Uncertainty
- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule
- 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 ... )
- 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:
  - $A_{25} \rightarrow 0.3 \text{ AtAirportOnTime}$
  - $\text{Sprinkler} \rightarrow 0.99 \text{ WetGrass}$
  - $\text{WetGrass} \rightarrow 0.7 \text{ Rain}$

- 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)
- 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., \( P(A_{25} \mid \text{no reported accidents}) = 0.06 \)
- 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., \( P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15 \)
Suppose I believe the following:
- $P(A_{25} \text{ gets me there on time | ...}) = 0.04$
- $P(A_{90} \text{ gets me there on time | ...}) = 0.70$
- $P(A_{120} \text{ gets me there on time | ...}) = 0.95$
- $P(A_{1440} \text{ gets me there on time | ...}) = 0.9999$

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

Making Decisions Under Uncertainty
Begin with a set \( \Omega \) - 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
  
  - \( 0 \leq P(\omega) \leq 1 \)
  - \( \sum_{\omega} P(\omega) = 1 \)

  e.g.,
  
  \[ P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6} \]

An event \( A \) is any subset of \( \Omega \)
  
  - \( P(A) = \sum_{\omega \in A} P(\omega) \)

  E.g., \( P(\text{die roll} < 4) = P(1) + P(2) + P(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \)
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:
  - \( P(X = x_i) = \sum_{\omega : X(\omega) = x_i} P(\omega) \)
  - e.g., \( P(\text{Odd}=\text{true}) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \)
Think of a proposition as the event (set of sample points) where the proposition is true

Given Boolean random variables A and B:

- event $a = \text{set of sample points where } A(\omega)=true$
- event $\neg a = \text{set of sample points where } A(\omega)=false$
- event $a \land b = \text{points where } A(\omega)=true \text{ and } B(\omega)=true$

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., $(a \lor b) \equiv (\neg a \land b) \lor (a \land \neg b) \lor (a \land b) \Rightarrow P(a \lor b) = P(\neg a \land b) + P(a \land \neg b) + P(a \land b)$
The definitions imply that certain logically related events must have related probabilities.

E.g., \( P(a \lor b) = P(a) + P(b) - P(a \land b) \)

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.
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 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:

<table>
<thead>
<tr>
<th>Weather</th>
<th>sunny</th>
<th>rain</th>
<th>cloudy</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = true</td>
<td>0.144</td>
<td>0.02</td>
<td>0.016</td>
<td>0.02</td>
</tr>
<tr>
<td>Cavity = false</td>
<td>0.576</td>
<td>0.08</td>
<td>0.064</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points.
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

\[
\lim_{dx \to 0} \frac{P(20.5 \leq X \leq 20.5 + dx)}{dx} = 0.125
\]
\[ P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

Gaussian Density
Conditional or posterior probabilities
- e.g., \( P(\text{cavity}|\text{toothache}) = 0.8 \)
- i.e., given that toothache is all I know
- NOT “if toothache then 80% chance of cavity”

Notation for conditional distributions:
- \( P(\text{Cavity}|\text{Toothache}) = 2\)-element vector of 2-element vectors

If we know more, e.g., cavity is also given, then we have
- \( P(\text{cavity}|\text{toothache, cavity}) = 1 \)

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.,
- \( P(\text{cavity}|\text{toothache, 49ersWin}) = P(\text{cavity}|\text{toothache}) = 0.8 \)

This kind of inference, sanctioned by domain knowledge, is crucial
Definition of conditional probability:

\[ P(a|b) = \frac{P(a \land b)}{P(b)} \quad \text{if } P(b) \neq 0 \]

Product rule gives an alternative formulation:

\[ P(a \land b) = P(a|b)P(b) = P(b|a)P(a) \]

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

\[ P(\text{Weather, Cavity}) = P(\text{Weather} | \text{Cavity})P(\text{Cavity}) \]

(View as a 4 x 2 set of equations, not matrix multiplication)

Chain rule is derived by successive application of product rule:

\[
P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1}) P(X_n | X_1, \ldots, X_{n-1}) \\
= P(X_1, \ldots, X_{n-2}) P(X_{n-1} | X_1, \ldots, X_{n-2}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \ldots \\
= \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\]
Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>( \neg \text{toothache} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>0.108</td>
<td>0.072</td>
</tr>
<tr>
<td>( \neg \text{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>( \neg \text{cavity} )</td>
<td>0.064</td>
<td>0.576</td>
</tr>
</tbody>
</table>

Inference by Enumeration

For any proposition \( \phi \), sum the atomic events where it is true:

\[
P(\phi) = \Sigma_{\omega: \omega \models \phi} P(\omega)
\]
• Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>¬catch</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>cavity</td>
<td>.016</td>
<td>.064</td>
</tr>
<tr>
<td>¬cavity</td>
<td>.144</td>
<td>.576</td>
</tr>
</tbody>
</table>

Inference by Enumeration

• For any proposition $\phi$, sum the atomic events where it is true:

\[ P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega) \]

\[ P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2 \]
Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>(\neg) toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\neg) cavity</td>
<td>catch</td>
<td>(\neg) catch</td>
</tr>
<tr>
<td>cavity</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>(\neg) cavity</td>
<td>.016</td>
<td>.064</td>
</tr>
</tbody>
</table>

Inference by Enumeration

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

\[
P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28
\]
Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>¬ catch</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>cavity</td>
<td>.016</td>
<td>.064</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>.144</td>
<td>.576</td>
</tr>
</tbody>
</table>

Inference by Enumeration

Can also compute conditional probabilities

\[
P(\neg cavity | toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)} = \frac{\frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064}}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]
<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th></th>
<th>¬ toothache</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>

**Normalization**

- Denominator can be viewed as a normalization constant

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

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

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

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

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???
- A and B are independent iff
- \( P(A|B) = P(A) \) or \( P(B|A) = P(B) \) or \( P(A, B) = P(A)P(B) \)
  - \( P(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather}) = P(\text{Toothache}, \text{Catch}, \text{Cavity})P(\text{Weather}) \)
- 32 entries reduced to 12; for \( n \) independent biased coins, \( 2^n \rightarrow n \)
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?
\[ P(\text{Toothache}, \text{Cavity}, \text{Catch}) \text{ has } 2^3 - 1 = 7 \text{ independent entries} \]

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

1. \[ P(\text{catch} | \text{toothache, cavity}) = P(\text{catch} | \text{cavity}) \]

The same independence holds if I haven't got a cavity:

2. \[ P(\text{catch} | \text{toothache, \neg cavity}) = P(\text{catch} | \neg \text{cavity}) \]

Catch is conditionally independent of Toothache given Cavity:

\[ P(\text{Catch} | \text{Toothache, Cavity}) = P(\text{Catch} | \text{Cavity}) \]

Equivalent statements:

\[ P(\text{Toothache} | \text{Catch, Cavity}) = P(\text{Toothache} | \text{Cavity}) \]
\[ P(\text{Toothache, Catch | Cavity}) = P(\text{Toothache | Cavity})P(\text{Catch | Cavity}) \]
- Write out full joint distribution using chain rule:
  - \( P(\text{Toothache}, \text{Catch}, \text{Cavity}) = \)
  - \( P(\text{Toothache} | \text{Catch}, \text{Cavity})P(\text{Catch}, \text{Cavity}) = \)
  - \( P(\text{Toothache} | \text{Catch}, \text{Cavity})P(\text{Catch} | \text{Cavity})P(\text{Cavity}) = \)
  - \( P(\text{Toothache} | \text{Cavity})P(\text{Catch} | \text{Cavity})P(\text{Cavity}) \)

- 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  \( P(a \land b) = P(a|b)P(b) = P(b|a)P(a) \)

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

or in distribution form

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha P(X|Y)P(Y) \]

Useful for assessing diagnostic probability from causal probability:

\[ P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})} \]

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

\[ P(\text{Cavity}|\text{toothache} \land \text{catch}) \]
\[ = \alpha P(\text{toothache} \land \text{catch}|\text{Cavity})P(\text{Cavity}) \]
\[ = \alpha P(\text{toothache}|\text{Cavity})P(\text{catch}|\text{Cavity})P(\text{Cavity}) \]

This is an example of a naive Bayes model:

\[ P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i|\text{Cause}) \]

Total number of parameters is \textbf{linear} in \( n \)
### An Example

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temp</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>True</td>
<td>No</td>
</tr>
<tr>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Mild</td>
<td>High</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cool</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cool</td>
<td>Normal</td>
<td>True</td>
<td>No</td>
</tr>
<tr>
<td>Overcast</td>
<td>Cold</td>
<td>Normal</td>
<td>True</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>False</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Mild</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>True</td>
<td>Yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>True</td>
<td>Yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Mild</td>
<td>High</td>
<td>True</td>
<td>No</td>
</tr>
</tbody>
</table>
## Estimated Probabilities for Weather Data

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Overcast</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Rainy</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Sunny</td>
<td>2/9</td>
<td>3/5</td>
<td>2/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Overcast</td>
<td>4/9</td>
<td>0/5</td>
<td>4/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Rainy</td>
<td>3/9</td>
<td>2/5</td>
<td>3/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>
A new day:

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>2/9 3/5</td>
<td>2/9 2/5</td>
<td>False</td>
<td>9/5</td>
</tr>
<tr>
<td>Overcast</td>
<td>4/9 0/5</td>
<td>4/9 2/5</td>
<td>True</td>
<td>3/3</td>
</tr>
<tr>
<td>Rainy</td>
<td>3/9 2/5</td>
<td>3/9 1/5</td>
<td>True</td>
<td>3/3</td>
</tr>
</tbody>
</table>

Likelihood of the two classes

For “yes” = \( \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0053 \)

For “no” = \( \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0206 \)

Conversion into a probability by normalization:

\[
P(\text{"yes"}) = \frac{0.0053}{(0.0053 + 0.0206)} = 0.205
\]

\[
P(\text{"no"}) = \frac{0.0206}{(0.0053 + 0.0206)} = 0.795
\]
Probability of event $H$ given evidence $E$:

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

- *A priori* probability of $H$:
  
  - $Pr[H]$  
  
    - Probability of event before evidence is seen

- *A posteriori* probability of $H$:
  
  - $Pr[H | E]$  
  
    - Probability of event after evidence is seen

Bayes’s Rule
Classification learning: what’s the probability of the class given an instance?

- Evidence $E = \text{instance}$
- Event $H = \text{class value for instance}$

Naïve Bayes for Classification

Naïve assumption: evidence splits into parts (i.e. attributes) that are independent

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

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temp.</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Cool</td>
<td>High</td>
<td>True</td>
<td>?</td>
</tr>
</tbody>
</table>

\[
Pr[yes \mid E] = Pr[Outlook = Sunny \mid yes] \\
\times Pr[Temperature = Cool \mid yes] \\
\times Pr[Humidity = High \mid yes] \\
\times Pr[Windy = True \mid yes] \\
\times \frac{Pr[yes]}{Pr[E]} = \frac{2 \times 3 \times 3 \times 3 \times 9}{9 \times 9 \times 9 \times 14} \frac{Pr[E]}{Pr[E]}
\]
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)

The “Zero-Frequency Problem”

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

\[ Pr[yes \mid E] = 0 \]
Example: attribute *outlook* for class *yes*

\[
\frac{2 + \mu/3}{9 + \mu} \quad \frac{4 + \mu/3}{9 + \mu} \quad \frac{3 + \mu/3}{9 + \mu}
\]

*Sunny* \quad *Overcast* \quad *Rainy*

*Weights don’t need to be equal (but they must sum to 1)*

\[
\frac{2 + \mu p_1}{9 + \mu} \quad \frac{4 + \mu p_2}{9 + \mu} \quad \frac{3 + \mu p_3}{9 + \mu}
\]
Training: instance is not included in frequency count for attribute value-class combination

Classification: attribute will be omitted from calculation

Example:

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temp.</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>Cool</td>
<td>High</td>
<td>True</td>
<td>?</td>
</tr>
</tbody>
</table>

Likelihood of “yes” = \( \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0238 \)

Likelihood of “no” = \( \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0343 \)

\( P(\text{“yes”}) = \frac{0.0238}{0.0238 + 0.0343} = 41\% \)

\( P(\text{“no”}) = \frac{0.0343}{0.0238 + 0.0343} = 59\% \)
Wumpus World

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

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

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:

\[ P(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} P(P_{i,j}) = 0.2^n \times 0.8^{16-n} \]

for \( n \) pits.
We know the following facts:
\[ b = \neg b_{1,1} \land b_{1,2} \land b_{2,1} \]
\[ known = \neg p_{1,1} \land \neg p_{1,2} \land \neg p_{2,1} \]
Query is \( P(P_{1,3}|known, b) \)
Define \( Unknown = P_{ij} \)s other than \( P_{1,3} \) and \( Known \)
For inference by enumeration, we have
\[
P(P_{1,3}|known, b) = \alpha \sum_{unknown} P(P_{1,3}, unknown, known, b)
\]
Grows exponentially with number of squares!
Using Conditional Independence

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

Define $\text{Unknown} = \text{Fringe} \cup \text{Other}$

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

Manipulate query into a form where we can use this!
Using Conditional Independence

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

\[
P(P_{2,2}|\text{known}, b) \approx \langle 0.86, 0.14 \rangle
\]
● 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