Hidden Markov Models

Zhiping Weng

Differece Between MM and HMM

For HMM, there is not a one-to-one correspondence between the states and observed sequence. The states are hidden. There is no way to tell for certain what state the model was in when $x_i$ was generated. Although it is no longer possible to tell what state the system is in by looking at the corresponding sequence, it is often the underlying state path we are interested in.
Defining of an HMM

- **Sequence of states**, also called a path ($\pi$), is not observable. It follows a simple Markov Model: $a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$
- Each state "emits" a symbol from a distribution over all possible symbols. The **emission probability** of symbol $b$ at state $k$ is defined as: $e_k(b) = P(x_i = b \mid \pi_i = k)$
- **Sequence of symbols**, $x$, is observable. Each symbol $x_i$ is emitted by a state $\pi_i$, however, we do not know what $\pi_i$ is.

---

The CpG Island Example
Another Example: A Rare AT-rich Site

The Three Basic HMM Questions

2. The likelihood or evaluation question: How does one compute the likelihood of a sequence, given the model?

1. The optimal state path or alignment or decoding question: How does one compute the most likely state path associated with a sequence, given the model?

3. The learning question: How does one estimate the emission and transition parameters of an HMM, from a set of representative sequences?
If We Know the State Path, It Is Easy to Evaluate a Sequence

\[ P(x \cap \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i,\pi_{i+1}} \]

\( x = (x_1, x_2, \ldots, x_L) \) is the sequence of length \( L \)
\( \pi = (\pi_1, \pi_2, \ldots, \pi_L) \) is the state path and \( \pi_{L+1} = E \)
\( P(x \cap \pi) \) is the joint probability of sequence \( x \) and state \( \pi \)
\( e_{\pi_i}(x_i) = P(x_i \mid \pi_i) \) is the emission probability
\( a_{\pi_i,\pi_{i+1}} = P(\pi_{i+1} \mid \pi_i) \) is the transition probability

Examples

For CpG island HMM:
\[ P(CGCG \cap C, G, C, G.) = a_{BC} \cdot (e_{C, C} \cdot a_{C,G} \cdot (e_{G, C} \cdot a_{G,C} \cdot (e_{C, (C) \cdot a_{C,G} \cdot (e_{G, G)}) \cdot a_{G,E}) \cdot a_{G,E}) \]

For AT-rich Site HMM:
\[ P(ATG \cap NNS) = a_{BN} \cdot (e_{N, T} \cdot a_{N,S}) \cdot (e_{S, G} \cdot a_{S,E}) \cdot a_{S,E} \cdot a_{S,E} \]
\[ = .5 \cdot (.25 \cdot .85) \cdot (.25 \cdot .05) \cdot (.1 \cdot .1) \]

\( B \) : The "Begin" state that does not emit
\( E \) : The "End" state that does not emit
Most Probable Path

A sequence can be generated by different paths. The "most probable" path \( \pi^* \) is the one with the largest \( P(x \cap \pi) \).

How can we find \( \pi^* \)?
Dynamic Programming, of course!

Using the Viterbi Algorithm to Find the Most Probable Path

\[
v_k(i) = P(x_i, x_{i-1}, ..., x_1 \cap \pi_i^* = k)
\]

- The prob of the most probable path \( \pi^* \) ending in state \( k \) with observation \( x_i \)
- \( v_b(0) = 1; \quad v_b(i) = 0 \) for \( k \neq B; \quad v_b(i) = 0 \) for \( i > 0 \)
- \( v_j(i+1) = e_j(x_{i+1}) \cdot \max_k \left( v_j(i) \cdot a_{ij} \right) \)
- \( v_N(1) = e_N(A) \cdot v_b(0) \cdot a_{BN} = .25 \cdot 1 \cdot .5 = .125 \)
- \( v_3(1) = e_3(A) \cdot v_b(0) \cdot a_{BS} = .4 \cdot 1 \cdot .5 = .2 \)
- \( v_N(2) = e_N(T) \cdot \max \left( v_N(1) \cdot a_{NS}, v_3(1) \cdot a_{SN} \right) \)
  \[
  = .25 \cdot \max( .125 \cdot .85, .2 \cdot 1 ) = .25 \cdot 0.125 \cdot 0.85 \cdot 0.1 = .0266
  \]
**The Probability That a Sequence Is Generated by a HMM: the Forward Algorithm**

### Forward Algorithm

- **Objective**: Calculate the probability of a sequence given a hidden Markov model (HMM).

#### Algorithm Steps

1. **Initialization**:
   - Calculate $f_k(i)$ for the initial state $k$.
   - For $k = N, S$:
     - $f_N(0) = 1$ if $i = 0$ otherwise $f_N(0) = 0$.
     - $f_S(0) = 0$.

2. **Recursion**:
   - For $i = 1, 2, ..., k$:
     - $f_N(i) = e_N(x_i) \cdot \sum_k (f_K(i-1) \cdot a_{nk})$.
     - $f_S(i) = e_S(x_i) \cdot \sum_k (f_K(i-1) \cdot a_{sk})$.

3. **Termination**:
   - The probability of the entire sequence is $f_N(k)$.

#### Example

- **Sequence**: $x = A, T, G, T$
- **States**: $N, S$
- **Values**: $e_N(A) = .25, e_N(T) = .5, e_N(G) = .25$, $e_S(A) = .1, e_S(T) = .5, e_S(G) = .25$
- **Transition Probabilities**: $a_{NN} = .25, a_{NS} = .5$, $a_{SN} = .1, a_{SS} = .85$

- **Calculations**:
  - $f_N(1) = e_N(A) \cdot a_{NN} = .25 \cdot .25 = .0625$
  - $f_S(1) = e_S(A) \cdot a_{SN} = .1 \cdot .25 = .025$
  - $f_N(2) = e_N(T) \cdot a_{NT} + e_N(G) \cdot a_{NG} + e_N(T) \cdot a_{NS} = .5 \cdot .5 + .25 \cdot .25 + .5 \cdot .5 = .525$
  - $f_S(2) = e_S(T) \cdot a_{ST} + e_S(G) \cdot a_{SG} + e_S(T) \cdot a_{SN} + e_S(G) \cdot a_{SN} = .5 \cdot .1 + .25 \cdot .85 + .5 \cdot .1 + .25 \cdot .85 = .4025$

**Overall Probability**:

$$P(x) = f_N(4) \cdot a_{NE} + f_S(4) \cdot a_{SE}$$
The Backward Algorithm

\[
b_{k}(i) = P(x \mid \pi_i = k) = P(x_{i+1}x_{i+2}\ldots x_{L} \mid \pi_i = k) \text{: The probability of the observed sequence after } x_i \text{ given that } \pi_i = k
\]

\[
b_{k}(i) = \sum_{l} (e_l(x_{i+l}) \cdot b_{l}(i+1) \cdot a_{lk})
\]

\[
b_{k}(L) = a_{lE} \text{ for all } k
\]

\[
b_{s}(4) = P(E \mid \pi_3 = N) = a_{sN} = .1
\]

\[
b_{s}(4) = P(E \mid \pi_3 = S) = a_{sS} = .1
\]

\[
b_{S}(3) = P(x_{3} \mid \pi_3 = N) = P(T \mid \pi_3 = N) = \varepsilon_{s}(T) \cdot b_{s}(4) \cdot a_{sN} + \varepsilon_{s}(T) \cdot b_{s}(4) \cdot a_{sS}
\]

\[
= .25 \cdot .1 \cdot .85 + .4 \cdot 1 \cdot .05 = .02125 + .00225 = .02325
\]

Decoding: The Posterior State Probabilities

\[
P(\pi_i = k \mid x) = \frac{P(x \cap \pi_i = k)}{P(x)}
\]

\[
= \frac{P(x_{i+1}x_{i+2}\ldots x_{L}, x_{i} \cap \pi_i = k)}{\sum_{k} f_k(L) a_{ik}}
\]

\[
= \frac{P(x_{i+1}x_{i+2}\ldots x_{L} \mid \pi_i = k)}{\sum_{k} f_k(L) a_{ik}}
\]

\[
= \frac{f_k(i) \cdot b_k(i)}{\sum_{k} f_k(L) a_{ik}}
\]

BE561 11/20/00
Zhiping Weng
Learning: Estimating the Transition and Emission Probabilities

It is easy when the state paths for training data are known.

\[ a_{kl} = \frac{A_{kl}}{\sum_l A_{kl}} \]

\[ e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \]

\[
\text{AATTATGGGCGATTATGGCCCTAGCAATGGCTGGCGA} \\
\text{NNSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
\[ P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) \]

\[ f_k(i) \quad k \quad l \quad b_l(i+1) \]

1 \quad i \quad i+1 \quad i+2 \quad L

A Profile-HMM

A hidden Markov model derived from the alignment discussed. The transitions are shown with arrows whose thickness indicate their probability. In each state the histogram shows the probabilities of the four nucleotides.
Input to HMM: Multiple Alignment

QUERY 1 ACDYTNCYSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 55
       1433 ACDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 55
       194574 CDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 76
       132831 CDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 81
       171184 SCEYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 54
       222171 CDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 76

QUERY 2 ACDYTNCYSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 55
       1433 ACDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 55
       194574 CDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 76
       132831 CDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 81
       171184 SCEYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 54
       222171 CDYTNCYSSSEEDVSTQAAGYKLHEDGETVGSNSYPHKNYYEGDFSE---V----SSP 76

From Sequence to Profile HMM

• Start with begin, end state & # of AAs you think are in main pattern – e.g., if AA sequence family has 30 clearly aligned columns, you would start with 30 main states
• Each state (dot) emits (generates) an AA with a certain probability. For instance, if we are in the beginning state, B, and 90% of the consensus sequences generate an L, we would emit “L” with probability 0.9
• Insert states allow for insertions of arbitrary # of AAs at that position (this is to capture ‘fuzzy’ sections of the protein sequence family.
• Delete states do not emit any symbols, but allow for gaps and so differences in length between two protein sequences.
Profile HMM SH3 domain, 30AAs
Shaded most conserved
Chosen to be states
Unshaded=insert/delete