CS 188: Artificial Intelligence

Probability

Instructors: Angela Liu and Yanlai Yang
University of California, Berkeley
(Slides adapted from Stuart Russell and Dawn Song)
Uncertainty

- The real world is rife with uncertainty!
  - E.g., if I leave for SFO 2 hours before my flight, will I be there in time?
- Sources of Uncertainty:
  - partial observability (road state, other drivers’ plans, etc.)
  - noisy sensors (radio traffic reports, Google maps)
  - immense complexity of modelling and predicting traffic, security line, etc.
  - lack of knowledge of world dynamics (will tire burst?)
- Probabilistic assertions summarize effects of *ignorance* and *laziness*
- Combine probability theory + utility theory -> decision theory
  - *Maximize expected utility*: \( a^* = \text{argmax}_a \sum_s P(s \mid a) U(s) \)
Basic laws of (discrete) probability

- Begin with a set $\Omega$ of possible worlds
  - E.g., 6 possible rolls of a die, \{1, 2, 3, 4, 5, 6\}

- A probability model assigns a number $P(\omega)$ to each world $\omega$

- These numbers must satisfy
  - $0 \leq P(\omega) \leq 1$
  - $\sum_{\omega \in \Omega} P(\omega) = 1$
An event is any subset of $\Omega$
- E.g., “roll < 4” is the set {1,2,3}
- E.g., “roll is odd” is the set {1,3,5}

The probability of an event is the sum of probabilities over its worlds
- $P(A) = \sum_{\omega \in A} P(\omega)$
- E.g., $P(\text{roll} < 4) = P(1) + P(2) + P(3) = 1/2$
A random variable (usually denoted by a capital letter) is some aspect of the world about which we (may) be uncertain.

Formally a **deterministic function** of $\omega$.

The **range** of a random variable is the set of possible values:

- $\text{Odd} =$ Is the dice roll an odd number? $\rightarrow \{\text{true, false}\}$
  - e.g. $\text{Odd}(1)=$true, $\text{Odd}(6)=$false
  - often write the event $\text{Odd}=\text{true}$ as $\text{odd}$, $\text{Odd}=\text{false}$ as $\neg \text{odd}$

- $T =$ Is it hot or cold? $\rightarrow \{\text{hot, cold}\}$

- $D =$ How long will it take to get to the airport? $\rightarrow [0, \infty)$

- $L_{\text{Ghost}} =$ Where is the ghost? $\rightarrow \{(0,0), (0,1), \ldots\}$

The **probability distribution** of a random variable $X$ gives the probability for each value $x$ in its range (probability of the event $X=x$):

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

$P(x)$ for short (when unambiguous)

$P(X)$ refers to the entire distribution (think of it as a vector or table)
Probability Distributions

- Associate a probability with each value; sums to 1

  - Temperature:
    
    | T   | P  |
    |-----|----|
    | hot | 0.5|
    | cold| 0.5|

  - Weather:
    
    | W    | P  |
    |------|----|
    | sun  | 0.6|
    | rain | 0.1|
    | fog  | 0.3|
    | meteor | 0.0|

  - Joint distribution
    
    | Temperature |       |
    |-------------|-------|
    | hot         | 0.45  |
    | cold        | 0.15  |
    | Weather     |       |
    | sun         | 0.02  |
    | rain        | 0.08  |
    | fog         | 0.27  |
    | meteor      | 0.00  |
Making possible worlds

- In many cases we
  - begin with random variables and their domains
  - construct possible worlds as assignments of values to all variables
- E.g., two dice rolls $Roll_1$ and $Roll_2$
  - How many possible worlds?
  - What are their probabilities?
- Size of distribution for $n$ variables with range size $d$?
- For all but the smallest distributions, cannot write out by hand!
Recall that the probability of an event is the sum of probabilities of its worlds:

\[ P(A) = \sum_{\omega \in A} P(\omega) \]

So, given a joint distribution over all variables, can compute any event probability!

- Probability that it’s hot AND sunny?
- Probability that it’s hot?
- Probability that it’s hot OR not foggy?

<table>
<thead>
<tr>
<th>Weather</th>
<th>Temperature</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hot</td>
<td>0.45</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>sun</td>
<td></td>
<td>0.45</td>
<td>0.15</td>
</tr>
<tr>
<td>rain</td>
<td></td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>fog</td>
<td></td>
<td>0.03</td>
<td>0.27</td>
</tr>
<tr>
<td>meteor</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- **Marginalization** *(summing out)*: Collapse a dimension by adding

\[ P(X=x) = \sum_y P(X=x, Y=y) \]

<table>
<thead>
<tr>
<th>Weather</th>
<th>Temperature</th>
<th>( P(T) )</th>
<th>( P(W) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>hot</td>
<td>0.45</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>rain</td>
<td>hot</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>fog</td>
<td>hot</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.27</td>
<td>0.00</td>
</tr>
<tr>
<td>meteor</td>
<td>hot</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\[ P(T) = \begin{bmatrix} 0.50 & 0.50 \end{bmatrix} \]

\[ P(W) = \begin{bmatrix} 0.60 & 0.10 & 0.30 & 0.00 \end{bmatrix} \]
A simple relation between joint and conditional probabilities

In fact, this is taken as the definition of a conditional probability

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

\[ P(W=s | T=c) = \frac{P(W=s, T=c)}{P(T=c)} = \frac{0.15}{0.50} = 0.3 \]

\[ = P(W=s, T=c) + P(W=r, T=c) + P(W=f, T=c) + P(W=m, T=c) 
= 0.15 + 0.08 + 0.27 + 0.00 = 0.50 \]
## Conditional Distributions

- **Distributions for one set of variables given another set**

### Weather vs. Temperature

| Weather | Temperature | $P(W|T=h)$ | $P(W|T=c)$ | $P(W|T)$ |
|---------|-------------|-----------|-----------|---------|
|         | hot | cold | hot | cold | hot | cold |
| sun     | 0.45 | 0.15 | 0.90 | 0.30 | 0.90 | 0.30 |
| rain    | 0.02 | 0.08 | 0.04 | 0.16 | 0.04 | 0.16 |
| fog     | 0.03 | 0.27 | 0.06 | 0.54 | 0.06 | 0.54 |
| meteor  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
Normalizing a distribution

- (Dictionary) To bring or restore to a normal condition

- Procedure:
  - Multiply each entry by $\alpha = 1/(\text{sum over all entries})$

### Normalization of Weather and Temperature

| Weather | Temperature | $P(W,T)$ | $P(W | T=c)$ | Normalize |
|---------|-------------|----------|--------------|-----------|
| sun     | hot         | 0.45     | 0.15         | 0.30      |
|         | cold        | 0.15     | 0.16         | 0.16      |
| rain    | hot         | 0.02     | 0.08         | 0.08      |
|         | cold        | 0.08     | 0.08         | 0.08      |
| fog     | hot         | 0.03     | 0.27         | 0.27      |
|         | cold        | 0.27     | 0.54         | 0.54      |
| meteor  | hot         | 0.00     | 0.00         | 0.00      |
|         | cold        | 0.00     | 0.00         | 0.00      |

$\alpha = 1/0.50 = 2$
The Product Rule

- Sometimes have conditional distributions but want the joint

\[ P(a \mid b) \ P(b) = P(a, b) \]

\[ P(a \mid b) = \frac{P(a, b)}{P(b)} \]
The Product Rule: Example

\[ P(W \mid T) \, P(T) = P(W, T) \]

### \( P(W \mid T) \)

<table>
<thead>
<tr>
<th></th>
<th>hot</th>
<th>cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

### \( P(T) \)

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

### \( P(W, T) \)

<table>
<thead>
<tr>
<th></th>
<th>Temperature</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather</td>
<td></td>
<td>hot</td>
</tr>
<tr>
<td>sun</td>
<td>0.45</td>
<td>0.15</td>
</tr>
<tr>
<td>rain</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>fog</td>
<td>0.03</td>
<td>0.27</td>
</tr>
<tr>
<td>meteor</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
A joint distribution can be written as a product of conditional distributions by repeated application of the product rule:

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

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

- Probabilistic inference: compute a desired probability from a probability model
  - Typically for a \textit{query variable} given \textit{evidence}
  - E.g., $P(\text{airport on time} \mid \text{no accidents}) = 0.90$
  - These represent the agent’s \textit{beliefs} given the evidence

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

- Probability model $P(X_1, ..., X_n)$ is given
- Partition the variables $X_1, ..., X_n$ into sets as follows:
  - Evidence variables: $E = e$
  - Query variables: $Q$
  - Hidden variables: $H$

- We want: $P(Q | e)$

- Step 1: Select the entries consistent with the evidence

- Step 2: Sum out $H$ from model to get joint of query and evidence
  $$P(Q, e) = \sum_h P(Q, h, e)$$
  $X_1, ..., X_n$

- Step 3: Normalize
  $$P(Q | e) = \alpha P(Q, e)$$
Inference by Enumeration

- $P(W)$?

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

<table>
<thead>
<tr>
<th>Season</th>
<th>Temp</th>
<th>Weather</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>summer</td>
<td>hot</td>
<td>sun</td>
<td>0.35</td>
</tr>
<tr>
<td>summer</td>
<td>hot</td>
<td>rain</td>
<td>0.01</td>
</tr>
<tr>
<td>summer</td>
<td>hot</td>
<td>fog</td>
<td>0.01</td>
</tr>
<tr>
<td>summer</td>
<td>hot</td>
<td>meteor</td>
<td>0.00</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>summer</td>
<td>cold</td>
<td>fog</td>
<td>0.09</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>meteor</td>
<td>0.00</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>0.01</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>fog</td>
<td>0.02</td>
</tr>
<tr>
<td>winter</td>
<td>hot</td>
<td>meteor</td>
<td>0.00</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>
<tr>
<td>winter</td>
<td>cold</td>
<td>fog</td>
<td>0.18</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>meteor</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Inference by Enumeration

- **Obvious problems:**
  - Worst-case time complexity $O(d^n)$ (exponential in #hidden variables)
  - Space complexity $O(d^n)$ to store the joint distribution
  - $O(d^n)$ data points to estimate the entries in the joint distribution
Bayes’ Rule

- Write the product rule both ways:
  \[ P(a | b) \ P(b) = P(a, b) = P(b | a) \ P(a) \]

- Dividing left and right expressions, we get:
  \[ P(a | b) = \frac{P(b | a) \ P(a)}{P(b)} \]

- Why is this at all helpful?
  - Let us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Describes an “update” step from prior \( P(a) \) to posterior \( P(a | b) \)
Example: We know that meningitis causes stiff neck 80% of the time. Prior probability of any patient having meningitis is 0.0001. Prior probability of any patient having stiff neck is 0.01.

- \( M: \) meningitis, \( S: \) stiff neck

\[
\begin{align*}
P(s | m) &= 0.8 \\
P(m) &= 0.0001 \\
P(s) &= 0.01
\end{align*}
\]

\[
P(m | s) = \frac{P(s | m) \cdot P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.01}
\]

- Note: posterior probability of meningitis still very small: 0.008
Independence

- Two variables $X$ and $Y$ are (absolutely) **independent** if
  \[ \forall x, y \quad P(x, y) = P(x) \ P(y) \]

  - I.e., the joint distribution **factors** into a product of two simpler distributions

- Equivalently, via the product rule $P(x, y) = P(x | y) P(y)$,
  \[ P(x \ | \ y) = P(x) \quad \text{or} \quad P(y \ | \ x) = P(y) \]

- Example: two dice rolls $Roll_1$ and $Roll_2$
  - $P(Roll_1=5, Roll_2=3) = P(Roll_1=5) \ P(Roll_2=3) = 1/6 \times 1/6 = 1/36$
  - $P(Roll_2=3 \ | \ Roll_1=5) = P(Roll_2=3)$
Example: Independence

- $n$ fair, independent coin flips:

$$P(X_1) = \begin{array}{c|c}
H & 0.5 \\
T & 0.5 \\
\end{array}$$

$$P(X_2) = \begin{array}{c|c}
H & 0.5 \\
T & 0.5 \\
\end{array}$$

... 

$$P(X_n) = \begin{array}{c|c}
H & 0.5 \\
T & 0.5 \\
\end{array}$$

$$P(X_1, X_2, ..., X_n) \leq 2^n$$
Conditional Independence
**Conditional Independence**

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

- $X$ is conditionally independent of $Y$ given $Z$ if and only if:
  \[
  \forall x, y, z \quad P(x \mid y, z) = P(x \mid z)
  \]

  or, equivalently, if and only if
  \[
  \forall x, y, z \quad P(x, y \mid z) = P(x \mid z) P(y \mid z)
  \]
Conditional Independence Examples

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

- What about this domain:
  - Fire
  - Smoke
  - Alarm
Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: usually red
  - 1 or 2 away: mostly orange
  - 3 or 4 away: typically yellow
  - 5+ away: often green
- Click on squares until confident of location, then “bust”
Video of Demo Ghostbusters with Probability
Ghostbusters model

- Variables and ranges:
  - $G$ (ghost location) in $\{(1,1),\ldots,(3,3)\}$
  - $C_{x,y}$ (color measured at square $x,y$) in
    \{red,orange,yellow,green\}

- Ghostbuster physics:
  - *Uniform prior distribution* over ghost location: $P(G)$
  - *Sensor model*: $P(C_{x,y} \mid G)$ (depends only on distance to $G$)
    - E.g. $P(C_{1,1} = \text{yellow} \mid G = (1,1)) = 0.1$
Ghostbusters model, contd.

- $P(G, C_{1,1}, \ldots C_{3,3})$ has $9 \times 4^9 = 2,359,296$ entries!!

- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - E.g., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$?

- Ghostbuster physics again:
  - $P(C_{x,y} \mid G)$ depends only on distance to $G$
    - So $P(C_{1,1} = \text{yellow} \mid G = (2,3)) = P(C_{1,1} = \text{yellow} \mid G = (2,3), C_{1,2} = \text{orange})$
    - I.e., $C_{1,1}$ is conditionally independent of $C_{1,2}$ given $G$
Apply the chain rule to decompose the joint probability model:

\[ P(G, C_{1,1}, ... C_{3,3}) = P(G) P(C_{1,1} \mid G) P(C_{1,2} \mid G, C_{1,1}) P(C_{1,3} \mid G, C_{1,1}, C_{1,2}) ... P(C_{3,3} \mid G, C_{1,1}, ..., C_{3,2}) \]

Now simplify using conditional independence:

\[ P(G, C_{1,1}, ... C_{3,3}) = P(G) P(C_{1,1} \mid G) P(C_{1,2} \mid G) P(C_{1,3} \mid G) ... P(C_{3,3} \mid G) \]

I.e., conditional independence properties of ghostbuster physics simplify the probability model from *exponential* to *quadratic* in the number of squares

This is called a *Naïve Bayes* model:

- One discrete query variable; all other variables are evidence variables
- Evidence variables are conditionally independent given the query variable
To Summarize ...

- Basic laws: $0 \leq P(\omega) \leq 1$, $\sum_{\omega \in \Omega} P(\omega) = 1$, $P(A) = \sum_{\omega \in A} P(\omega)$
- Random variable $X(\omega)$ has a value in each $\omega$
  - Distribution $P(X)$ gives probability for each possible value $x$
  - Joint distribution $P(X,Y)$ gives total probability for each combination $x,y$
- Summing out/marginalization: $P(X=x) = \sum_y P(X=x,Y=y)$
- Conditional probability: $P(X|Y) = P(X,Y)/P(Y)$
- Chain rule: $P(X_1,..,X_n) = \prod_i P(X_i | X_1,..,X_{i-1})$
- Bayes Rule: $P(X|Y) = P(Y|X)P(X)/P(Y)$
- Independence: $P(X,Y) = P(X) P(Y)$ or $P(X|Y) = P(X)$ or $P(Y|X) = P(Y)$
- Conditional Independence: $P(X|Y,Z) = P(X|Z)$ or $P(X,Y|Z) = P(X|Z) P(Y|Z)$
Next time

- Bayes nets!