Today

- Uncertainty
- Probability Basics
  - Joint and Condition Distributions
  - Models and Independence
  - Bayes Rule
  - Estimation
- Utility Basics
  - Value Functions
  - Expectations

Uncertainty

- Let action $A_t = \text{leave for airport } t \text{ minutes before flight}$
- Will $A_t$ get me there on time?

Problems:
- partial observability (road state, other drivers' plans, etc.)
- noisy sensors (KCBS traffic reports)
- uncertainty in action outcomes (flat tire, etc.)
- immense complexity of modeling and predicting traffic

- A purely logical approach either
- Risks falsehood: "$A_{25}$ will get me there on time" or
- Leads to conclusions that are too weak for decision making:
  - "$A_{25}$ will get me there on time if there's no accident on the bridge, it doesn't rain, and my tires remain intact, etc., etc."

- $A_{1440}$ might reasonably be said to get me there on time but I'd have to stay overnight in the airport…

Probabilities

- Probabilistic approach
  - Given the available evidence, $A_{25}$ will get me there on time with probability 0.04
  - $P(A_{25} | \text{no reported accidents}) = 0.04$

- Probabilities change with new evidence:
  - $P(A_{25} | \text{no reported accidents}, 5 \text{ a.m.}) = 0.15$
  - $P(A_{25} | \text{no reported accidents}, 5 \text{ a.m., raining}) = 0.08$
  - i.e., observing evidence causes beliefs to be updated

Probabilistic Models

<table>
<thead>
<tr>
<th>CSPs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables with domains</td>
</tr>
<tr>
<td>Constraints: map from assignments to true/false</td>
</tr>
<tr>
<td>Ideally: only certain variables directly interact</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probabilistic models:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Random) variables with domains</td>
</tr>
<tr>
<td>Joint distributions: map from assignments (or outcomes) to positive numbers</td>
</tr>
<tr>
<td>Normalized: sum to 1.0</td>
</tr>
<tr>
<td>Ideally: only certain variables are directly correlated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>sun</td>
<td>T</td>
</tr>
<tr>
<td>warm</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>

What Are Probabilities?

- Objectivist / frequentist answer:
  - Averages over repeated experiments
  - E.g. empirically estimating $P(\text{rain})$ from historical observation
  - Assertion about how future experiments will go (in the limit)
  - New evidence changes the reference class
  - Makes one think of inherently random events, like rolling dice

- Subjectivist / Bayesian answer:
  - Degrees of belief about unobserved variables
  - E.g. an agent’s belief that it’s raining, given the temperature
  - Often estimate probabilities from past experience
  - New evidence updates beliefs

- Unobserved variables still have fixed assignments (we just don’t know what they are)
Probabilities Everywhere?

- Not just for games of chance!
  - I'm snuffling: am I sick?
  - Email contains "FREE!": is it spam?
  - Tooth hurts: have cavity?
  - Safe to cross street?
  - 60 min enough to get to the airport?

- Why can a random variable have uncertainty?
  - Inherently random process (dice, etc)
  - Insufficient or weak evidence
  - Unmodeled variables
  - Ignorance of underlying processes
  - The world’s just noisy!

- Compare to fuzzy logic, which has degrees of truth, or soft assignments

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 outcome, or atomic event) to reals:
  \[
  P(X_1 \equiv x_1, X_2 \equiv x_2, \ldots, X_n \equiv x_n)
  \]

- Size of distribution if \( n \) variables with domain sizes \( d \):
  \[
  \text{Size of distribution} = 2^n 
  \]

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

Examples

- An event is a set \( E \) of assignments (or outcomes):
  \[
  P(E) = \sum_{(x_1, x_2, \ldots, x_n) \in E} P(x_1, \ldots, x_n)
  \]

  - 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

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

- Notation for conditional distributions:
  - \( P(\text{cavity} \mid \text{toothache}) = \) a single number
  - \( P(\text{cavity} \mid \text{toothache}) = 4\)-element vector summing to 1
  - \( P(\text{cavity} \mid \text{toothache}) = \) Two 2-element vectors, each summing to 1

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

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

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

Marginalization

- Marginalization (or summing out) is projecting a joint distribution to a sub-distribution over subset of variables:
  \[
  P(X_1, X_3) = \sum_{X_2} P(X_1, X_2, X_3)
  \]

  - Marginalization is fixing some variables and renormalizing over the rest:
    \[
    P(X_1, X_3 \mid x_2) = \frac{P(X_1, x_2, X_3)}{\sum_{x_2} P(x_1, x_2, X_3)}
    \]

Conditioning

- Conditioning is fixing some variables and renormalizing over the rest:
  \[
  P(X_1, X_3 \mid x_2) = \frac{P(X_1, x_2, X_3)}{P(x_2)}
  \]

  - New evidence may be irrelevant, allowing simplification:
    - \( P(\text{cavity} \mid \text{toothache, traffic}) = P(\text{cavity} \mid \text{toothache}) = 0.8 \)
    - This kind of inference, sanctioned by domain knowledge, is crucial
Inference by Enumeration

- $P(R)$?
- $P(R|\text{winter})$?
- $P(R|\text{winter}, \text{warm})$?

<table>
<thead>
<tr>
<th>S</th>
<th>T</th>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum</td>
<td>sun</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>sun</td>
<td>warm</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>rain</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>warm</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rain</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rain</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Chain Rule I

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

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

- Example: $P(\text{Sun}, \text{Dry})$?

<table>
<thead>
<tr>
<th>S</th>
<th>D</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>wet</td>
<td>0.1</td>
</tr>
<tr>
<td>rain</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>wet</td>
<td>sun</td>
<td>0.1</td>
</tr>
<tr>
<td>dry</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>wet</td>
<td>rain</td>
<td>0.72</td>
</tr>
<tr>
<td>dry</td>
<td>rain</td>
<td>0.14</td>
</tr>
<tr>
<td>dry</td>
<td>rain</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Lewis Carroll’s Sack Problem

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

<table>
<thead>
<tr>
<th>F</th>
<th>D</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>r</td>
<td>0.5</td>
</tr>
<tr>
<td>b</td>
<td>r</td>
<td>0.0</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>0.25</td>
</tr>
</tbody>
</table>

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 [Sun, Dry, Toothache, Cavity]?

- How many parameters in the full joint model?
- How many parameters in the independent model?

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

- N fair, independent coins:
  \[ P(X_1, X_2, \ldots, X_n) = P(X_1) P(X_2) \cdots P(X_n) = 2^n \]

Example: Independence?

- Arbitrary joint distributions can be (poorly) modeled by independent factors

![Joint distributions](image)

Conditional Independence

- \( P(\text{Toothache}, \text{Cavity}, \text{Catch}) \) 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:
  \[ P(\text{catch} | \text{toothache}, \text{cavity}) = P(\text{catch} | \text{cavity}) \]
- The same independence holds if I haven't got 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:
  - \( P(\text{Toothache} | \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) \)
  - \( P(\text{Toothache}, \text{Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity}) \)

The Chain Rule II

- Can always factor any joint distribution as a product of incremental conditional distributions
  \[ P(X_1, X_2, \ldots, X_n) = P(X_2) P(X_3 | X_2) P(X_4 | X_3, X_2) \ldots P(X_n | X_1 \ldots X_{n-1}) \]
- Why?
  - This actually claims nothing…
  - What are the sizes of the tables we supply?

The Chain Rule III

- Write out full joint distribution using chain rule:
  \[ P(\text{Toothache}, \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Catch}, \text{Cavity}) P(\text{Catch} | \text{Cavity}) \]
- Graphical model notation:
  - Each variable is a node
  - The parents of a node are the other variables which the decomposed joint conditions on
  - MUCH more on this to come!
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's invert a conditional distribution
  - Often 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:
  \[ P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})}. \]
- Example:
  - \( m \) is meningitis, \( s \) is stiff neck
  \[ P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008 \]
  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

Combining Evidence

\[ P(\text{Cavity} | \text{toothache, catch}) = \frac{P(\text{toothache, catch} | \text{Cavity})P(\text{Cavity})}{\alpha} \]

This is an example of a naive Bayes model:

\[ P(\text{Cause}, \text{Effect}_1 \ldots \text{Effect}_n) = P(\text{Cause}) \prod_{i} P(\text{Effect}_i | \text{Cause}). \]

- Total number of parameters is linear in \( n \)!
- We'll see much more of naive Bayes next week

Expectations

\[ f : X \rightarrow \mathbb{R} \]

- Real valued functions of random variables:
  \[ E_P[f(X)] = \sum x f(x)P(x) \]
- Example: Expected value of a fair die roll
  \[
  \begin{array}{c|c|c|c|c}
  \text{Face} & 1 & 2 & 3 & 4 \\
  \text{Probability} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\
  \text{Value} & 1 & 2 & 3 & 4 \\
  \end{array}
  \]

  \[ 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{35}{6} \]

Expectations

- Expected seconds wasted because of spam filter

<table>
<thead>
<tr>
<th>Filter</th>
<th>Schedule</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strict</td>
<td>spam</td>
<td>0.43</td>
</tr>
<tr>
<td>Lax</td>
<td>spam</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>ham</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>ham</td>
<td>0.05</td>
</tr>
</tbody>
</table>

- We'll use the expected cost of actions to drive classification, decision networks, and reinforcement learning...

Utilities

- Preview of utility theory (later)

- Utilities:
  - Function from events to real numbers (payoffs)
  - E.g. spam
  - E.g. airport
Estimation

- How to estimate the distribution of a random variable $X$?
- **Maximum likelihood:**
  - Collect observations from the world
  - For each value $x$, look at the empirical rate of that value:
    $$\hat{p}(x) = \frac{\text{count}(x)}{\text{total sample}}$$
  - This estimate is the one which maximizes the likelihood of the data
- **Elicitation:** ask a human!
  - Harder than it sounds
  - E.g. what’s $P(\text{raining} \mid \text{cold})$?
  - Usually need domain experts, and sophisticated ways of eliciting probabilities (e.g. betting games)

Estimation

- **Problems with maximum likelihood estimates:**
  - If I flip a coin once, and it’s heads, what’s the estimate for $P(\text{heads})$?
  - What if I flip it 50 times with 27 heads?
  - What if I flip 10M times with 8M heads?
- **Basic idea:**
  - We have some prior expectation about parameters (here, the probability of heads)
  - Given little evidence, we should skew towards our prior
  - Given a lot of evidence, we should listen to the data
- **How can we accomplish this? Stay tuned!**