Genome 540 Discussion

February 27th, 2024
Clifford Rostomily
Assignment 7 Questions?

- Part 1: Use your predicted D-segments from hw6 to
  - Generate a new scoring scheme
  - Simulate background sequence
- Part 2: Run your D-segment program on the background and compare to the real data
- Part 3: Answer some questions
Assignment 8
HMM Tasks

Rabiner 1989:

**Likelihood**: Given an HMM $\lambda = (A, B)$ and an observation sequence $O$, determine the likelihood $P(O|\lambda)$.

**Decoding**: Given an observation sequence $O$ and an HMM $\lambda = (A, B)$, discover the best hidden state sequence $Q$.

**Learning**: Given an observation sequence $O$ and the set of states in the HMM, learn the HMM parameters $A$ and $B$. 
Example

Your dog is very moody and you want to know when they **like** or **hate** you so you start recording what they are doing when you get home everyday...

- **Waiting**
  - I’ve been sitting next to the door all day!

- **Lounging**
  - Oh it’s you...

- **Sleeping**
  - Out cold
Model

Initiation probabilities ($p_k$)

Transition Matrix ($A$)

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>H</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Set of Hidden States ($S = \{k\}$)

Emission Probabilities of the observations ($B = \{b\}$) given the state

P(waiting|like) = .6
P(lounging|like) = .3
P(sleeping|like) = .1
P(waiting|hate) = .1
P(lounging|hate) = .4
P(sleeping|hate) = .5

$p_{init} = [0.5, 0.5]$
Graphical representation with data

start

- L
- H

W|L

- L
- H

L|L

- L
- H

L

S|L

S|H

Waiting

Lounging

Sleeping
Graphical representation with data

Start

Initiation

Waiting

Lounging

Sleeping
Graphical representation with data

State 1

- Start
  - L
  - H

- Waiting
  - W|L
  - W|H

- Lounging
  - L|L
  - L|H

- Sleeping
  - L
  - H
  - S|L
  - S|H
Graphical representation with data

Emission

- L
- W|L
- L
- L|L
- L
- S|L

Waiting

- H
- W|H
- H
- L|H
- H
- S|H

Lounging

Sleeping
Graphical representation with data

Transition

- **Start**
  - L
  - H

- **Waiting**
  - W|L
  - W|H

- **Lounging**
  - L|L
  - L|H

- **Sleeping**
  - L
  - S|L
  - S|H
1. Step 1: Expectation
   a. Compute the forward probabilities
   b. Compute the backward probabilities

2. Step 3: Maximization
   a. Update the transition and emission probabilities
Forward Algorithm - **Likelihood** of an observed sequence

3 steps:
1. Initialization
2. Recursion
3. Termination
Forward Algorithm - **Likelihood** of an observed sequence

\[ f_1(L) = p(L) \cdot e(W|L) \]

\[ f_2(L) = (f_1(L) \cdot a_{LL} + f_1(H) \cdot a_{HL}) \cdot e(L|L) \]

\[ f_1(H) = p(H) \cdot e(W|H) \]

***Emission may also be written \( e_H(W) \)***
Computing the backward probabilities

Backward probabilities: probability of seeing the observations from time $t + 1$ to the end
Computing the backward probabilities

Backward probabilities: probability of seeing the observations from time $t + 1$ to the end
Computing the backward probabilities

\[
\begin{align*}
p(L) & \quad e(W|L) \quad a_{LL} \quad e(L|L) \quad a_{LL} \quad e(S|L) \quad S|L \\
p(H) & \quad e(W|H) \quad a_{HL} \quad e(L|H) \quad a_{HL} \quad e(S|H) \quad S|H \\
\end{align*}
\]
Computing the backward probabilities

\[ b_t(i) = b_{t+1}(L) * a_{LL} * e(L|L) + b_{t+1}(H) * a_{LH} * e(L|H) \]

**Initialize assuming** \( b_T(i) = 1 \)

\[ b_t(i) = \sum_j b_{t+1}(j) * a_{ij} * e_j(t+1) \]
Calculating the transition probabilities

\[ P_{t}(i,j) = \frac{f_{t}(i) * a_{ij} * e_{t}(o_{t+1}) * b_{t+1}(j)}{\sum_{j=1}^{N} f_{t}(j)b_{t}(j)} \]
Calculating the transition probabilities

\[ P_t(i,j) = \frac{f_t(i) \cdot a_{ij} \cdot e_j(o_{t+1}) \cdot b_{t+1}(j)}{\sum_{j=1}^{N} f_t(j) b_t(j)} \]

\[ a(i,j) = \frac{\sum_{t=1}^{T-1} P_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} P_t(i,k)} \]

- Probability of observations constrained on a specific transition
- Probability of observations given the model
Calculating the emission probabilities

\[ y_t(j) = \frac{f_t(j)b_t(j)}{P(O|\lambda)} = \frac{f_t(j)b_t(j)}{\sum_{j=1}^{N} f_t(j)b_t(j)} \]

Probability of being in state \( j \) at time \( t \) given the observation sequence \( O \) and the model

\[ e_t(v_k|j) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} \]

\[ e_t(v_k|j) = e_j(v_k) = \frac{\sum_{t=1,O_t=v_k}^{T} y_t(j)}{\sum_{t=1}^{T} y_t(j)} \]

Sum of all \( y_t(j) \) where the observed symbol = \( v_k \)
Avoiding vanishing probabilities

- Scaling
  - Good tutorial

- Work in log space
  - Mann 2006
Scaling

- When computing forward probabilities, also compute a scaling factor $c_t$.
  \[ c_t = \frac{1}{\sum_{i=1}^{N} f_t(i)} \]

- New forward probabilities at time $t$ are multiplied by $c_t$.

- Use $c_t$ for scaling backward probabilities as well.

- To get back true forward/backward probabilities

  \[ f^*_t(i) = (\prod_{t=1}^{T} c_t)f_t(i) \]
Reminders

- HW7 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template