CS 561: Artificial Intelligence

Instructor: Sofus A. Macskassy, macskass@usc.edu
TAs: Nadeesha Ranasinghe (nadeeshr@usc.edu)
    William Yeoh (wyeoh@usc.edu)
    Harris Chiu (chiciu@usc.edu)

Lectures: MW 5:00-6:20pm, OHE 122 / DEN
Office hours: By appointment
Class page: http://www-rcf.usc.edu/~macskass/CS561-Spring2010/

This class will use http://www.uscden.net/ and class webpage
    - Up to date information
    - Lecture notes
    - Relevant dates, links, etc.

Course material:
Outline [AIMA Ch 13]

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule
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 (at 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
- 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)
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., \( P(A_{25} | \text{no reported accidents, 5 a.m.}) = 0.15 \)
- (Analogous to logical entailment status \( KB \models \alpha \), not truth.)
Making decisions under uncertainty

- 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
Probability basics

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 $\omega \in \Omega$ s.t.
$0 \leq P(\omega) \leq 1$
$\sum_{\omega} P(\omega) = 1$
e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event $A$ is any subset of $\Omega$

$P(A) = \sum_{\omega \in A} P(\omega)$
e.g., $P(\text{die roll } < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$
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 r.v. $X$:

$$P(X = x_i) = \sum_{\omega: X(\omega) = x_i} P(\omega)$$

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 $\neg a =$ 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., $(a \lor b) \equiv (\neg a \land b) \lor (a \land \neg b) \lor (a \land b)$

\[ P(a \lor b) = P(\neg a \land b) + P(a \land \neg b) + P(a \land b) \]
Why use probability?

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

- 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.
Syntax for propositions

- Propositional or Boolean random variables
  e.g., $\text{Cavity}$ (do I have a cavity?)
  $\text{Cavity} = \text{true}$ is a proposition, also written $\text{cavity}$

- Discrete random variables (finite or infinite)
  e.g., $\text{Weather}$ is one of $\langle \text{sunny}, \text{rain}, \text{cloudy}, \text{snow} \rangle$
  $\text{Weather} = \text{rain}$ is a proposition
  Values must be exhaustive and mutually exclusive

- Continuous random variables (bounded or unbounded)
  e.g., $\text{Temp} = 21.6$; also allow, e.g., $\text{Temp} < 22.0$. 

- Arbitrary Boolean combinations of basic propositions
Prior probability

- Prior or unconditional probabilities of propositions
e.g., \( P(\text{Cavity} = \text{true}) = 0.1 \) and \( P(\text{Weather} = \text{sunny}) = 0.72 \) correspond to belief prior to arrival of any (new) evidence.

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

- Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)
  \( P(\text{Weather}, \text{Cavity}) \) = a 4 \times 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.
Probability for continues variables

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

- Here \( P \) is a density; integrates to 1.
- \( P(X = 20.5) = 0.125 \) really means
  \[ \lim_{dx \to 0} \frac{P(20.5 \leq X \leq 20.5 + dx)}{dx} = 0.125 \]
Gaussian density

\[ P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Conditional probability

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

- (Notation for conditional distributions:
  \( P(\text{Cavity}|\text{Toothache}) = \) 2-element vector of 2-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.,
  \( P(\text{cavity}|\text{toothache}, \text{49ersWin}) = P(\text{cavity}|\text{toothache}) = 0.8 \)

- This kind of inference, sanctioned by domain knowledge, is crucial
Conditional Probability

Definition of conditional probability:

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

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 \(4 \times 2\) set of equations, not matrix mult.)

Chain rule is derived by successive application of product rule:

\[ P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1}) \ P(X_n|X_1, \ldots, X_{n-1}) \]
\[ = P(X_1, \ldots, X_{n-2}) \ P(X_{n-1}|X_1, \ldots, X_{n-2}) \ P(X_n|X_1, \ldots, X_{n-1}) \]
\[ = \ldots \]
\[ = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \]
Inference by enumeration

Start with the joint distribution:

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

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

$$P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega)$$
Inference by enumeration

Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>\neg toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>0.108</td>
<td>0.072</td>
</tr>
<tr>
<td>\neg cavity</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 cavity</td>
<td>0.064</td>
<td>0.576</td>
</tr>
</tbody>
</table>

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

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

\[
P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
\]
Inference by enumeration

Start with the joint distribution:

<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>0.072</td>
<td>0.008</td>
</tr>
<tr>
<td>cavity</td>
<td>0.016</td>
<td>0.064</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>0.144</td>
<td>0.576</td>
</tr>
</tbody>
</table>

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

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

$$P(\text{cavity} \lor \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$
### Inference by enumeration

Start with the joint distribution:

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

Can also compute conditional probabilities:

\[
P(\neg \text{cavity}|\text{toothache}) = \frac{P(\neg \text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]
Normalization

Start with the joint distribution:

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

Denominator can be viewed as a normalization constant $\alpha$

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

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables
Inference by enumeration, contd.

- 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:
  \[
P(Y \mid E = e) = \alpha \sum_h P(Y, E = e, H = h)
  \]

- 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???
Independence

- $A$ and $B$ are independent iff
- $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A,B) = P(A)P(B)$

\[
P(\text{Toothache; Catch; Cavity; Weather}) = P(\text{Toothache; Catch; Cavity}) \times P(\text{Weather})
\]

- 32 entries reduced to 12; for $n$ independent biased counts, $2^n \rightarrow n$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?
Conditional Independence

- $P(\text{Toothache}, \text{Cavity}, \text{Catch})$ 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:
  \[ P(\text{catch}|\text{toothache, cavity}) = P(\text{catch}|\text{cavity}) \]
  \[ P(\text{catch}|\text{toothache, :cavity}) = P(\text{catch}|\text{cavity}) \]
- Catch is conditionally independent of Toothache given Cavity:
  \[ P(\text{Catch}|\text{Toothache, Cavity}) = P(\text{Catch}|\text{Cavity}) \]
- Equivalent statements:
  \[ P(\text{Toothache}|\text{Catch, Cavity}) = P(\text{Toothache}|\text{Cavity}) \]
  \[ P(\text{Toothache, Catch}|\text{Cavity}) = P(\text{Toothache}|\text{Cavity})P(\text{Catch}|\text{Cavity}) \]
Conditional independence contd.

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

$\Rightarrow$ 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:

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

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

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

- This is an example of a naive Bayes model:

\[
P(Cause, Effect_1, \ldots, Effect_n) = P(Cause) \prod_i P(Effect_i | Cause)
\]

Total number of parameters linear in \(n\)
Wumpus World

- \( P_{ij} = \text{true} \) iff \([i, j]\) contains a pit
- \( B_{ij} = \text{true} \) iff \([i, j]\) is breezy
- Include only \( B_{1,1}, B_{1,2}, B_{2,1} \) in the probability model
Specifying the probability model

- The full joint distribution is \( P(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1}) \)
- Apply product rule: \( P(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4}) P(P_{1,1}, \ldots, P_{4,4}) \)
- (Do it this way to get \( P(\text{Effect} \mid \text{Cause}) \).)
- 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.
Observations and query

- 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 neighboring hidden squares.

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

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

- Manipulate query into a form where we can use this!
Using conditional independence contd.

\[
P(P_{1,3}|\text{known}, b) = \alpha \sum_{\text{unknown}} P(P_{1,3}, \text{unknown}, \text{known}, b) \\
= \alpha \sum_{\text{unknown}} P(b|P_{1,3}, \text{known}, \text{unknown})P(P_{1,3}, \text{known}, \text{unknown}) \\
= \alpha \sum_{\text{fringe other}} P(b|\text{known}, P_{1,3}, \text{fringe}, \text{other})P(P_{1,3}, \text{known}, \text{fringe}, \text{other}) \\
= \alpha \sum_{\text{fringe other}} P(b|\text{known}, P_{1,3}, \text{fringe})P(P_{1,3}, \text{known}, \text{fringe}, \text{other}) \\
= \alpha \sum_{\text{fringe}} P(b|\text{known}, P_{1,3}, \text{fringe}) \sum_{\text{other}} P(P_{1,3}, \text{known}, \text{fringe}, \text{other}) \\
= \alpha \sum_{\text{fringe}} P(b|\text{known}, P_{1,3}, \text{fringe}) \sum_{\text{other}} P(P_{1,3})P(\text{known})P(\text{fringe})P(\text{other}) \\
= \alpha P(\text{known})P(P_{1,3}) \sum_{\text{fringe}} P(b|\text{known}, P_{1,3}, \text{fringe})P(\text{fringe}) \sum_{\text{other}} P(\text{other}) \\
= \alpha' P(P_{1,3}) \sum_{\text{fringe}} P(b|\text{known}, P_{1,3}, \text{fringe})P(\text{fringe})
Using conditional independence contd.

- $P(P_{1,3}|\text{known, } b) = 0.2(0.04+0.16+0.16), 0.8(0.04+0.16) 
  \approx 0.31, 0.69$

- $P(P_{2,2}|\text{known, } b) \approx 0.86, 0.14$
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