Probabilistic Reasoning

Doug Downey, Northwestern EECS 348
Spring 2010
Limitations of logic-based agents

• Qualification Problem
  – Action’s preconditions can be complex
  – Action(Grab, t) => Holding(t)
    ....unless gold is slippery or nailed down or too heavy or our hands are full or...

• Brittleness
  – One contradiction in KB => KB entails everything
Limitations of logic-based agents

• Qualification Problem
  – Action’s preconditions can be complex
  – Action(Grab, t) => Holding(t)
    ....unless gold is slippery or nailed down or too heavy or our hands are full or...

  \[ P(\text{success}) = 0.97 \]

• Brittleness
  – One contradiction in KB => KB entails everything

  Instead of \( a \land \neg a \),
  \[ P(a) + P(\neg a) = 1 \]
Events

• Event space $\Omega$
  - E.g. for dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$

• Set of measurable events $S \subseteq 2^\Omega$
  - E.g.,
    $\alpha = \text{event we roll an even number} = \{2, 4, 6\} \in S$
  - $S$ must:
    • Contain the empty event $\emptyset$ and the trivial event $\Omega$
    • Be closed under union & complement

  - $\alpha, \beta \in S \rightarrow \alpha \cup \beta \in S$ and $\alpha \in S \rightarrow \Omega - \alpha \in S$
Probability Distributions

Can visualize probability as fraction of area
A probability distribution $P$ over $(\Omega, \mathcal{S})$ is a mapping from $\mathcal{S}$ to real values such that:

- $P(\alpha) \geq 0$
- $P(\Omega) = 1$
- $\alpha, \beta \in \mathcal{S} \land \alpha \cap \beta = \emptyset \rightarrow P(\alpha \cup \beta) = P(\alpha) + P(\beta)$
Probability: Interpretations & Motivation

• Interpretations
  – Frequentist
  – Bayesian/subjective

• Why use probability for subjective beliefs?
  – Beliefs that violate the axioms can lead to bad decisions *regardless* of the outcome [de Finetti, 1931]
  – Example: \( P(A) = 0.6, \ P(\text{not } A) = 0.8 \) ?
  – Example: \( P(A) > P(B) \) and \( P(B) > P(A) \) ?
Random Variables

• A random variable is a function from $\Omega$ to a value
  – A short-hand for referring to attributes of events.
• E.g., your grade in this course
  – Let $\Omega =$ set of possible scores on hmwks and test
  – Cumbersome to have separate events GradeA, GradeB, GradeC
  – So instead define a random variable Grade
    • Deterministic function from $\Omega$ to $\{A, B, C\}$
Distributions

- Called “marginal” because they apply to only one r.v.
Joint Distribution

\[ P(\text{Intelligence, Grade}) \]

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence=Low</th>
<th>Intelligence=High</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.07</td>
<td>0.18</td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>0.18</td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Joint Distribution specified with $2 \times 3 - 1 = 5$ values
### Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
</tr>
</tbody>
</table>

P(Grade = A, Intelligence = Low) = 0.07
Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>0.07</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>0.18</td>
<td>0.09</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0.35</td>
<td>0.03</td>
</tr>
</tbody>
</table>

P(Grade = A)? 0.07 + 0.18 = 0.25
# Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

\[
P(\text{Grade} = A \lor \text{Intelligence} = \text{High}) = 0.07 + 0.18 + 0.09 + 0.03 = 0.37
\]

=> Given the joint distribution, we can compute probabilities for any proposition by summing events.
Conditional Probability

• $P(Grade = A \mid Intelligence = High) = 0.6$
  
  – the probability of getting an A given only $Intelligence = High$, and nothing else.

  • If we know $Motivation = High$ or $OtherInterests = Many$, the probability of an A changes even given high $Intelligence$

• Formal Definition:
  
  – $P(\alpha \mid \beta) = P(\alpha, \beta) / P(\beta)$

  • When $P(\beta) > 0$
Conditional Probability

P(Grade = A | Intelligence = High) ?

P(Grade = A, Intelligence = High) = 0.18
P(Intelligence = High) = 0.18 + 0.09 + 0.03 = 0.30
=> P(Grade = A | Intelligence = High) = 0.18 / 0.30 = 0.6
P(Intelligence | Grade = A)?

### Conditional Probability

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
<td>0.18</td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Conditional Probability

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td>High</td>
</tr>
<tr>
<td>A</td>
<td>0.28</td>
<td></td>
<td>0.72</td>
</tr>
<tr>
<td>B</td>
<td>0.76</td>
<td></td>
<td>0.24</td>
</tr>
<tr>
<td>C</td>
<td>0.92</td>
<td></td>
<td>0.08</td>
</tr>
</tbody>
</table>

\[P(\text{Intelligence} \mid \text{Grade})?\]

Actually three separate distributions, one for each \textit{Grade} value
(has three independent parameters total)
Chain Rule

\[
P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1)
\]

- E.g., \(P(\text{Grade}=B, \text{Int.} = \text{High})\)
  \[= P(\text{Grade}=B | \text{Int.}=\text{High})P(\text{Int.} = \text{High})\]
- Can be used for distributions...
  - \(P(A, B) = P(A | B)P(B)\)
Queries

• Given subsets of random variables $Y$ and $E$, and assignments $e$ to $E$
  – Find $P(Y \mid E = e)$

• Answering queries = inference
  – The whole point of probabilistic models, more or less
  – $P(Disease \mid Symptoms)$
  – $P(StockMarketCrash \mid RecentPriceActivity)$
  – $P(CodingRegion \mid DNASequence)$
  – $P(PlayTennis \mid Weather)$
  – ...(the other key task is learning)
### Answering Queries: Summing Out

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence = Low</th>
<th>Intelligence = High</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time = Lots</td>
<td>Time = Little</td>
</tr>
<tr>
<td></td>
<td>Time = Lots</td>
<td>Time = Little</td>
</tr>
<tr>
<td>A</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>B</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>C</td>
<td>0.10</td>
<td>0.25</td>
</tr>
</tbody>
</table>

\[
P(\text{Grade} | \text{Time} = \text{Lots}) = \sum_{v \in \text{Val}(\text{Intelligence})} P(\text{Grade}, \text{Intelligence} = v | \text{Time} = \text{Lots})
\]
Answering Queries: Solved?

• Given the joint distribution, we can answer any query by summing

• ...but, joint distribution of 500 Boolean variables has $2^{500} - 1$ parameters (about $10^{150}$)

• For non-trivial problems (~25 boolean r.v.s or more), using the joint distribution requires
  – Way too much **computation** to compute the sum
  – Way too many **observations** to learn the parameters
  – Way too much **space** to store the joint distribution
Conditional Independence (1 of 3)

- Independence
  - $P(A, B) = P(A) \times P(B)$, denoted $A \perp B$
  - E.g. consecutive dice rolls
    - Gambler’s fallacy
  - Rare in (real) applications

Note: Book calls this “marginal independence” when applied to r.v.s, but just “independence” when applied to events
Conditional Independence (2 of 3)

- Conditional Independence
  - $P(A, B | C) = P(A | C) P(B | C)$, denoted $(A \perp B | C)$
  - Much more common
  - E.g.,
    - $(GetIntoNU \perp GetIntoStanford | Application)$,
    - but NOT $(GetIntoNU \perp GetIntoStanford)$
Conditional Independence (3 of 3)

• How does Conditional Independence save the day?

\[ P(NU, Stanford, App) = \]
\[ P(NU | Stanford, App) \times P(Stanford | App) \times P(App) \]

Now, \((A \perp B | C)\) means \(P(A | B, C) = P(A | C)\)

So since \((NU \perp Stanford | App)\), we have

\[ P(NU, Stanford, App) = \]
\[ P(NU | App) \times P(Stanford | App) \times P(App) \]

Say \(App \in \{\text{Good, Bad}\}\) and \(School \in \{\text{Yes, No, Wait}\}\)

All we need is 4+4+1=9 numbers

(vs. \(3 \times 3 \times 2 - 1 = 17\) for the full joint)

• Full joint has size \textbf{exponential} in \# of r.v.s

Conditional independence eliminates this!
Bayes’ Rule

- $P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$

- Example:
  - $P(\text{symptom} \mid \text{disease}) = 0.95$, $P(\text{symptom} \mid \neg\text{disease}) = 0.05$
  - $P(\text{disease} = 0.0001)$

  $P(\text{disease} \mid \text{symptom})$
  
  $$= \frac{P(\text{symptom} \mid \text{disease}) \cdot P(\text{disease})}{P(\text{symptom})}$$

  $$= \frac{0.95 \times 0.0001}{0.95 \times 0.0001 + 0.05 \times 0.9999} = 0.002$$
What have we learned?

- Probability – a calculus for dealing with uncertainty
  - Built from small set of axioms (ignore at your peril)
- Joint Distribution $P(A, B, C, ...)$
  - Specifies probability of all combinations of r.v.s
  - Intractable to compute exhaustively for non-trivial problems
- Conditional Probability $P(A \mid B)$
  - Specifies probability of A given B
- Conditional Independence
  - Can radically reduce number of variable combinations we must assign unique probabilities to.
- Bayes’ Rule