Announcements

- Written 3 due on Thursday night
  - Extra OHs before then: see web page

- Review session?  TBA

- Project 4 up!
  - Due 11/19

- Course contest update
  - You can qualify for the final tournament starting tonight!
Today

- HMMs: Most likely explanation queries
- Speech recognition
  - A massive HMM!
  - Details of this section not required
- Start machine learning

Speech and Language

- Speech technologies
  - Automatic speech recognition (ASR)
  - Text-to-speech synthesis (TTS)
  - Dialog systems
- Language processing technologies
  - Machine translation
  - Information extraction
  - Web search, question answering
  - Text classification, spam filtering, etc…
HMMs: MLE Queries

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

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

State Path Trellis

- State trellis: graph of states and transitions over time

- 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 the seq’s probability
- Can think of the Forward (and now Viterbi) algorithms as computing sums of all paths (best paths) in this graph
Viterbi Algorithm

\[
x^*_{1:T} = \arg \max_{x_{1:T}} P(x_{1:T} | e_{1:T}) = \arg \max_{x_{1:T}} P(x_{1:T}, e_{1:T})
\]

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

\[
= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) \max_{x_{1:t-2}} P(x_{1:t-1}, e_{1:t-1})\]

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

Example

<table>
<thead>
<tr>
<th>Rain_1</th>
<th>Rain_2</th>
<th>Rain_3</th>
<th>Rain_4</th>
<th>Rain_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>slate space paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
</tr>
<tr>
<td>false</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>umbrella</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
</tr>
<tr>
<td>false</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>most likely path</th>
</tr>
</thead>
<tbody>
<tr>
<td>.8182</td>
</tr>
<tr>
<td>.1818</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>m_{1:1}</th>
<th>m_{1:2}</th>
<th>m_{1:3}</th>
<th>m_{1:4}</th>
<th>m_{1:5}</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5155</td>
<td>.0491</td>
<td>.1237</td>
<td>.0173</td>
<td>.0024</td>
</tr>
</tbody>
</table>
Digitizing Speech

Continuous Sound pressure wave

Speech input is an acoustic wave form

"l" to "a" transition:

Speech in an Hour

Speech input is an acoustic wave form

Graphs from Simon Arnfield's web tutorial on speech, Sheffield:
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

Adding 100 Hz + 1000 Hz Waves
Part of [ae] from “lab”

- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves
Back to Spectra

- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.

- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

Resonances of the vocal tract

- The human vocal tract as an open tube

![Diagram of vocal tract with closed and open ends]

- 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.

Figure from W. Barry Speech Science slides
Acoustic Feature Sequence

- Time slices are translated into acoustic feature vectors (~39 real numbers per slice)

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

- $P(E|X)$ encodes which acoustic vectors are appropriate for each phoneme (each kind of sound)
- $P(X|X')$ encodes how sounds can be strung together
- We will have one state for each sound in each word
- From some state $x$, can only:
  - Stay in the same state (e.g. speaking slowly)
  - Move to the next position in the word
  - At the end of the word, move to the start of the next word
- We build a little state graph for each word and chain them together to form our state space $X$

HMMs for Speech
Transitions with Bigrams

Decoding

- While there are some practical issues, finding the words given the acoustics is an HMM inference problem.

- We want to know 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.
End of Part II!

- Now we’re done with our unit on probabilistic reasoning
- Last part of class: machine learning

Parameter Estimation

- Estimating the distribution of a random variable
- *Elicitation:* ask a human!
  - Usually need domain experts, and sophisticated ways of eliciting probabilities (e.g. betting games)
  - Trouble calibrating
- *Empirically:* use training data
  - For each outcome $x$, look at the *empirical rate* of that value:
    $$ P_{ML}(x) = \frac{\text{count}(x)}{\text{total samples}} $$
    - $P_{ML}(r) = 1/3$
  - This is the estimate that maximizes the *likelihood of the data*
    $$ L(x, \theta) = \prod_i P_\theta(x_i) $$
Estimation: Smoothing

- Relative frequencies are the maximum likelihood estimates

\[
\begin{align*}
\theta_{ML} &= \arg \max_{\theta} P(X|\theta) \\
&= \arg \max_{\theta} \prod_i P_\theta(X_i) \\
\implies P_{ML}(x) &= \frac{\text{count}(x)}{\text{total samples}}
\end{align*}
\]

- In Bayesian statistics, we think of the parameters as just another random variable, with its own distribution

\[
\begin{align*}
\theta_{MAP} &= \arg \max_{\theta} P(\theta|X) \\
&= \arg \max_{\theta} P(X|\theta) P(\theta)/P(X) \\
&= \arg \max_{\theta} P(X|\theta) P(\theta)
\end{align*}
\]

Estimation: Laplace Smoothing

- Laplace’s estimate:
  - Pretend you saw every outcome once more than you actually did

\[
\begin{align*}
P_{LAP}(x) &= \frac{c(x) + 1}{\sum_x [c(x) + 1]} \\
&= \frac{c(x) + 1}{N + |X|}
\end{align*}
\]

- Can derive this as a MAP estimate with Dirichlet priors (see cs281a)
Estimation: Laplace Smoothing

- **Laplace’s estimate (extended):**
  - Pretend you saw every outcome \( k \) extra times
  
  \[
  P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}
  \]

- What’s Laplace with \( k = 0 \)?
- \( k \) is the strength of the prior

- **Laplace for conditionals:**
  - Smooth each condition independently:
  
  \[
  P_{LAP,k}(x|y) = \frac{c(x,y) + k}{c(y) + k|X|}
  \]