CS489/698
Lecture 18: March 12, 2018

Hidden Markov Models

[RN] Sec. 15.3 [B] Sec. 13.1-13.2
[M] 17.3-17.5
Sequence Data

• So far, we assumed that the data instances are classified independently
  – More precisely, we assumed that the data is iid (identically and independently distributed)
    • E.g., text categorization, digit recognition in separate images, etc.

• In many applications, the data arrives sequentially and the classes are correlated
  – E.g., weather prediction, robot localization, speech recognition, activity recognition
Speech Recognition

![Speech Waveform Diagram]

- Frequency (Hz)
- Amplitude
- Time (sec)
- Speech Transcription: bayes' theorem
Classification

- Extension of some classification models for sequence data

<table>
<thead>
<tr>
<th>Generative models</th>
<th>Discriminative models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent classification</td>
<td>Correlated classification</td>
</tr>
<tr>
<td>Mixture of Gaussians</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Conditional Random Field</td>
</tr>
<tr>
<td>Feed Forward Neural Network</td>
<td>Recurrent Neural Network</td>
</tr>
</tbody>
</table>
Hidden Markov Model

Mixture of Gaussians       HMMs
Assumptions

• **Stationary Process**: transition and emission distributions are identical at each time step

\[
\begin{align*}
\Pr(x_t | y_t) &= \Pr(x_{t+1} | y_{t+1}) \quad \forall t \\
\Pr(y_t | y_{t-1}) &= \Pr(y_{t+1} | y_t) \quad \forall t
\end{align*}
\]

• **Markovian Process**: next state is independent of previous states given the current state

\[
\Pr(y_{t+1} | y_t, y_{t-1}, \ldots, y_1) = \Pr(y_{t+1} | y_t) \quad \forall t
\]
Hidden Markov Model

• Graphical Model

• Parameterization
  – Transition distribution:
  – Emission distribution:

• Joint distribution:
Mobile Robot Localisation

- Example of a Markov process

- Problem: uncertainty grows over time...
Mobile Robot Localisation

• Hidden Markov Model:
  \( y \): coordinates of the robot on a map
  \( x \): distances to surrounding obstacles (measured by laser range finders or sonars)
  \( \Pr(y_t|y_{t-1}) \): movement of the robot with uncertainty
  \( \Pr(x_t|y_t) \): uncertainty in the measurements provided by laser range finders and sonars

• Localisation: \( \Pr(y_t|x_t, \ldots, x_1) \)?
Inference in temporal models

• Four common tasks:
  – **Monitoring**: $\Pr(y_t | x_{1..t})$
  – **Prediction**: $\Pr(y_{t+k} | x_{1..t})$
  – **Hindsight**: $\Pr(y_k | x_{1..t})$ where $k < t$
  – **Most likely explanation**: 
    $$\argmax_{y_1, \ldots, y_t} \Pr(y_{1..t} | x_{1..t})$$

• What algorithms should we use?
Monitoring

• $\Pr(y_t|x_{1..t})$: distribution over current state given observations

• Examples: robot localisation, patient monitoring

• Recursive computation:

  \[
  \Pr(y_t|x_{1..t}) \propto \Pr(x_t|y_t, x_{1..t-1}) \Pr(y_t|x_{1..t-1}) \text{ by Bayes' thm}
  \]

  \[
  = \Pr(x_t|y_t) \Pr(y_t|x_{1..t-1}) \text{ by conditional independence}
  \]

  \[
  = \Pr(x_t|y_t) \sum_{y_{t-1}} \Pr(y_t, y_{t-1}|x_{1..t-1}) \text{ by marginalization}
  \]

  \[
  = \Pr(x_t|y_t) \sum_{y_{t-1}} \Pr(y_t|y_{t-1}, x_{1..t-1}) \Pr(y_{t-1}|x_{1..t-1}) \text{ by chain rule}
  \]

  \[
  = \Pr(x_t|y_t) \sum_{y_{t-1}} \Pr(y_t|y_{t-1}) \Pr(y_{t-1}|x_{1..t-1}) \text{ by cond ind}
  \]
Forward Algorithm

• Compute $\Pr(y_t|x_{1..t})$ by forward computation

$$\Pr(y_1|x_1) \propto \Pr(x_1|y_1) \Pr(y_1)$$

For $i = 2$ to $t$ do

$$\Pr(y_i|x_{1..i}) \propto \Pr(x_i|y_i) \sum_{y_{i-1}} \Pr(y_i|y_{i-1}) \Pr(y_{i-1}|x_{1..i-1})$$

End

• Linear complexity in $t$
Prediction

- $\Pr(y_{t+k}|x_{1..t})$: distribution over future state given observations
- Examples: weather prediction, stock market prediction

Recursive computation

\[
\Pr(y_{t+k}|x_{1..t}) = \sum_{y_{t+k-1}} \Pr(y_{t+k}, y_{t+k-1}|x_{1..t}) \text{ by marginalization}
\]

\[
= \sum_{y_{t+k-1}} \Pr(y_{t+k}|y_{t+k-1}, x_{1..t}) \Pr(y_{t+k-1}|x_{1..t}) \text{ by chain rule}
\]

\[
= \sum_{y_{t+k-1}} \Pr(y_{t+k}|y_{t+k-1}) \Pr(y_{t+k-1}|x_{1..t}) \text{ by cond ind}
\]
Forward Algorithm

1. Compute $\Pr(y_t | x_{1..t})$ by forward computation

   $\Pr(y_1 | x_1) \propto \Pr(x_1 | y_1) \Pr(y_1)$

   For $i = 1$ to $t$ do

   
   $\Pr(y_i | x_{1..i}) \propto \Pr(x_i | y_i) \sum_{y_{i-1}} \Pr(y_i | y_{i-1}) \Pr(y_{i-1} | x_{1..i-1})$

   End

2. Compute $\Pr(y_{t+k} | x_{1..t})$ by forward computation

   For $j = 1$ to $k$ do

   
   $\Pr(y_{t+j} | x_{1..t}) = \sum_{y_{t+j-1}} \Pr(y_{t+j} | y_{t+j-1}) \Pr(y_{t+j-1} | x_{1..t})$

   End

• Linear complexity in $t + k$
Hindsight

- \( \Pr(y_k|x_{1..t}) \) for \( k < t \): distribution over a past state given observations
- Example: delayed activity/speech recognition
- computation:
  \[
  \Pr(y_k|x_{1..t}) \propto \Pr(y_k, x_{k+1..t}|x_{1..k}) \quad \text{by conditioning}
  = \Pr(y_k|x_{1..k}) \Pr(x_{k+1..t}|y_k) \quad \text{by chain rule}
  \]
- Recursive computation
  \[
  \Pr(x_{k+1..t}|y_k) = \sum_{y_{k+1}} \Pr(y_{k+1}, x_{k+1..t}|y_k) \quad \text{by marginalization}
  = \sum_{y_{k+1}} \Pr(y_{k+1}|y_k) \Pr(x_{k+1..t}|y_{k+1}) \quad \text{by chain rule}
  = \sum_{y_{k+1}} \Pr(y_{k+1}|y_k) \Pr(x_{k+1}|y_{k+1}) \Pr(x_{k+2..t}|y_{k+1}) \quad \text{by cond ind}
  \]
Forward-backward algorithm

1. Compute $\text{Pr}(y_k | x_{1..k})$ by forward computation

   $\text{Pr}(y_1 | x_1) \propto \text{Pr}(x_1 | y_1) \text{Pr}(y_1)$

   For $i = 2$ to $k$ do
   
   $\text{Pr}(y_i | x_{1..i}) \propto \text{Pr}(x_i | y_i) \sum_{y_{i-1}} \text{Pr}(y_i | y_{i-1}) \text{Pr}(y_{i-1} | x_{1..i-1})$

   End

2. Compute $\text{Pr}(x_{k+1..t} | y_k)$ by backward computation

   $\text{Pr}(x_t | y_{t-1}) = \sum_{y_t} \text{Pr}(y_t | y_{t-1}) \text{Pr}(x_t | y_t)$

   For $j = t - 1$ downto $k$ do

   $\text{Pr}(x_{j..t} | y_{j-1}) = \sum_{y_j} \text{Pr}(y_j | y_{j-1}) \text{Pr}(x_j | y_j) \text{Pr}(x_{j+1..t} | y_j)$

   End

3. $\text{Pr}(y_k | x_{k+1..t}) \propto \text{Pr}(y_k | x_{1..k}) \text{Pr}(x_{k+1..t} | y_k)$

   • Linear complexity in $t$
Most likely explanation

- \( \arg \max_{y_{1..t}} \Pr(y_{1..t}|x_{1..t}) \): most likely state sequence given observations
- Example: speech recognition
- Computation:
  \[
  \max_{y_{1..t}} \Pr(y_{1..t}|x_{1..t}) = \max_{y_t} \Pr(x_t|y_t) \max_{y_{1..t-1}} \Pr(y_{1..t}|x_{1..t-1})
  \]
- Recursive computation:
  \[
  \max_{y_{1..i}} \Pr(y_{1..i}|x_{1..i-1}) \propto \max_{y_{i-1}} \Pr(y_i|y_{i-1}) \Pr(x_{i-1}|y_{i-1}) \max_{y_{1..i-2}} \Pr(y_{1..i-1}|x_{1..i-2})
  \]
Viterbi Algorithm

1. Compute $\max_{y_{1..t}} \Pr(y_{1..t} | x_{1..t})$ by dynamic programming

$$\max_{y_1} \Pr(y_{1..2} | x_1) \propto \max_{y_1} \Pr(y_2 | y_1) \Pr(x_1 | y_1) \Pr(y_1)$$

For $i = 2$ to $t - 1$ do

$$\max_{y_{1..i}} \Pr(y_{1..i+1} | x_{1..i}) \propto \max_{y_i} \Pr(y_{i+1} | y_i) \Pr(x_i | y_i) \max_{y_{1..i-1}} \Pr(y_{1..i} | x_{1..i-1})$$

End

$$\max_{y_{1..t}} \Pr(y_{1..t} | x_{1..t}) \propto \max_{y_t} \Pr(x_t | y_t) \max_{y_{1..t-1}} \Pr(y_{1..t} | x_{1..t-1})$$

• Linear complexity in $t$
Case Study: Activity Recognition

- Task: infer activities performed by a user of a smart walker
  - Inputs: sensor measurements
  - Output: activity
Inputs: Raw Sensor Data

• 8 channels:
  – Forward acceleration
  – Lateral acceleration
  – Vertical acceleration
  – Load on left rear wheel
  – Load on right rear wheel
  – Load on left front wheel
  – Load on right front wheel
  – Wheel rotation counts (speed)

• Data recorded at 50 Hz and digitized (16 bits)
Data Collection

- 8 walker users at Winston Park (84-97 years old)
- 12 older adults (80-89 years old) in the Kitchener-Waterloo area who do not use walkers

Output: Activities

- Not Touching Walker (NTW)
- Standing (ST)
- Walking Forward (WF)
- Turning Left (TL)
- Turning Right (TR)
- Walking Backwards (WB)
- Sitting on the Walker (SW)
- Reaching Tasks (RT)
- Up Ramp/Curb (UR/UC)
- Down Ramp/Curb (DR/DC)
Hidden Markov Model (HMM)

- Parameters
  - Initial state distribution: $\pi_{\text{class}} = \Pr(y_1 = \text{class})$
  - Transition probabilities: $\theta_{\text{class}|\text{class}} = \Pr(y_{t+1} = \text{class'}|y_t = \text{class})$
  - Emission probabilities: $\phi^i_{\text{val}|\text{class}} = \Pr(x^i_t = \text{val}|y_t = \text{class})$
    or $N(\mu^i_{\text{class}}, \sigma^i_{\text{class}}) = \Pr(x^i_t = \text{val}|y_t = \text{class})$

- Maximum likelihood:
  - Supervised: $\pi^*, \theta^*, \phi^* = \arg\max_{\pi, \theta, \phi} \Pr(y_{1:T}, x_{1:T}|\pi, \theta, \phi)$
  - Unsupervised: $\pi^*, \theta^*, \phi^* = \arg\max_{\pi, \theta, \phi} \Pr(x_{1:T}|\pi, \theta, \phi)$
Demo
Maximum Likelihood

• Supervised Learning: $y$’s are known
• Objective: $\underset{\pi, \theta, \phi}{\text{argmax}} \Pr(y_{1..t}, x_{1..t} | \pi, \theta, \phi)$

• Derivation:
  – Set derivative to 0
  – Isolate parameters $\pi, \theta, \phi$

• Consider a single input $x$ per time step
• Let $y \in \{c_1, c_2\}$ and $x \in \{v_1, v_2\}$
Multinomial emissions

- Let $c_i^{start}$ be # times of that process starts in class $c_i$
- Let $c_i$ be # of times that process is in class $c_i$
- Let $(c_i, c_j)$ be # of times that $c_i$ follows $c_j$
- Let $(v_i, c_j)$ be # of times that $v_i$ occurs with $c_j$
- $\Pr(y_{0..t}, x_{1..t})$
  
  $= \Pr(y_0) \prod_{i=1}^{t} \Pr(y_i | y_{i-1}) \Pr(x_i | y_i)$
  
  $= (\pi_{c_1})^{c_1^{start}} (1 - \pi_{c_1})^{c_2^{start}} (\theta_{c_1|c_1})^{(c_1,c_1)} (1 - \theta_{c_1|c_1})^{(c_2,c_1)}$
  
  $(\theta_{c_1|c_2})^{(c_1,c_2)} (1 - \theta_{c_1|c_2})^{(c_2,c_2)} (\phi_{v_1|c_1})^{(v_1,c_1)} (1 - \phi_{v_1|c_1})^{(v_2,c_1)}$
  
  $(\phi_{v_1|c_2})^{(v_1,c_2)} (1 - \phi_{v_1|c_2})^{(v_2,c_2)}$
Multinomial emissions

\[ \arg\max_{\pi, \theta, \phi} \Pr(y_{1..t}, x_{1..t} | \pi, \theta, \phi) \]

\[ \arg\max_{\pi_c_1} (\pi_{c_1})^{#c_1^{start}} (1 - \pi_{c_1})^{#c_2^{start}} \]

\[ \arg\max_{\theta_{c_1|c_1}} (\theta_{c_1|c_1})^{(#c_1,c_1)} (1 - \theta_{c_1|c_1})^{(#c_2,c_1)} \]

\[ \arg\max_{\theta_{c_1|c_2}} (\theta_{c_1|c_2})^{(#c_1,c_2)} (1 - \theta_{c_1|c_2})^{(#c_2,c_2)} \]

\[ \arg\max_{\phi_{v_1|c_1}} (\phi_{v_1|c_1})^{(#v_1,c_1)} (1 - \phi_{v_1|c_1})^{(#v_2,c_1)} \]

\[ \arg\max_{\phi_{v_1|c_2}} (\phi_{v_1|c_2})^{(#v_1,c_2)} (1 - \phi_{v_1|c_2})^{(#v_2,c_2)} \]
Multinomial emissions

• Optimization problem:

\[
\max_{\pi_{c_1}} \left( \pi_{c_1}^{#c_1^{\text{start}}} \right) \left( 1 - \pi_{c_1} \right)^{#c_2^{\text{start}}}
\]

\[
\Rightarrow \max_{\pi_{c_1}} \left( #c_1^{\text{start}} \right) \log(\pi_{c_1}) + \left( #c_2^{\text{start}} \right) \log(1 - \pi_{c_1})
\]

• Set derivative to 0:

\[
0 = \frac{#c_1^{\text{start}}}{\pi_{c_1}} - \frac{#c_2^{\text{start}}}{1 - \pi_{c_1}}
\]

\[
\Rightarrow (1 - \pi_{c_1})(#c_1^{\text{start}}) = (\pi_{c_1})(#c_2^{\text{start}})
\]

\[
\Rightarrow \pi_{c_1} = \frac{#c_1^{\text{start}}}{#c_1^{\text{start}} + #c_2^{\text{start}}}
\]
Relative Frequency Counts

- Maximum likelihood solution

\[ \pi_{c_1}^{\text{start}} = \frac{\#c_1^{\text{start}}}{\#c_1^{\text{start}} + \#c_2^{\text{start}}} \]

\[ \theta_{c_1|c_1} = \frac{\#(c_1, c_1)}{\#(c_1, c_1) + \#(c_2, c_1)} \]

\[ \theta_{c_1|c_2} = \frac{\#(c_1, c_2)}{\#(c_1, c_2) + \#(c_2, c_2)} \]

\[ \phi_{v_1|c_1} = \frac{\#(v_1, c_1)}{\#(v_1, c_1) + \#(v_2, c_1)} \]

\[ \phi_{v_1|c_2} = \frac{\#(v_1, c_2)}{\#(v_1, c_2) + \#(v_2, c_2)} \]
Gaussian Emissions

- Maximum likelihood solution

\[
\pi_{c_1}^{\text{start}} = \frac{\#c_1^{\text{start}}}{\#c_1^{\text{start}} + \#c_2^{\text{start}}} \\
\theta_{c_1|c_1} = \frac{\#(c_1, c_1)}{\#(c_1, c_1) + \#(c_2, c_1)} \\
\theta_{c_1|c_2} = \frac{\#(c_1, c_2)}{\#(c_1, c_2) + \#(c_2, c_2)} \\
\mu_{c_1} = \frac{1}{\#c_1} \sum_{t|y_t=c_1} x_t, \quad \sigma_{c_1}^2 = \frac{1}{\#c_1} \sum_{t|y_t=c_1} (x_t - \mu_{c_1})^2 \\
\mu_{c_2} = \frac{1}{\#c_2} \sum_{t|y_t=c_2} x_t, \quad \sigma_{c_2}^2 = \frac{1}{\#c_2} \sum_{t|y_t=c_2} (x_t - \mu_{c_2})^2
\]