CS 188: Artificial Intelligence
Fall 2006

Lecture 14: Probability
10/17/2006

Dan Klein – UC Berkeley

Announcements

- Grades:
  - Check midterm, p1.1, and p1.2 grades in glookup
  - Let us know if there are problems, so we can calculate useful preliminary grade estimates
  - If we missed you, tell us your partner’s login

- Readers request:
  - List partners and logins on top of your readme files
  - Turn off your debug output

- Project 3.1 up: written probability problems, start now!
- Extra office hours: Thursday 2-3pm (if people use them)

Today

- Probability
  - Random Variables
  - Joint and Conditional Distributions
  - Bayes Rule
  - Independence

- You’ll need all this stuff for the next few weeks, so make sure you go over it!

Uncertainty

- General situation:
  - Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - Agent needs to reason about other aspects (e.g., where an object is or what disease is present)
  - Agent knows something about how the known variables relate to the unknown variables

- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

Random Variables

- A random variable is some aspect of the world about which we have uncertainty
  - R = is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?

- We denote random variables with capital letters

- Like in a CSP, each random variable has a domain
  - R in {true, false}
  - D in [0, ∞]
  - L in possible locations

Probabilities

- We generally calculate conditional probabilities
  - P(on time | no reported accidents) = 0.90

- Probabilities change with new evidence:
  - P(on time | no reported accidents, 5 a.m.) = 0.95
  - P(on time | no reported accidents, 5 a.m., raining) = 0.80
  - i.e., observing evidence causes beliefs to be updated
Probabilistic Models

- CSPs:
  - Variables with domains
  - Constraints: map from assignments to true/false
  - Ideally: only certain variables directly interact

- Probabilistic models:
  - (Random) variables with domains
  - Joint distributions: map from assignments (or outcomes) to positive numbers
  - Normalized: sum to 1.0
  - Ideally: only certain variables directly interact
  - Assignments are called outcomes

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>warm</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Distributions on Random Vars

- A joint distribution over a set of random variables: $X_1, X_2, \ldots, X_n$
is a map from assignments (or outcomes, or atomic events) to reals:
\[
P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]

\[
P(x_1, x_2, \ldots, x_n)
\]

- Size of distribution if $n$ variables with domain sizes $d$?
- Must obey:
  \[
  0 < P(x_1, x_2, \ldots, x_n) < 1
  \]
  \[
  \sum_{(x_1, x_2, \ldots, x_n)} P(x_1, x_2, \ldots, x_n) = 1
  \]
- For all but the smallest distributions, impractical to write out

Examples

- An event is a set $E$ of assignments (or outcomes)

\[
P(E) = \sum_{(x_1, x_2, \ldots, x_n) \in E} P(x_1, x_2, \ldots, x_n)
\]

- From a joint distribution, we can calculate the probability of any event

- Probability that it’s warm AND sunny?
- Probability that it’s warm?
- Probability that it’s warm OR sunny?

Conditional Probabilities

- A conditional probability is the probability of an event given another event (usually evidence)

\[
P(a | b) = \frac{P(a, b)}{P(b)}
\]

<table>
<thead>
<tr>
<th>T</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>warm</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Marginalization

- Marginalization (or summing out) is projecting a joint distribution to a sub-distribution over subset of variables

\[
P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)
\]

\[
P(T, S') = \sum_{t} P(t, s')
\]

<table>
<thead>
<tr>
<th>T</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>warm</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Conditional Probabilities

- Conditional or posterior probabilities:
  - E.g., $P(\text{cavity} | \text{toothache}) = 0.8$
  - Given that toothache is all I know...

- Notation for conditional distributions:
  - $P(\text{cavity} | \text{toothache}) = \text{a single number}$
  - $P(\text{cavity}, \text{toothache}) = \text{a 2x2 table summing to 1}$
  - $P(\text{cavity}, \text{toothache}) = \text{Two 2-element vectors, each summing to 1}$

- If we know more:
  - $P(\text{cavity} | \text{toothache, catch}) = 0.9$
  - $P(\text{cavity} | \text{toothache, cavity}) = 1$

- Note: the less specific belief remains valid after more evidence arrives, but is not always useful

- New evidence may be irrelevant, allowing simplification:
  - $P(\text{cavity} | \text{toothache, traffic}) = P(\text{cavity} | \text{toothache}) = 0.8$
  - This kind of inference, guided by domain knowledge, is crucial
**Conditioning**

- Conditional probabilities are the ratio of two probabilities:
  \[ P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} \]
  \[ P(x_1|x_2) = \frac{\sum_{x_1} P(x_1, x_2)}{P(x_2)} \]

\[ P(w|r) = \frac{P(w, r)}{P(r)} \]

<table>
<thead>
<tr>
<th>T</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>warm</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Normalization Trick**

- A trick to get the whole conditional distribution at once:
  - Get the joint probabilities for each value of the query variable
  - Renormalize the resulting vector

\[ P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{\sum_{x_1} P(x_1, x_2)}{P(x_2)} \]

<table>
<thead>
<tr>
<th>T</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>warm</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**The Product Rule**

- Sometimes joint \( P(X,Y) \) is easy to get
- Sometimes easier to get conditional \( P(X|Y) \)

\[ P(x|y) = \frac{P(x,y)}{P(y)} \]

\[ P(x,y) = P(x|y)P(y) \]

**Example: \( P(\text{sun}, \text{dry}) \)**

\[
\begin{array}{c|c|c|c}
 & P & D & \bar{D} \\
 R & 0.5 & 0.2 & 0.3 \\
 B & 0.5 & 0.8 & 0.7 \\
\end{array}
\]

**Lewis Carroll's Sack Problem**

- Sack contains a red or blue token, 50/50
- We add a red token
- If we draw a red token, what’s the chance of drawing a second red token?
- Variables:
  - \( F=\{r,b\} \) the original token
  - \( D=\{r,b\} \) the first token we draw
- Query: \( P(F=r|D=r) \)

**Lewis Carroll’s Sack Problem**

- Now we have \( P(F,D) \)
- Want \( P(F=r|D=r) \)

<table>
<thead>
<tr>
<th>F</th>
<th>D</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>r</td>
<td>0.8</td>
</tr>
<tr>
<td>r</td>
<td>b</td>
<td>0.2</td>
</tr>
<tr>
<td>b</td>
<td>r</td>
<td>0.5</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Bayes’ Rule**

- Two ways to factor a joint distribution over two variables:
  \[ P(x,y) = P(x|y)P(y) = P(y|x)P(x) \]

- Dividing, we get:
  \[ P(x|y) = \frac{P(y|x)P(x)}{P(y)} \]

- Why is this at all helpful?
  - Let us invert a conditional distribution
  - Often the one conditional is tricky but the other simple
  - Foundation of many systems we’ll see later (e.g., ASR, MT)

- In the running for most important AI equation!
### More Bayes’ Rule

- **Diagnostic probability from causal probability:**
  \[
P(C | E) = \frac{P(E | C) P(C)}{P(E)}
  \]
- **Example:**
  - \( m \) is meningitis, \( s \) is stiff neck
  \[
P(m | s) = \frac{P(s | m) P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.00008
  \]
  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

### Battleship

- **Let’s say we have two distributions:**
  - Prior distribution over ship locations: \( P(L) \)
  - Say this is uniform
  - Sensor reading model: \( P(R | L) \)
    - Given by some known black box
    - E.g. \( P(R = \text{yellow} | L = (1, 1)) = 0.1 \)
    - For now, assume the reading is always for the lower left corner
  - We can calculate the posterior distribution over ship locations using (conditionalized) Bayes’ rule:
  \[
P(l | r) \propto P(r | l) P(l)^*
  \]

### Inference by Enumeration

- **P(sun)?**
  - \( S \)
  - \( T \)
  - \( R \)
  - \( P \)
    
    | S  | T | R | P   |
    |----|---|---|-----|
    | sum | win | sun | 0.30 |
    | sum | win | rain | 0.05 |
    | sum | cold | sun | 0.10 |
    | sum | cold | rain | 0.05 |
    | win | win | sun | 0.15 |
    | win | win | rain | 0.05 |
    | win | cold | sun | 0.15 |
    | win | cold | rain | 0.20 |

- **P(sun | winter)?**
- **P(sun | winter, warm)?**

### Independence

- **Two variables are independent if:**
  \[
P(X, Y) = P(X) P(Y)
  \]
  - This says that their joint distribution factors into a product two simpler distributions
  - **Independence is a modeling assumption**
  - Empirical joint distributions: at best “close” to independent
  - What could we assume for \{Weather, Traffic, Cavity\}?  
  - How many parameters in the joint model?  
  - How many parameters in the independent model?  
  - Independence is like something from CSPs: what?
Example: Independence?

- Arbitrary joint distributions can be poorly modeled by independent factors

<table>
<thead>
<tr>
<th>T</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>warm</td>
<td>rain</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Conditional Independence

- P(Toothache, Cavity, Catch) has $2^3 = 8$ entries (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(catch | toothache, cavity) = P(catch | cavity)
- The same independence holds if I don’t have a cavity:
  - 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

- Unconditional (absolute) independence is very rare (why?)
- Conditional independence is our most basic and robust form of knowledge about uncertain environments:
  \[ P(X, Y | Z) = P(X | Z)P(Y | Z); \]
- What about this domain:
  - Traffic
  - Umbrella
  - Raining
- What about fire, smoke, alarm?

The Chain Rule II

- Can always factor any joint distribution as an incremental product of conditional distributions
  \[ P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2 | X_1)P(X_3 | X_2, X_1) \ldots \]
  \[ P(X_n | X_{n-1}, \ldots, X_1); \]
- Why?
- This actually claims nothing…
- What are the sizes of the tables we supply?

The Chain Rule III

- Trivial decomposition:
  \[ P(\text{Traffic}, \text{Rain}, \text{Umbrella}) = \]
  \[ P(\text{Rain})P(\text{Traffic} | \text{Rain})P(\text{Umbrella} | \text{Rain}, \text{Traffic}) \]
- With conditional independence:
  \[ P(\text{Traffic}, \text{Rain}, \text{Umbrella}) = \]
  \[ P(\text{Rain})P(\text{Traffic} | \text{Rain})P(\text{Umbrella} | \text{Rain}); \]
- Conditional independence is our most basic and robust form of knowledge about uncertain environments
- Graphical models (next class) will help us work with independence

Combining Evidence

- Imagine we have two sensor readings
- We want to calculate the distribution over ship locations given those observations:
  - List the probabilities we are given
- We first calculate the non-conditional (joint) probabilities we care about
- Then we renormalize