Genome 540 Discussion

February 29th, 2024 - Happy Leap Day!
Clifford Rostomily
Assignment 8
Example - 2 state HMM for genomic sequences

Transition Matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Initiation probabilities

\[ P_{\text{init}} = [0.5, 0.5] \]

Set of Hidden States

State 1
- \( P(A|S1) \)
- \( P(C|S1) \)
- \( P(G|S1) \)
- \( P(T|S1) \)

State 2
- \( P(A|S2) \)
- \( P(C|S2) \)
- \( P(G|S2) \)
- \( P(T|S2) \)

Emission probabilities of the observations
Some notation

\[ S = \{ S_1, S_2, \ldots, S_N \} \quad q_t = \text{state at time } t \quad \text{Set of states (Size N)} \]

\[ V = \{ v_1, v_2, \ldots, v_m \} \quad \text{Set of emitted symbols (vocabulary) (size M)} \]

\[ A = \{ a_{ij} \} \quad a_{ij} = P( q_{t+1} = S_j | q_t = S_i ) \quad 1 \leq i, j \leq N \quad \text{Transition matrix (N x N)} \]

\[ O = O_1 O_2 \ldots O_T \quad O_t = \text{output symbol at time } t \quad \text{Sequence of observed symbols (Length T)} \]

\[ B = \{ b_j(k) \} \quad b_j(k) = P( O_t = v_k | q_t = S_j ) \quad 1 \leq j \leq N, 1 \leq k \leq M \quad \text{Emission probabilities} \]

\[ \pi = \{ \pi_i \} \quad \pi_i = P( q_1 = S_i ) \quad 1 \leq i \leq N \quad \text{Initiation probabilities} \]
Baum Welch Steps

1. Compute (scaled) forward probabilities and scaling factors using the forward trellis
2. Compute (scaled) backward probabilities using the backward trellis
3. Compute updated parameter estimates
4. Repeat until convergence
Baum Welch Steps

1. Compute (scaled) forward probabilities and scaling factors using the forward trellis
2. Compute (scaled) backward probabilities using the backward trellis
3. Compute updated parameter estimates
4. Repeat until convergence
Forward trellis

Initiation

Transition

Emission

π₁

π₂

start

A

C

G

b₁(A)

b₁(C)

b₁(G)

b₂(A)

b₂(C)

b₂(G)

a₁₁

a₁₂

a₁₁

a₁₂

a₂₁

a₂₁

a₂₂

a₂₂
Compute forward probabilities: $\alpha_t(i)$

$$\alpha_t(i) = P(O_1O_2...O_T, q_t = S_i | \lambda)$$

Initiation

$$\alpha_1(i) = \pi_i b_i(O_1)$$

Extension

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1})$$
Example

\[ \alpha_1(1) = \pi_1 \times b_1(A) \]

\[ \alpha_2(1) = (\alpha_1(A) \times a_{11} + \alpha_1(2) \times a_{21}) \times b_1(C) \]
Example... with scaling

1. \( \alpha_1(1) = \pi_1 \cdot b_1(A) \)
2. \( c_1 = \frac{1}{\alpha_1(1) + \alpha_1(2)} \)
3. \( \hat{\alpha}_1(1) = \pi_1 \cdot b_1(A) \cdot c_1 \)
4. \( \dot{\alpha}_2(A) = (\hat{\alpha}_1(1) \cdot \alpha_{11} + \hat{\alpha}_1(2) \cdot \alpha_{21}) \cdot b_1(C) \)
5. Repeat 2-3 to get \( \hat{\alpha} \)

Scaling factor equation:
\[
c_t = \frac{1}{\sum_{i=1}^{N} \dot{\alpha}_i(i)}
\]
Compute the backward probabilities: $\beta_t(i)$

$$\beta_t(i) = P(O_{t+1}O_{t+2}\cdots O_T|q_t = S_i, \lambda)$$

$$\beta_T(i) = 1$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$
\[ \pi_1(1) = (\pi_2(1) b_1(C) a_{11} + \pi_2(2) b_2(C) a_{12}) c_2 \]

\[ \pi_2(1) = (1 b_1(G) a_{11} + 1 b_2(G) a_{12}) c_2 \]

*For scaling just multiply by the corresponding scaling factors from the forward probabilities*
Calculating the updated probabilities

Once you have the forward and backward probabilities you can also calculate:

\[ P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \]

\[ \gamma_t(i) = P(q_t = S_i|O, \lambda) = \frac{\alpha_t(i) \beta_t(i)}{P(O|\lambda)} \]

\[ \xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j|O, \lambda) = \frac{\alpha_t(i)a_{ij} b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \]
Calculating the updated probabilities

...and then you can compute the updated transition, initiation, and emission probabilities

\[
\bar{\pi}_i = \gamma_1(i);
\]

\[
\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)};
\]

\[
\bar{b}_j(k) = \frac{\sum_{t=1, O_t=v_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)}
\]
But what about scaling?

\[
\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i) \cdot a_{ij} b_{j}(O_{t+1}) \cdot \hat{\beta}_{t+1}(j)}{\sum_{t=1}^{T-1} \hat{\alpha}_t(i) \cdot \hat{\beta}_t(i) / c_t}
\]

\[
\bar{b}_{j}(k) = \frac{\sum_{t=1, O_t = v_k}^{T} \hat{\alpha}_t(j) \cdot \hat{\beta}_t(j) / c_t}{\sum_{t=1}^{T} \hat{\alpha}_t(j) \cdot \hat{\beta}_t(j) / c_t}
\]
Avoiding vanishing probabilities

- These slides follow the following tutorial:
  - Shen Scaling Tutorial
- Alternatively you can skip scaling and work in log space. How to do this is described here:
  - Mann 2006
Reminders

- HW8 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template