Announcements

- HW 8 is out, due 4/16
  - We’ll cover the last of the material you need for this on Wednesday
- Project 5 due 4/16
CS 188: Artificial Intelligence

Hidden Markov Models

Instructor: Sergey Levine and Anca Dragan --- University of California, Berkeley

[These slides were created by Dan Klein, Pieter Abbeel, and Sergey Levine. http://ai.berkeley.edu]
Markov Models

- Basic conditional independence:
  - Past and future independent of the present
  - Each time step only depends on the previous
  - This is called the (first order) Markov property

- Note that the chain is just a (growable) BN
  - We can always use generic BN reasoning on it if we truncate the chain at a fixed length
Example Markov Chain: Weather

- States: \( X = \{ \text{rain, sun} \} \)

- Initial distribution: 1.0 sun

- CPT \( P(X_t \mid X_{t-1}) \):

<table>
<thead>
<tr>
<th>( X_{t-1} )</th>
<th>( X_t )</th>
<th>( P(X_t \mid X_{t-1}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>sun</td>
<td>0.9</td>
</tr>
<tr>
<td>sun</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>rain</td>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>rain</td>
<td>rain</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Two new ways of representing the same CPT
Mini-Forward Algorithm

- Question: What’s \( P(X) \) on some day \( t \)?

\[
P(x_1) = \text{known}
\]

\[
P(x_t) = \sum_{x_{t-1}} P(x_{t-1}, x_t)
\]

\[
= \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1})
\]

Forward simulation
Stationary Distributions

- **Question:** What’s $P(X)$ at time $t = \infty$?

$$P_\infty(\text{sun}) = P(\text{sun}|\text{sun})P_\infty(\text{sun}) + P(\text{sun}|\text{rain})P_\infty(\text{rain})$$

$$P_\infty(\text{rain}) = P(\text{rain}|\text{sun})P_\infty(\text{sun}) + P(\text{rain}|\text{rain})P_\infty(\text{rain})$$

\[
egin{align*}
P_\infty(\text{sun}) &= 0.9P_\infty(\text{sun}) + 0.3P_\infty(\text{rain}) \\
P_\infty(\text{rain}) &= 0.1P_\infty(\text{sun}) + 0.7P_\infty(\text{rain}) \\
P_\infty(\text{sun}) &= 3P_\infty(\text{rain}) \\
P_\infty(\text{rain}) &= 1/3P_\infty(\text{sun})
\end{align*}
\]

Also:  
\[
P_\infty(\text{sun}) + P_\infty(\text{rain}) = 1
\]

**Table:**

| $X_{t-1}$ | $X_t$ | $P(X_t|X_{t-1})$ |
|----------|-------|------------------|
| sun      | sun   | 0.9              |
| sun      | rain  | 0.1              |
| rain     | sun   | 0.3              |
| rain     | rain  | 0.7              |

Also: 
\[
\begin{align*}
P_\infty(\text{sun}) &= 3/4 \\
P_\infty(\text{rain}) &= 1/4
\end{align*}
\]
Hidden Markov Models

- Markov chains not so useful for most agents
  - Need observations to update your beliefs

- Hidden Markov models (HMMs)
  - Underlying Markov chain over states $X$
  - You observe outputs (effects) at each time step
Example: Weather HMM

An HMM is defined by:

- Initial distribution: \( P(X_1) \)
- Transitions: \( P(X_t | X_{t-1}) \)
- Emissions: \( P(E_t | X_t) \)

| \( R_t \) | \( R_{t+1} \) | \( P(R_{t+1} | R_t) \) |
|---------|---------|----------------|
| +r      | +r      | 0.7            |
| +r      | -r      | 0.3            |
| -r      | +r      | 0.3            |
| -r      | -r      | 0.7            |

| \( R_t \) | \( U_t \) | \( P(U_t | R_t) \) |
|---------|---------|----------------|
| +r      | +u      | 0.9            |
| +r      | -u      | 0.1            |
| -r      | +u      | 0.2            |
| -r      | -u      | 0.8            |
Inference: Find State Given Evidence

- We are given evidence at each time and want to know

\[ B_t(X) = P(X_t|e_{1:t}) \]

- Idea: start with \( P(X_1) \) and derive \( B_t \) in terms of \( B_{t-1} \)
  - equivalently, derive \( B_{t+1} \) in terms of \( B_t \)
Two Steps: Passage of Time + Observation

\[ B(X_t) = P(X_t|e_{1:t}) \quad B'(X_{t+1}) \]

\[ B(X_{t+1}) \]
Passage of Time

- Assume we have current belief $P(X \mid \text{evidence to date})$
  
  
  \[
  B(X_t) = P(X_t | e_{1:t})
  \]

- Then, after one time step passes:

  \[
  P(X_{t+1} | e_{1:t}) = \sum_{x_t} P(X_{t+1}, x_t | e_{1:t})
  \]

  \[
  = \sum_{x_t} P(X_{t+1} | x_t, e_{1:t}) P(x_t | e_{1:t})
  \]

  \[
  = \sum_{x_t} P(X_{t+1} | x_t) P(x_t | e_{1:t})
  \]

- Basic idea: beliefs get “pushed” through the transitions
  - With the “B” notation, we have to be careful about what time step $t$ the belief is about, and what evidence it includes

- Or compactly:

  \[
  B'(X_{t+1}) = \sum_{x_t} P(X' | x_t) B(x_t)
  \]
Assume we have current belief $P(X \mid \text{previous evidence})$:

$$B'(X_{t+1}) = P(X_{t+1} | e_{1:t})$$

Then, after evidence comes in:

$$P(X_{t+1} | e_{1:t+1}) = \frac{P(X_{t+1}, e_{t+1} | e_{1:t})}{P(e_{t+1} | e_{1:t})} \,
\propto X_{t+1} \, P(X_{t+1}, e_{t+1} | e_{1:t})$$

$$= P(e_{t+1} | e_{1:t}, X_{t+1}) P(X_{t+1} | e_{1:t})$$

$$= P(e_{t+1} | X_{t+1}) P(X_{t+1} | e_{1:t})$$

Or, compactly:

$$B(X_{t+1}) \propto X_{t+1} \, P(e_{t+1} | X_{t+1}) B'(X_{t+1})$$

Basic idea: beliefs “rewighted” by likelihood of evidence

Unlike passage of time, we have to renormalize
The Forward Algorithm

- We are given evidence at each time and want to know

\[ B_t(X) = P(X_t|e_{1:t}) \]

- We can derive the following updates

\[
P(x_t|e_{1:t}) \propto_X P(x_t, e_{1:t}) = \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t})
= \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1}) P(x_t|x_{t-1}) P(e_t|x_t)
= P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}, e_{1:t-1})
\]

We can normalize as we go if we want to have \( P(x|e) \) at each time step, or just once at the end...
Online Belief Updates

- Every time step, we start with current $P(X \mid \text{evidence})$
- We update for time:
  \[ P(x_t|e_{1:t-1}) = \sum_{x_{t-1}} P(x_{t-1}|e_{1:t-1}) \cdot P(x_t|x_{t-1}) \]
- We update for evidence:
  \[ P(x_t|e_{1:t}) \propto_X P(x_t|e_{1:t-1}) \cdot P(e_t|x_t) \]
- The forward algorithm does both at once (and doesn’t normalize)
Particle Filtering
Inference with Samples: Particle Filtering

- **Particles:** track samples of states rather than an explicit distribution

![Diagram showing the process of particle filtering with steps for elapse, weight, and resampling.](image)
Example: Robot Localization
Most Likely Explanation
HMMs: MLE Queries

- HMMs defined by
  - States $X$
  - Observations $E$
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X|X_{-1})$
  - Emissions: $P(E|X)$

- New query: most likely explanation: $\arg\max_{x_{1:t}} P(x_{1:t}|e_{1:t})$

- New method: the Viterbi algorithm
State Trellis

- State trellis: graph of states and transitions over time

```
\begin{array}{cccc}
\text{sun} & \text{sun} & \text{sun} & \text{sun} \\
\text{rain} & \text{rain} & \text{rain} & \text{rain} \\
\end{array}
```

- $X_1$ $X_2$ $\cdots$ $X_N$

- Each arc represents some transition $x_{t-1} \rightarrow x_t$
- Each arc has weight $P(x_t|x_{t-1})P(e_t|x_t)$
- Each path is a sequence of states
- The product of weights on a path is that sequence’s probability along with the evidence
- Forward algorithm computes sums of paths, Viterbi computes best paths
Finding the Most Likely Path

Forward Algorithm (Sum)

\[ f_t[x_t] = P(x_t, e_{1:t}) = P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1}) f_{t-1}[x_{t-1}] \]

Viterbi Forward Phase (Max)

\[ m_t[x_t] = \max_{x_{1:t-1}} P(x_{1:t-1}, x_t, e_{1:t}) = P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) m_{t-1}[x_{t-1}] \]

probability of best path 1 : t that ends at \( x_t \)
Why is This True?

\[
m_1[x_1] = P(e_1|x_1)P(x_1)
\]

probability of best path 1 : t that ends at \( x_t \)

\[
m_2[x_2] = \max \{ P(e_2|x_2)P(x_2|x_1 = \text{rain})m_1[x_1 = \text{rain}], P(e_2|x_2)P(x_2|x_1 = \text{sun})m_1[x_1 = \text{sun}] \}
\]

\[
= \max_{x_1} P(e_2|x_2)P(x_2|x_1)m_1[x_1]
\]

\[
m_t[x_t] = \max_{x_{1:t-1}} P(x_{1:t-1}, x_t, e_{1:t})
\]

\[
= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1})m_{t-1}[x_{t-1}]
\]
Now What?

\[ m_N[x_N] \]

probability of best path 1 : N that ends at \( x_N \)

what is the last state on the most likely path?

\[ \text{arg max}_{x_N} m_N[x_N] \]

what is the second to last state on the most likely path?
A Tricky Counter-Example

\[ P(e_1 | x_1) P(x_1) \quad P(e_2 | x_2) P(x_2 | x_1) \]

\[ \text{arg max}_{x_1} m_1[x_1] = ? \quad \text{sun!} \]

\[ m_2[x_2] = \max_{x_1} P(e_2 | x_2) P(x_2 | x_1) m_1[x_1] \]

\[ m_2[x_2 = \text{sun}] = \max\{0.7 \times 0.01, 0.1 \times 0.5\} = 0.05 \]

\[ m_2[x_2 = \text{rain}] = \max\{0.7 \times 0.01, 0.1 \times 0.8\} = 0.08 \]

\[ \text{arg max}_{x_2} m_2[x_2] = \text{rain} \]

\[ P(x_1 = \text{sun}, x_2 = \text{rain}, e_1, e_2) = 0.7 \times 0.01 = 0.007 \]

best path 1 : t that ends at \( x_t \)  \neq \  \text{best path 1 : N that goes through } x_t
### What do We Do?

\[ P(e_1|x_1)P(x_1) \quad P(e_2|x_2)P(x_2|x_1) \]

0.7 \[ \text{sun} \]

0.1 \[ \text{rain} \]

0.01 \[ \text{sun} \]

0.3 \[ \text{rain} \]

\[ m_2[x_2] = \max_{x_1} P(e_2|x_2)P(x_2|x_1)m_1[x_1] \]
\[ m_2[x_2 = \text{sun}] = \max\{0.7 \times 0.01, 0.1 \times 0.5\} = 0.05 \]
\[ m_2[x_2 = \text{rain}] = \max\{0.7 \times 0.01, 0.1 \times 0.8\} = 0.08 \]

\[ \arg \max_{x_2} m_2[x_2] = \text{rain} \]

idea: what if we also save where the best path came from?

\[ m_t[x_t] = \max_{x_{t-1}} P(e_t|x_t)P(x_t|x_{t-1})m_{t-1}[x_{t-1}] \]
\[ a_t[x_t] = \arg \max_{x_{t-1}} P(e_t|x_t)P(x_t|x_{t-1})m_{t-1}[x_{t-1}] \]
Follow the Breadcrumbs...

\[ X_1 \quad X_2 \quad \cdots \quad X_N \]

for \( t = 1 \) to \( N \):

\[
m_t[x_t] = \max_{x_{t-1}} P(e_t|x_t)P(x_t|x_{t-1})m_{t-1}[x_{t-1}]
\]

\[
a_t[x_t] = \arg\max_{x_{t-1}} P(e_t|x_t)P(x_t|x_{t-1})m_{t-1}[x_{t-1}]
\]

last state on most likely path: \( x_N^* = \arg\max_{x_N} m_N[x_N] \)

second to last state on most likely path: \( x_{N-1}^* = a_N[x_N^*] \)

third to last state on most likely path: \( x_{N-2}^* = a_{N-1}[x_{N-1}^*] \)
Follow the Breadcrumbs...

\[
\begin{align*}
X_1 & \quad X_2 & \cdots & \quad X_N \\
\text{sun} & \quad \text{sun} & \quad \text{sun} & \quad \text{sun} \\
\text{rain} & \quad \text{rain} & \quad \text{rain} & \quad \text{rain}
\end{align*}
\]

for \( t = 1 \) to \( N \):

\[
m_t[x_t] = \max_{x_{t-1}} P(e_t|x_t)P(x_t|x_{t-1})m_{t-1}[x_{t-1}]
\]

\[
a_t[x_t] = \arg\max_{x_{t-1}} P(e_t|x_t)P(x_t|x_{t-1})m_{t-1}[x_{t-1}]
\]

\[
x_N^* = \arg\max_{x_N} m_N[x_N]
\]

for \( t = N \) to \( 2 \):

\[
x_{t-1}^* = a_t[x_t^*]
\]
Most Likely Explanation
Dynamic Bayes Nets
We want to track multiple variables over time, using multiple sources of evidence.

Idea: Repeat a fixed Bayes net structure at each time.

Variables from time $t$ can condition on those from $t-1$.

Dynamic Bayes nets are a generalization of HMMs.
**DBN Particle Filters**

- A particle is a complete sample for a time step

- **Initialize**: Generate prior samples for the \( t=1 \) Bayes net
  - Example particle: \( G_1^a = (3,3) \quad G_1^b = (5,3) \)

- **Elapse time**: Sample a successor for each particle
  - Example successor: \( G_2^a = (2,3) \quad G_2^b = (6,3) \)

- **Observe**: Weight each *entire* sample by the likelihood of the evidence conditioned on the sample
  - Likelihood: \( P(E_1^a \mid G_1^a) \times P(E_1^b \mid G_1^b) \)

- **Resample**: Select prior samples (tuples of values) in proportion to their likelihood
Exact Inference in DBNs

- Variable elimination applies to dynamic Bayes nets
- Procedure: “unroll” the network for $T$ time steps, then eliminate variables until $P(X_T|e_{1:T})$ is computed

Online belief updates: Eliminate all variables from the previous time step; store factors for current time only
Application: Speech Recognition
Speech Recognition in Action

[Video: NLP – ASR tvsample.avi (from Lecture 1)]
Digitizing Speech
Speech input is an acoustic waveform

“l” to “a” transition:

Figure: Simon Arnfield, http://www.psyc.leeds.ac.uk/research/cogn/speech/tutorial/
Spectral Analysis

- **Frequency gives pitch; amplitude gives volume**
  - Sampling at ~8 kHz (phone), ~16 kHz (mic) (kHz=1000 cycles/sec)

- **Fourier transform of wave displayed as a spectrogram**
  - Darkness indicates energy at each frequency

Human ear figure: depion.blogspot.com
Part of [ae] from “lab”

- Complex wave repeating nine times
  - Plus smaller wave that repeats 4x for every large cycle
  - Large wave: freq of 250 Hz (9 times in .036 seconds)
  - Small wave roughly 4 times this, or roughly 1000 Hz
Why These Peaks?

- **Articulator process:**
  - Vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others
Resonances of the Vocal Tract

- The human vocal tract as an open tube

- Air in a tube of a given length will tend to vibrate at resonance frequency of tube

- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end

![Diagram of the vocal tract as an open tube with closed and open ends, and a length of 17.5 cm.](image)

Figure: W. Barry Speech Science slides
Figure: Mark Liberman

[Demo: speech synthesis]
Video of Demo Speech Synthesis
Vowel [i] sung at successively higher pitches

Graphs: Ratre Wayland
Time slices are translated into acoustic feature vectors (~39 real numbers per slice)

These are the observations $E$, now we need the hidden states $X$
Speech State Space

- HMM Specification
  - $P(E|X)$ encodes which acoustic vectors are appropriate for each phoneme (each kind of sound)
  - $P(X|X')$ encodes how sounds can be strung together

- State Space
  - We will have one state for each sound in each word
  - Mostly, states advance sound by sound
  - Build a little state graph for each word and chain them together to form the state space $X$
States in a Word
Transitions with a Bigram Model

Training Counts

198015222 the first
194623024 the same
168504105 the following
158562063 the world
...
14112454 the door
-----------------------------------
23135851162 the *

\[
\hat{P}(\text{door}|\text{the}) = \frac{14112454}{23135851162} = 0.0006
\]
Decoding

- Finding the words given the acoustics is an HMM inference problem
- Which state sequence $x_{1:T}$ is most likely given the evidence $e_{1:T}$?

$$x_{1:T}^* = \arg \max_{x_{1:T}} P(x_{1:T}|e_{1:T}) = \arg \max_{x_{1:T}} P(x_{1:T}, e_{1:T})$$

- From the sequence $x$, we can simply read off the words