    • $P(A|B) + P(\neg A|B) = 1$
  – Product axiom: $P(AB|C) = P(A|C)P(B|AC) = P(B|C)P(A|BC)$

• Notations
  – Variable $X$, value $x_i$, $P(X=x_i)$,
  – $P(X)$ denotes for all values of $X$ as shown in the table, aka "joint probability distribution"
  – Mixed variables and values: $P(X,Y)$, $P(X,y)$
For any propositions $A$, $B$

1. $0 \leq P(A) \leq 1$
2. $P(True) = 1$ and $P(False) = 0$
3. $P(A \lor B) = P(A) + P(B) - P(A \land B)$

**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.
Conditional Probability $P(A \mid B)$

Definition of conditional probability:

$$P(A \mid B) = \frac{P(A \land B)}{P(B)} \text{ if } P(B) \neq 0$$

Product rule gives an alternative formulation:

$$P(A \land B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

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

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

(View as a $4 \times 2$ set of equations, not matrix mult.)

Chain rule is derived by successive application of product rule:

$$P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1}) P(X_n \mid X_1, \ldots, X_{n-1})$$

$$= P(X_1, \ldots, X_{n-2}) P(X_{n-1} \mid X_1, \ldots, X_{n-2}) P(X_n \mid X_1, \ldots, X_{n-1})$$

$$= \ldots$$

$$= \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$$
Bayes’ Rule

Product rule \( P(A \land B) = P(A|B)P(B) = P(B|A)P(A) \)

\[ \Rightarrow \text{Bayes’ rule } P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

Why is this 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!
Normalization

Suppose we wish to compute a posterior distribution over $A$ given $B = b$, and suppose $A$ has possible values $a_1 \ldots a_m$

We can apply Bayes’ rule for each value of $A$:
\[
P(A = a_1 | B = b) = P(B = b | A = a_1) P(A = a_1) / P(B = b)
\]
\[
\ldots
\]
\[
P(A = a_m | B = b) = P(B = b | A = a_m) P(A = a_m) / P(B = b)
\]

Adding these up, and noting that $\sum_i P(A = a_i | B = b) = 1$:

\[
1 / P(B = b) = 1 / \sum_i P(B = b | A = a_i) P(A = a_i)
\]

This is the normalization factor, constant w.r.t. $i$, denoted $\alpha$:

\[
P(A | B = b) = \alpha P(B = b | A) P(A)
\]

Typically compute an unnormalized distribution, normalize at end

e.g., suppose $P(B = b | A) P(A) = \langle 0.4, 0.2, 0.2 \rangle$

then $P(A | B = b) = \alpha \langle 0.4, 0.2, 0.2 \rangle = \frac{\langle 0.4, 0.2, 0.2 \rangle}{0.4 + 0.2 + 0.2} = \langle 0.5, 0.25, 0.25 \rangle$
### Normalization Example

<table>
<thead>
<tr>
<th>Cavity</th>
<th>Catch</th>
<th>Toothache</th>
<th>Logic Truth</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>{0,1}</td>
<td>0.576</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>{0,1}</td>
<td>0.064</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>{0,1}</td>
<td>0.144</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>{0,1}</td>
<td>0.016</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>{0,1}</td>
<td>0.008</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>{0,1}</td>
<td>0.012</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>{0,1}</td>
<td>0.072</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>{0,1}</td>
<td>0.108</td>
</tr>
</tbody>
</table>

P(Toothache | Cavity=1) = \(\alpha P(\text{Cavity=1} \land \text{Toothache})\) = \(\alpha < 0.012+0.108, 0.008+0.072> = \alpha <0.08, 0.12> = <0.08/0.2, 0.12/0.2> = <0.4, 0.6>

<table>
<thead>
<tr>
<th>Toothache</th>
<th>~Toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catch</td>
<td>~Catch</td>
</tr>
<tr>
<td>Cavity</td>
<td>0.108</td>
</tr>
<tr>
<td>~Cavity</td>
<td>0.016</td>
</tr>
</tbody>
</table>
Normalization once for all?

- If you use the proper inference rules, then normalization will be preserved
- For the last example, we may use product rule
  - \( P(\text{Toothache} | Cavity=1) = \alpha P(\text{Cavity}=1 \ ^\land \ \text{Toothache}) \)
  - \( P(\text{Toothache} | Cavity=1) = P(\text{Cavity}=1^\land\text{Toothache}) / P(\text{Cavity}=1) \)
  - The results will be the same
  - Try it
Introducing a variable as an extra condition:

\[ P(X|Y) = \sum_z P(X|Y, Z = z) P(Z = z|Y) \]

Intuition: often easier to assess each specific circumstance, e.g.,
\[ P(\text{RunOver}|\text{Cross}) \]
\[ = P(\text{RunOver}|\text{Cross}, \text{Light} = \text{green}) P(\text{Light} = \text{green}|\text{Cross}) \]
\[ + P(\text{RunOver}|\text{Cross}, \text{Light} = \text{yellow}) P(\text{Light} = \text{yellow}|\text{Cross}) \]
\[ + P(\text{RunOver}|\text{Cross}, \text{Light} = \text{red}) P(\text{Light} = \text{red}|\text{Cross}) \]

When \( Y \) is absent, we have summing out or marginalization:

\[ P(X) = \sum_z P(X|Z = z) P(Z = z) = \sum_z P(X, Z = z) \]

In general, given a joint distribution over a set of variables, the distribution over any subset (called a marginal distribution for historical reasons) can be calculated by summing out the other variables.
Marginalization Example

<table>
<thead>
<tr>
<th></th>
<th>Toothache</th>
<th>~Toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catch</td>
<td>0.108</td>
<td>0.072</td>
</tr>
<tr>
<td>~Catch</td>
<td>0.012</td>
<td>0.008</td>
</tr>
<tr>
<td>Cavity</td>
<td>0.016</td>
<td>0.144</td>
</tr>
<tr>
<td>~Cavity</td>
<td>0.064</td>
<td>0.576</td>
</tr>
</tbody>
</table>

\[
P( \text{Toothache} \mid \text{Cavity} ) = \frac{P(\text{Toothache} \mid \text{Cavity, Catch}=1)P(\text{Catch}=1 \mid \text{Cavity})}{\# \text{ possible worlds} = 2} \\
+ P(\text{Toothache} \mid \text{Cavity, Catch}=0)P(\text{Catch}=0 \mid \text{Cavity})
\]

\[
P(\text{Toothache}) = \frac{P(\text{Toothache} \mid \text{Cavity}=1, \text{Catch}=0)}{\# \text{ possible worlds} = 2} \\
+ P(\text{Toothache} \mid \text{Cavity}=0, \text{Catch}=0) \\
+ P(\text{Toothache} \mid \text{Cavity}=1, \text{Catch}=1) \\
+ P(\text{Toothache} \mid \text{Cavity}=0, \text{Catch}=1)
Inference from joint distributions

Typically, we are interested in the posterior joint distribution of the query variables $Y$ given specific values $e$ for the evidence variables $E$

(Let $X$ be all variables)

Let the hidden variables be $H = X - Y - E$

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

$$P(Y|E = e) = \alpha P(Y, E = e) = \alpha \sum_h P(Y, E = e, H = h)$$

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

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???
Inference for any Proposition

- $0 \leq P(A) \leq 1$
- $P(\text{False}) = 0; P(\text{True}) = 1$
- $P(A \land B) = P(AB) = P(A)P(B|A) = P(B)P(A|B)$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$
Inference probability for sentence

1) For any proposition \( \phi \) defined on the random variables \( \phi(w_i) \) is true or false

2) \( \phi \) is equivalent to the disjunction of \( w_i \)'s where \( \phi(w_i) \) is true

Hence \( P(\phi) = \sum_{w_i: \phi(w_i)} P(w_i) \)

i.e., the unconditional probability of any proposition is computable as the sum of entries from the full joint distribution

Conditional probabilities can be computed in the same way as a ratio:

\[
P(\phi|\xi) = \frac{P(\phi \land \xi)}{P(\xi)}
\]

E.g.,

\[
P(Cavity|Toothache) = \frac{P(Cavity \land Toothache)}{P(Toothache)} = \frac{0.04}{0.04 + 0.01} = 0.8
\]
Inference Example

Assume the full joint distribution (like a truth table!)

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>~toothache</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>catch</td>
<td>~catch</td>
<td>catch</td>
<td>~catch</td>
</tr>
<tr>
<td>cavity</td>
<td>0.108</td>
<td>0.012</td>
<td>0.072</td>
<td>0.008</td>
</tr>
<tr>
<td>~cavity</td>
<td>0.016</td>
<td>0.064</td>
<td>0.144</td>
<td>0.576</td>
</tr>
</tbody>
</table>

Y=Cavity: we want to know whether we have a cavity
E=Toothache: we know we have a toothache
H=Catch: we don’t know whether probe would catch or not

\[ P(Y \mid e) = \alpha \quad P(Y, e) = \alpha \sum_h P(Y, e, h) \]
Inference of Probability vs Logic

• \( P(AB | C) = P(A | C)P(B | AC) = P(B | C)P(A | BC) \)

• Probability can do much more than logics
  – Deductive reasoning:
    • If \( A \rightarrow B \) and \( A \), then \( B \)
    • If \( A \rightarrow B \) and \( \sim B \), then \( \sim A \)
    • If \( A \rightarrow B \) and \( B \), then “\( A \) become more plausible”
  – Inductive reasoning:
    • If \( A \rightarrow B \) and \( \sim A \), then “\( B \) become less plausible”
    • If \( A \rightarrow ”B \) becomes more plausible” and \( B \), then “\( A \) become more plausible”
Why Can Probability Do More?

• Can you prove the following using probability?
  1. If A→B and A, then B
  2. If A→B and ~B, then ~A
  3. If A→B and B, then “A become more plausible”
  4. If A→B and ~A, then “B become less plausible”
  5. If A→”B becomes more plausible” and B, then “A become more plausible”

• These proofs are your home works

As an example of the last inference rule, let C stands for the background knowledge. Then the premise on the left-hand side takes the form \( P(B|AC) > P(B|C) \), and the Bayesian theorem tells us immediately \( P(A|BC') > P(A|C) \).
Inference in a Wumpus world

A = Agent
B = Breeze
S = Smell
P = Pit
W = Wumpus
OK = Safe
V = Visited
G = Glitter

Which one has a PIT?
Inferences In the world Wumpus

• Logic can only guess randomly which of [1,3], [2,2],[3,1] has a pit
• Using probability you can calculate which one is more likely have a pit than others
Details in Wumpus World

• Any of [1,3], [2,2] or [3,1] may have a pit, but are any riskier/safer than others to try if pits are relatively rare?
  — Either [1,3] or [3,1] should be less risky than [2,2] because most probable pattern given the evidence is one pit at [2,2]

• Need a probabilistic rather than a logical model

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

\[
B_{ij} = \text{true} \text{ iff } [i,j] \text{ is breezy}
\]

}\text{Boolean random variables}

For simplicity, will include only \(B_{12}, B_{21}, P_{11}, \ldots, P_{44}\) in probability model
Construct a Probability Model (for the Wumpus world)

• Construct the probability model
  – Select your random (binary) variables: $B_{12}, B_{21}, P_{11}, \ldots, P_{44}$
    • (see the last slide)
  – Construct the full joint probability distribution (exclusive, exhaustive, sum=1)

\[ P(B_{12}, B_{21}, P_{11}, \ldots, P_{44}) \]

• Use the model to compute the probabilities in interest
  – $P(P_{11}, \ldots, P_{44})$ is the probability of pit distribution (uniform)
    • Assume that pits are placed randomly with $P(p_{ij})=.2$ for all $i,j$
    • Assume that pits are placed independently

\[ P(P_{11}, \ldots, P_{44}) = P(P_{11}) \cdot P(P_{12}) \cdot \ldots \cdot P(P_{44}) \]

  – $P(B_{12}, B_{21} \mid P_{11}, \ldots, P_{44})$ is the probability of breeze at [1,2] [21]. We can compute this using the product rule

\[ P(B_{12}, B_{21} \mid P_{11}, \ldots, P_{44}) = \frac{P(B_{12}, B_{21}, P_{11}, \ldots, P_{44})}{P(P_{11}, \ldots, P_{44})} \]
Observations and Query

- Have observed these facts:
  \[ b = b_{12} \land b_{21} \]
  \[ \text{known} = \neg p_{11} \land \neg p_{12} \land \neg p_{21} \]
- Query is \( P(P_{13} \mid \text{known}, b) \)
  - Would also want to query \( P_{22} \) and \( P_{31} \)
- Define unknown = set of \( P_{ij} \)s other than \( P_{13} \) & known
- For inference by enumeration, we have
  \[ P(P_{13} \mid \text{known}, b) = \alpha \sum_{\text{unknown}} P(P_{13}, \text{unknown}, \text{known}, b) \]
  
  *Grows exponentially with number of squares*
Conditional Independence to the Rescue

• Basic Insight
  – If partition unknown variables into fringe and other, then observed breezes are conditionally independent of other variables, given known, query and fringe variables

• Reformulate query into usable form
  – If Unknown = Fringe or Other, then
    \[
    P(b \mid P_{13}, \text{Known,Unknown}) = P(b \mid P_{13}, \text{Known,Fringe})
    \]
  – Use to convert
    \[
    P(P_{13} \mid \text{known,b}) = \alpha \Sigma_{\text{unknown}} P(P_{13}, \text{unknown,known,b})
    \]
    Into (via sequence of transformations)
    \[
    \alpha' P(P_{13}) \Sigma_{\text{fringe}} P(b \mid \text{known,P}_{13},\text{fringe}) P(\text{fringe})
    \]
    [where \(\alpha' = \alpha P(\text{known})\)]
  – Greatly reduces computation
Results

\[ P(P_{13} \mid \text{known}, b) = \alpha' \cdot P(P_{13}) \cdot \sum_{\text{fringe}} P(b \mid \text{known}, P_{13}, \text{fringe}) \cdot P(\text{fringe}) \]

- \( P(b \mid \text{known}, P_{13}, \text{fringe}) \) is 0 or 1 depending on if fringe is consistent with b
- \( P(\text{fringe}) \) is shown in figure for situations in which fringe is consistent with b
- Now need to partition action events according to whether \( P_{13} \) is true or false
  \[ = \alpha' [0.2(0.04 + 0.16 + 0.16), 0.8(0.04 + 0.16)] = [0.31, 0.69] \]

\[ P(P_{22} \mid \text{known}, b) = [0.86, 0.14] \]

\( P_{13} = \text{true} \)

\( P_{13} = \text{false} \)
Alternative: Using Bayesian Rule!

• Can compute it nicely by comparing:
  - $P(P_{13} \mid \text{known}, b)$
  - $P(P_{31} \mid \text{known}, b)$
  - $P(P_{22} \mid \text{known}, b)$

• Let’s use Bayesian Rule
  - $P(P_{13} \mid \text{known}, b) = P(P_{13})P(\text{known}, b \mid P_{13})/P(\text{known}, b)$
  - $P(P_{22} \mid \text{known}, b) = P(P_{22})P(\text{known}, b \mid P_{22})/P(\text{known}, b)$
  - $P(P_{31} \mid \text{known}, b) = P(P_{31})P(\text{known}, b \mid P_{31})/P(\text{known}, b)$

• Since $P(P_{13}) = P(P_{22}) = P(P_{31})$, only need compare
  - $P(\text{known}, b \mid P_{13})$
  - $P(\text{known}, b \mid P_{22})$  ← this is larger than the other two, why?
  - $P(\text{known}, b \mid P_{31})$
Compare the three choices

• Using chain rules
  – \( P(\text{known}, b | p_{13}) = P(\text{known} | b, p_{13})P(b | p_{13}) \)
  – \( P(\text{known}, b | p_{22}) = P(\text{known} | b, p_{22})P(b | p_{22}) \)
  – \( P(\text{known}, b | p_{31}) = P(\text{known} | b, p_{31})P(b | p_{31}) \)

• The first terms are equal
  – because pits are independent from each other

• The second term:
  – \( P(b_{12} \land b_{21} | p_{13}) < 1 \)
  – \( P(b_{12} \land b_{21} | p_{22}) = 1 \)
  – \( P(b_{12} \land b_{21} | p_{13}) < 1 \)

  – Where \( \text{known} = \sim p_{11} \land \sim p_{12} \land \sim p_{21} \), and \( b = b_{12} \land b_{21} \)
Summary

• Probability is a rigorous formalism for uncertain knowledge
  – Conditional probabilities enable reasoning with uncertain evidence

• A full joint probability distribution specifies the probability of every atomic event
  – Queries answerable by summing over probabilities of atomic events

• Bayes’ theorem/rule provides basis for most modern diagnostic reasoning in AI
  – Converts uncertain causal information into diagnostic conclusions

• For nontrivial domains, we must find a way to reduce the size of the joint distribution
  – (Conditional) independence provides the tools