CSE 473: Artificial Intelligence

Probability

Steve Tanimoto
University of Washington

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All CS188 materials are available at http://ai.berkeley.edu.]
Topics from 30,000’

- We’re done with Part I Search and Planning!

- Part II: Probabilistic Reasoning
  - Diagnosis
  - Speech recognition
  - Tracking objects
  - Robot mapping
  - Genetics
  - Error correcting codes
  - ... lots more!

- Part III: Machine Learning
Outline

- **Probability**
  - Random Variables
  - Joint and Marginal Distributions
  - Conditional Distribution
  - Product Rule, Chain Rule, Bayes’ Rule
  - Inference
  - Independence

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

- **General situation:**
  - **Observed variables (evidence):** Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - **Unobserved variables:** Agent needs to reason about other aspects (e.g., where an object is or what disease is present)
  - **Model:** 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
What is...?

Random Variable

\[ P(W) \]

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.6</td>
</tr>
<tr>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>fog</td>
<td>0.3</td>
</tr>
<tr>
<td>meteor</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Value

Probability Distribution
A joint distribution over a set of random variables: $X_1, X_2, \ldots, X_n$ specifies a probability for each assignment (or outcome):

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

$$P(x_1, x_2, \ldots, x_n)$$

- Must obey: $P(x_1, x_2, \ldots, x_n) \geq 0$

$$\sum_{(x_1, x_2, \ldots, x_n)} P(x_1, x_2, \ldots, x_n) = 1$$

Size of joint distribution if $n$ variables with domain sizes $d$?

- For all but the smallest distributions, impractical to write out!

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</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>
A *probabilistic model* is a joint distribution over a set of random variables.

**Probabilistic models:**
- (Random) variables with domains
- Joint distributions: say whether assignments (called “outcomes”) are likely
- Normalized: sum to 1.0
- Ideally: only certain variables directly interact

**Constraint satisfaction problems:**
- Variables with domains
- Constraints: state whether assignments are possible
- Ideally: only certain variables directly interact

### Distribution over T,W

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</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>

### Constraint over T,W

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>T</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>F</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>F</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>T</td>
</tr>
</tbody>
</table>
Events

- An event is a set $E$ of outcomes
  \[ P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n) \]

- From a joint distribution, we can calculate the probability of any event
  - Probability that it’s hot AND sunny?
  - Probability that it’s hot?
  - Probability that it’s hot OR sunny?

- Typically, the events we care about are partial assignments, like $P(T=\text{hot})$

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Quiz: Events

- $P(+x, +y)$?
- $P(+x)$?
- $P(-y \text{ OR } +x)$?
Marginal Distributions

- Marginal distributions are **sub-tables** which eliminate variables.
- **Marginalization** (summing out): Combine collapsed rows by adding.

\[
P(T,W)
\]

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</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>

\[
P(T) = \sum_s P(t,s)
\]

\[
P(W) = \sum_t P(t,s)
\]

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

\[
P(T)
\]

<table>
<thead>
<tr>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>0.5</td>
</tr>
<tr>
<td>cold</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[
P(W)
\]

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.6</td>
</tr>
<tr>
<td>rain</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Quiz: Marginal Distributions

$$P(X, Y)$$

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td>+y</td>
<td>0.2</td>
</tr>
<tr>
<td>+x</td>
<td>-y</td>
<td>0.3</td>
</tr>
<tr>
<td>-x</td>
<td>+y</td>
<td>0.4</td>
</tr>
<tr>
<td>-x</td>
<td>-y</td>
<td>0.1</td>
</tr>
</tbody>
</table>

$$P(x) = \sum_y P(x, y)$$

$$P(y) = \sum_x P(x, y)$$

<table>
<thead>
<tr>
<th>X</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td></td>
</tr>
<tr>
<td>-x</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+y</td>
<td></td>
</tr>
<tr>
<td>-y</td>
<td></td>
</tr>
</tbody>
</table>
Conditional Probabilities

- A simple relation between joint and marginal probabilities
  - In fact, this is taken as the **definition** of a conditional probability

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

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</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>

\[
P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4
\]

\[
= P(W = s, T = c) + P(W = r, T = c)
= 0.2 + 0.3 = 0.5
\]
Quiz: Conditional Probabilities

\( P(X, Y) \)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td>+y</td>
<td>0.2</td>
</tr>
<tr>
<td>+x</td>
<td>-y</td>
<td>0.3</td>
</tr>
<tr>
<td>-x</td>
<td>+y</td>
<td>0.4</td>
</tr>
<tr>
<td>-x</td>
<td>-y</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- \( P(+x \mid +y) \)?
- \( P(-x \mid +y) \)?
- \( P(-y \mid +x) \)?
Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others.

**Conditional Distributions**

\[
P(W|T = h) = \begin{cases} 
  W & P \\
  
  \text{sun} & 0.8 \\
  \text{rain} & 0.2 \\
\end{cases}
\]

\[
P(W|T = c) = \begin{cases} 
  W & P \\
  
  \text{sun} & 0.4 \\
  \text{rain} & 0.6 \\
\end{cases}
\]

**Joint Distribution**

\[
P(T, W) = \begin{cases} 
  T & W & P \\
  
  \text{hot} & \text{sun} & 0.4 \\
  \text{hot} & \text{rain} & 0.1 \\
  \text{cold} & \text{sun} & 0.2 \\
  \text{cold} & \text{rain} & 0.3 \\
\end{cases}
\]
Conditional Distributions - The Slow Way...

\[
P(T, W)
\]

\[
P(W = s | T = c) = \frac{P(W = s, T = c)}{P(T = c)}
\]
\[
= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)}
\]
\[
= \frac{0.2}{0.2 + 0.3} = 0.4
\]

\[
P(W = r | T = c) = \frac{P(W = r, T = c)}{P(T = c)}
\]
\[
= \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)}
\]
\[
= \frac{0.3}{0.2 + 0.3} = 0.6
\]
Probabilistic Inference

- Probabilistic inference =
  "compute a desired probability from other known probabilities (e.g. conditional from joint)"

- We generally compute conditional probabilities
  - $P(\text{on time} \mid \text{no reported accidents}) = 0.90$
  - These represent the agent’s beliefs given the evidence

- Probabilities change with new evidence:
  - $P(\text{on time} \mid \text{no accidents, 5 a.m.}) = 0.95$
  - $P(\text{on time} \mid \text{no accidents, 5 a.m., raining}) = 0.80$
  - Observing new evidence causes beliefs to be updated
Inference by Enumeration

- **General case:**
  - Evidence variables: \( E_1 \ldots E_k = e_1 \ldots e_k \)
  - Query\(^*\) variable: \( Q \)
  - Hidden variables: \( H_1 \ldots H_r \)

\[
\begin{align*}
E_1 \ldots E_k &= e_1 \ldots e_k \\
Q &
H_1 \ldots H_r \\
\{ &
X_1, X_2, \ldots X_n \\
& All \ variables
\end{align*}
\]

- We want:

\[
P(Q|e_1 \ldots e_k)
\]

* Works fine with multiple query variables, too

- **Step 1:** Select the entries consistent with the evidence

\[
P(Q, e_1 \ldots e_k) = \sum_{h_1 \ldots h_r} P(Q, h_1 \ldots h_r, e_1 \ldots e_k) \\
X_1, X_2, \ldots X_n
\]

- **Step 2:** Sum out \( H \) to get joint of Query and evidence

- **Step 3:** Normalize

\[
Z = \sum_q P(Q, e_1 \ldots e_k) \\
P(Q|e_1 \ldots e_k) = \frac{1}{Z} P(Q, e_1 \ldots e_k)
\]

\[
x + \frac{1}{Z}
\]

* Works fine with multiple query variables, too
Inference by Enumeration

- $P(W)$?
- $P(W \mid \text{winter})$?
- $P(W \mid \text{winter, hot})$?

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>summer</td>
<td>hot</td>
<td>sun</td>
<td>0.30</td>
</tr>
<tr>
<td>2</td>
<td>summer</td>
<td>hot</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>summer</td>
<td>cold</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>4</td>
<td>summer</td>
<td>cold</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>winter</td>
<td>hot</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>6</td>
<td>winter</td>
<td>hot</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>7</td>
<td>winter</td>
<td>cold</td>
<td>sun</td>
<td>0.15</td>
</tr>
<tr>
<td>8</td>
<td>winter</td>
<td>cold</td>
<td>rain</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Inference by Enumeration

- Computational problems?
  - Worst-case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution
The Product Rule

- Sometimes have conditional distributions but want the joint

\[ P(y)P(x|y) = P(x, y) \quad \iff \quad P(x|y) = \frac{P(x, y)}{P(y)} \]
The Product Rule

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

- Example:

<table>
<thead>
<tr>
<th>( P(W) )</th>
<th>D</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td></td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>rain</td>
<td></td>
<td></td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P(D \mid W) )</th>
<th>D</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>wet sun</td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>dry sun</td>
<td></td>
<td></td>
<td>0.9</td>
</tr>
<tr>
<td>wet rain</td>
<td></td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>dry rain</td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P(D, W) )</th>
<th>D</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>wet sun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dry sun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wet rain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dry rain</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

\[ P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \]

\[ P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1}) \]
Independence

- Two variables are independent in a joint distribution if:

\[ P(X, Y) = P(X)P(Y) \]

\[ \forall x, y \quad P(x, y) = P(x)P(y) \]

- Says the joint distribution factors into a product of two simple ones
- Usually variables aren’t independent!

- Can use independence as a modeling assumption
  - Independence can be a simplifying assumption
  - Empirical joint distributions: at best “close” to independent
  - What could we assume for \{Weather, Traffic, Cavity\}?

- Independence is like something from CSPs: what?
Example: Independence?

\[
P_1(T, W) = \begin{array}{c|c|c}
T & W & P \\
\hline
\text{hot} & \text{sun} & 0.4 \\
\text{hot} & \text{rain} & 0.1 \\
\text{cold} & \text{sun} & 0.2 \\
\text{cold} & \text{rain} & 0.3 \\
\end{array}
\]

\[
P(T) = \begin{array}{c|c}
T & P \\
\hline
\text{hot} & 0.5 \\
\text{cold} & 0.5 \\
\end{array}
\]

\[
P_2(T, W) = P(T)P(W) = \begin{array}{c|c|c}
T & W & P \\
\hline
\text{hot} & \text{sun} & 0.3 \\
\text{hot} & \text{rain} & 0.2 \\
\text{cold} & \text{sun} & 0.3 \\
\text{cold} & \text{rain} & 0.2 \\
\end{array}
\]

\[
P(W) = \begin{array}{c|c}
W & P \\
\hline
\text{sun} & 0.6 \\
\text{rain} & 0.4 \\
\end{array}
\]
Example: Independence

- N fair, independent coin flips:

\[
\begin{array}{c|c}
P(X_1) & P(X_2) & P(X_n) \\
\hline
H & 0.5 & H & 0.5 & H & 0.5 \\
T & 0.5 & T & 0.5 & T & 0.5 \\
\end{array}
\]

\[
P(X_1, X_2, \ldots X_n) = 2^n
\]
Conditional Independence
Conditional Independence

- $P(\text{Toothache, Cavity, Catch})$

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - $P(+\text{catch} \mid +\text{toothache}, +\text{cavity}) = P(+\text{catch} \mid +\text{cavity})$

- The same independence holds if I don’t have a cavity:
  - $P(+\text{catch} \mid +\text{toothache}, -\text{cavity}) = P(+\text{catch} \mid -\text{cavity})$

- Catch is conditionally independent of Toothache given Cavity:
  - $P(\text{Catch} \mid \text{Toothache, Cavity}) = P(\text{Catch} \mid \text{Cavity})$

- Equivalent statements:
  - $P(\text{Toothache} \mid \text{Catch , Cavity}) = P(\text{Toothache} \mid \text{Cavity})$
  - $P(\text{Toothache, Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) P(\text{Catch} \mid \text{Cavity})$
  - One can be derived from the other easily
Conditional Independence

- Unconditional (absolute) independence very rare (why?)

- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

- X is conditionally independent of Y given Z
  \[ X \perp Y \mid Z \]

  if and only if:

  \[ \forall x, y, z : P(x, y \mid z) = P(x \mid z)P(y \mid z) \]

  or, equivalently, if and only if

  \[ \forall x, y, z : P(x \mid z, y) = P(x \mid z) \]
Conditional Independence

- What about this domain:
  - Traffic
  - Umbrella
  - Raining
Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm
Bayes Rule
Pacman – Sonar (P4)

[Demo: Pacman – Sonar – No Beliefs(L14D1)]
Video of Demo Pacman – Sonar (no beliefs)
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:

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

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we’ll see later (e.g. ASR, MT)

- In the running for most important AI equation!
Inference with Bayes’ Rule

- Example: Diagnostic probability from causal probability:

\[
P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}
\]

- Example:
  - M: meningitis, S: stiff neck
  - \(P(+m) = 0.0001\)
  - \(P(+s|m) = 0.8\)
  - \(P(+s|\neg m) = 0.01\)

\[
P(+m|s) = \frac{P(+s|m)P(+m)}{P(+s)} = \frac{P(+s|m)P(+m)}{P(+s|m)P(+m) + P(+s|\neg m)P(\neg m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999} = 0.0079
\]

- Note: posterior probability of meningitis still very small
- Note: you should still get stiff necks checked out! Why?
Ghostbusters Sensor Model

Values of Pacman’s Sonar Readings

|                | $P($red $|$ 3) | $P($orange $|$ 3) | $P($yellow $|$ 3) | $P($green $|$ 3) |
|----------------|---------------|------------------|------------------|------------------|
|                | 0.05          | 0.15             | 0.5              | 0.3              |

Real Distance = 3
Let’s say we have two distributions:

- Prior distribution over ghost location: $P(G)$
  - Let’s say this is uniform
- Sensor reading model: $P(R \mid G)$
  - Given: we know what our sensors do
  - $R$ = reading color measured at $(1,1)$
  - E.g. $P(R = \text{yellow} \mid G = (1,1)) = 0.1$

We can calculate the posterior distribution $P(G \mid r)$ over ghost locations given a reading using Bayes’ rule:

$$P(g \mid r) \propto P(r \mid g)P(g)$$
Video of Demo Ghostbusters with Probability
Probability Recap

- **Conditional probability**
  \[ P(x|y) = \frac{P(x, y)}{P(y)} \]

- **Product rule**
  \[ P(x, y) = P(x|y)P(y) \]

- **Chain rule**
  \[ P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \]

- **Bayes rule**
  \[ P(x|y) = \frac{P(y|x)}{P(y)} P(x) \]

- **X, Y independent if and only if:**
  \[ \forall x, y : P(x, y) = P(x)P(y) \]

- **X and Y are conditionally independent given Z:**
  \[ X \perp Y | Z \]
  \[ \forall x, y, z : P(x, y|z) = P(x|z)P(y|z) \]