Our Status in CS188

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

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

- Part III: Machine Learning
Part II: Probabilistic Reasoning

- Probability
- Distributions over LARGE Numbers of Random Variables
  - Representation
  - Independence
  - Inference
    - Variable Elimination
    - Sampling
  - Hidden Markov Models

---

Probability

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

You’ll need all this stuff A LOT for the next few weeks, so make sure you go over it now and know it inside out! The next few weeks we will learn how to make these work computationally efficiently for LARGE numbers of random variables.
Inference in Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: red
  - 1 or 2 away: orange
  - 3 or 4 away: yellow
  - 5+ away: green
- Sensors are noisy, but we know $P(\text{Color} \mid \text{Distance})$

<table>
<thead>
<tr>
<th>$P(\text{red} \mid 3)$</th>
<th>$P(\text{orange} \mid 3)$</th>
<th>$P(\text{yellow} \mid 3)$</th>
<th>$P(\text{green} \mid 3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.15</td>
<td>0.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Uncertainty

- General situation:
  - **Evidence**: Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - **Hidden 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
Random Variables

- A random variable is some aspect of the world about which we (may) 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 variables in a CSP, random variables have domains
  - $R$ in \{true, false\} (sometimes write as \{+$r$, $-r$\})
  - $D$ in $[0, \infty)$
  - $L$ in possible locations, maybe \{(0,0), (0,1), …\}

Probability Distributions

- Unobserved random variables have distributions
  - $P(T)$
    - $T$ | $P$
    - warm | 0.5
    - cold | 0.5
  - $P(W)$
    - $W$ | $P$
    - sun | 0.6
    - rain | 0.1
    - fog | 0.3
    - meteor | 0.0

- A distribution is a TABLE of probabilities of values
- A probability (lower case value) is a single number
  - $P(W = rain) = 0.1$  
  - $P(rain) = 0.1$

- Must have: $\forall x P(x) \geq 0$  
  - $\sum x P(x) = 1$
Joint Distributions

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

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

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

- Size of distribution if $n$ variables with domain sizes $d$?
- Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

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

- 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>

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables

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

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

<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>hot</td>
<td>sun</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

Marginal Distributions

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

  
  $P(T, W)
  
  \begin{array}{c|cc|c}
  \hline
  & 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 \\
  \hline
  \end{array}
  
  \begin{array}{c|c}
  \hline
  T & P \\
  \hline
  \text{hot} & 0.5 \\
  \text{cold} & 0.5 \\
  \hline
  \end{array}
  
  \begin{array}{c|c|c}
  \hline
  & W & P \\
  \hline
  \text{sun} & 0.6 \\
  \text{rain} & 0.4 \\
  \hline
  \end{array}

  \begin{align*}
  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)
  \end{align*}
Conditional Probabilities

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

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

\[
P(T, W)
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>W</td>
<td>P</td>
</tr>
<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>

\[
P(W = r|T = c) = ???
\]

Conditional Distributions

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

\[
P(W|T = hot)
\]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>P</td>
</tr>
<tr>
<td>sun</td>
<td>0.8</td>
</tr>
<tr>
<td>rain</td>
<td>0.2</td>
</tr>
</tbody>
</table>

\[
P(W|T = cold)
\]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>P</td>
</tr>
<tr>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>

\[
P(T, W)
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>W</td>
<td>P</td>
</tr>
<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>
Normalization Trick

- A trick to get a whole conditional distribution at once:
  - Select the joint probabilities matching the evidence
  - Normalize the selection (make it sum to one)

\[
P(T, W)
\]

\[
P(T, r)
\]

\[
P(T|r)
\]

- Why does this work? Sum of selection is \( P(\text{evidence}) \)\! (\( P(r) \), here)

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

Probabilistic Inference

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

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

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

- $P(\text{sun})$?
- $P(\text{sun} \mid \text{winter})$?
- $P(\text{sun} \mid \text{winter}, \text{warm})$?

<table>
<thead>
<tr>
<th>S</th>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>summer</td>
<td>hot</td>
<td>sun</td>
<td>0.30</td>
</tr>
<tr>
<td>summer</td>
<td>hot</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>sun</td>
<td>0.15</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>rain</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Inference by Enumeration

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

- We want: $P(Q \mid e_1 \ldots e_k)$

- First, select the entries consistent with the evidence
- Second, sum out $H$ to get joint of Query and evidence:

$$P(Q, e_1 \ldots e_k) = \sum_{h_1 \ldots h_r} P(Q, h_1 \ldots h_r, e_1 \ldots e_k)$$

- Finally, normalize the remaining entries to conditionalize

- Obvious problems:
  - Worst-case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution

* Works fine with multiple query variables, too
Inference by Enumeration Example 2: Model for Ghostbusters

- Reminder: ghost is hidden, sensors are noisy

- T: Top sensor is red
  B: Bottom sensor is red
  G: Ghost is in the top

- Queries:
  P( +g) = ??
  P( +g | +t) = ??
  P( +g | +t, -b) = ??

- Problem: joint distribution too large / complex

<table>
<thead>
<tr>
<th>T</th>
<th>B</th>
<th>G</th>
<th>P(T,B,G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>+b</td>
<td>+g</td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>+b</td>
<td>−g</td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>−b</td>
<td>+g</td>
<td>0.24</td>
</tr>
<tr>
<td>+t</td>
<td>−b</td>
<td>−g</td>
<td>0.04</td>
</tr>
<tr>
<td>−t</td>
<td>+b</td>
<td>+g</td>
<td>0.04</td>
</tr>
<tr>
<td>−t</td>
<td>+b</td>
<td>−g</td>
<td>0.24</td>
</tr>
<tr>
<td>−t</td>
<td>−b</td>
<td>+g</td>
<td>0.06</td>
</tr>
<tr>
<td>−t</td>
<td>−b</td>
<td>−g</td>
<td>0.06</td>
</tr>
</tbody>
</table>

The Product Rule

- Sometimes have conditional distributions but want the joint

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

- Example:

| P(W) | P(D|W) | P(D, W) |
|------|--------|---------|
| R    | D      | W      | P      |
| sun  | wet    | sun    | 0.1    |
|      | dry    | sun    | 0.9    |
|      | wet    | rain   | 0.7    |
|      | dry    | rain   | 0.3    |
| rain | wet    | sun    | 0.08   |
|      | dry    | sun    | 0.72   |
|      | wet    | rain   | 0.14   |
|      | dry    | rain   | 0.06   |
The Chain Rule

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

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

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

- Why is this always true?

- Can now build a joint distributions only specifying conditionals!
  - Bayesian networks essentially apply the chain rule plus make conditional independence assumptions.

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 is meningitis, s is stiff neck
    - \( P(s|m) = 0.8 \)
    - \( P(m) = 0.0001 \)
    - \( P(s) = 0.1 \)

\[ P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008 \]

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

Ghostbusters, Revisited

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

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

\[ P(g|r) \propto P(r|g)P(g) \]
Independence

- Two variables are independent if:
  \[\forall x, y : P(x, y) = P(x)P(y)\]
  - Says their joint distribution factors into a product two simpler ones.
  - Usually variables are not independent!
  - Equivalent definition of independence:
    \[\forall x, y : P(x|y) = P(x)\]
  - We write: \(X \perp Y\)

- Independence is a simplifying modeling assumption
  - Empirical joint distributions: at best “close” to independent
  - What could we assume for \{Weather, Traffic, Cavity, Toothache\}?

- Independence is like something from CSPs, what?

Example: Independence?

\[
\begin{array}{c|cc}
\text{T} & \text{P} \\
\text{warm} & 0.5 \\
\text{cold} & 0.5 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{T} & \text{W} & \text{P} \\
\text{warm} & \text{sun} & 0.4 \\
\text{warm} & \text{rain} & 0.1 \\
\text{cold} & \text{sun} & 0.2 \\
\text{cold} & \text{rain} & 0.3 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{W} & \text{P} \\
\text{sun} & 0.6 \\
\text{rain} & 0.4 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{T} & \text{W} & \text{P} \\
\text{warm} & \text{sun} & 0.3 \\
\text{warm} & \text{rain} & 0.2 \\
\text{cold} & \text{sun} & 0.3 \\
\text{cold} & \text{rain} & 0.2 \\
\end{array}
\]
Example: Independence

- N fair, independent coin flips:

<table>
<thead>
<tr>
<th>$P(X_1)$</th>
<th>$P(X_2)$</th>
<th>$P(X_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H 0.5</td>
<td>H 0.5</td>
<td>H 0.5</td>
</tr>
<tr>
<td>T 0.5</td>
<td>T 0.5</td>
<td>T 0.5</td>
</tr>
</tbody>
</table>

$P(X_1, X_2, \ldots X_n)$

2^n

Conditional Independence

- $P($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:
  - $P(+\text{catch} | +\text{toothache}, +\text{cavity}) = P(+\text{catch} | +\text{cavity})$

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

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

- Equivalent statements:
  - $P(\text{Toothache} | \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Cavity})$
  - $P(\text{Toothache}, \text{Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity})$
  - 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:

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

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

- What about fire, smoke, alarm?

The Chain Rule Revisited

\[
P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots
\]

- Trivial decomposition:

\[
P(\text{Traffic, Rain, Umbrella}) = \\
P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain, Traffic})
\]

- With assumption of conditional independence:

\[
P(\text{Traffic, Rain, Umbrella}) = \\
P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})
\]

- Representation size: $1 + 2 + 4$ versus $1 + 2 + 2$

- Bayes’ nets / graphical models are concerned with distributions with conditional independences
Ghostbusters Chain Rule

- Each sensor depends only on where the ghost is.
- That means, the two sensors are conditionally independent, given the ghost position.
- T: Top square is red
  B: Bottom square is red
  G: Ghost is in the top
- Givens:
  \[ P(+g) = 0.5 \]
  \[ P(+t | +g) = 0.8 \]
  \[ P(+t | -g) = 0.4 \]
  \[ P(+b | +g) = 0.4 \]
  \[ P(+b | -g) = 0.8 \]

\[
P(T,B,G) = P(G) P(T|G) P(B|G)
\]

<table>
<thead>
<tr>
<th>T</th>
<th>B</th>
<th>G</th>
<th>P(T,B,G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>+b</td>
<td>+g</td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>+b</td>
<td>¬g</td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>¬b</td>
<td>+g</td>
<td>0.24</td>
</tr>
<tr>
<td>+t</td>
<td>¬b</td>
<td>¬g</td>
<td>0.04</td>
</tr>
<tr>
<td>¬t</td>
<td>+b</td>
<td>+g</td>
<td>0.04</td>
</tr>
<tr>
<td>¬t</td>
<td>+b</td>
<td>¬g</td>
<td>0.24</td>
</tr>
<tr>
<td>¬t</td>
<td>¬b</td>
<td>+g</td>
<td>0.06</td>
</tr>
<tr>
<td>¬t</td>
<td>¬b</td>
<td>¬g</td>
<td>0.06</td>
</tr>
</tbody>
</table>