Hidden Markov Models

BMI/CS 576
www.biostat.wisc.edu/bmi576
Colin Dewey
colin.dewey@wisc.edu
Fall 2016
Classifying sequences

• Markov chains
  – useful for modeling a single class of sequence
  – likelihood ratios of different models can be used to classify sequences

• What if a sequence contains multiple classes of elements?
  – Example: a whole genome sequence
  – How can we model such sequences?
  – How can we partition these sequences into their component elements?
Revisiting the CpG question

• Given a sequence $x_1..x_L$ we can use two Markov chains to decide if $x_1..x_L$ is a CpG island or not.
• What do we do if we were asked to “find” the CpG islands in the genome?
• We have to search for these “islands” in a “sea” of non-islands
One attempt: merge Markov chains

Problem: when we observe a DNA sequence, we do not observe “A+” or “A-”, we simply observe “A”, for example.
Hidden State

- We’ll distinguish between the *observed* parts of a problem and the *hidden* parts.
- In the Markov models we’ve considered previously, it is clear which state accounts for each part of the observed sequence.
- In the model above, there are multiple states that could account for each part of the observed sequence.
  - This is the hidden part of the problem.
An HMM for an occasionally dishonest casino

What is hidden? Which die is rolled

What is observed? Number (1-6) on the die
Two HMM random variables

- Observed sequence
  \[ X = X_1, X_2, \ldots, X_L \]
- Hidden state sequence
  \[ \pi = \pi_1, \pi_2, \ldots, \pi_L \]
- HMM:
  - Markov chain over hidden sequence
  - Dependence between \( \pi_i \) and \( X_i \)
The Parameters of an HMM

• as in Markov chain models, we have transition probabilities

\[ a_{kl} = \Pr(\pi_i = l \mid \pi_{i-1} = k) \]

probability of a transition from state \( k \) to \( l \)

\( \pi \) represents a path (sequence of states) through the model

• since we’ve decoupled states and characters, we also have emission probabilities

\[ e_k(b) = \Pr(X_i = b \mid \pi_i = k) \]

probability of emitting character \( b \) in state \( k \)
A Simple HMM with Emission Parameters

\( a_{13} \) probability of a transition from state 1 to state 3

\( e_2(A) \) probability of emitting character \( A \) in state 2
Simple HMM for Gene Finding

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences
Three Important Questions

• How likely is a given sequence?
  the Forward algorithm

• What is the most probable “path” (sequence of hidden states) for generating a given sequence?
  the Viterbi algorithm

• How can we learn the HMM parameters given a set of sequences?
  the Forward-Backward (Baum-Welch) algorithm
How Likely is a Given Path and Sequence?

- the probability that the path $\pi_1 \ldots \pi_L$ is taken and the sequence $X_1 \ldots X_L$ is generated:

$$\Pr(X_1 \ldots X_L, \pi_1 \ldots \pi_L) = a_0 \pi_1 a_{\pi_L} N \prod_{i=1}^{L-1} a_{\pi_i \pi_{i+1}} \prod_{i=1}^{L} e_{\pi_i} (X_i)$$

(assuming begin/end are the only silent states on path)
How Likely Is A Given Path and Sequence?

\[
\Pr(AAC,113) = a_{01} \times e_1(A) \times a_{11} \times e_1(A) \times a_{13} \times e_3(C) \times a_{35}
\]

\[
= 0.5 \times 0.4 \times 0.2 \times 0.4 \times 0.8 \times 0.3 \times 0.6
\]
How Likely is a Given Sequence?

• We usually only observe the sequence, not the path
• To find the probability of a sequence, we must sum over all possible paths

\[ \Pr(X_1 \ldots X_L) = \sum_{\pi} \Pr(X_1 \ldots X_L, \pi_1 \ldots \pi_L) \]

• but the number of paths can be exponential in the length of the sequence...
• the Forward algorithm enables us to compute this efficiently
Example

- Consider an candidate CpG island
  \textbf{CGCGC}
- Considering our HMM for CpG island model, some possible paths that are consistent with this CpG island are
  \textbf{C}^+G^+C^+G^+C^+
  C^-G^-C^-G^-C^-
  C^-G^+C^-G^+C^-
Number of paths

- for a sequence of length $L$, how many possible paths through this HMM are there?

- the Forward algorithm enables us to compute the probability of a sequence by efficiently summing over all possible paths
How Likely is a Given Sequence: The Forward Algorithm

- A dynamic programming solution
- subproblem: define $f_k(i)$ to be the probability of generating the first $i$ characters and ending in state $k$
- we want to compute $f_N(L)$, the probability of generating the entire sequence (x) and ending in the end state (state N)
- can define this recursively
The Forward Algorithm

- because of the Markov property, don’t have to explicitly enumerate every path

- e.g. compute $f_4(i)$ using $f_2(i-1)$, $f_4(i-1)$
The Forward Algorithm

- initialization:

\[ f_0(0) = 1 \]

probability that we’re in start state and have observed 0 characters from the sequence

\[ f_k(0) = 0, \quad \text{for } k \text{ that are not silent states} \]
The Forward Algorithm

- recursion for emitting states ($i = 1 \ldots L$):

$$f_l(i) = e_l(i) \sum_k f_k(i - 1)a_{kl}$$

- recursion for silent states:

$$f_l(i) = \sum_k f_k(i)a_{kl}$$
The Forward Algorithm

- termination:

\[ Pr(X) = Pr(X_1 \ldots X_L) = f_N(L) = \sum_k f_k(L) a_{kN} \]

probability that we’re in the end state and have observed the entire sequence.
• given the sequence $x = \text{TAGA}$
Forward Algorithm Example

• given the sequence $x = \text{TAGA}$
• initialization

$$f_0(0) = 1 \quad f_1(0) = 0 \quad \ldots \quad f_5(0) = 0$$

• computing other values

$$f_1(1) = e_1(T) \times (f_0(0) \times a_{01} + f_1(0) a_{11}) = 0.3 \times (1 \times 0.5 + 0 \times 0.2) = 0.15$$

$$f_2(1) = 0.4 \times (1 \times 0.5 + 0 \times 0.8)$$

$$f_1(2) = e_1(A) \times (f_0(1) \times a_{01} + f_1(1) a_{11}) = 0.4 \times (0 \times 0.5 + 0.15 \times 0.2)$$

$$\cdots$$

$$\Pr(\text{TAGA}) = f_5(4) = (f_3(4) \times a_{35} + f_4(4) a_{45})$$
Three Important Questions

• How likely is a given sequence?
• What is the most probable “path” for generating a given sequence?
• How can we learn the HMM parameters given a set of sequences?
Finding the Most Probable Path: The Viterbi Algorithm

- Dynamic programming approach, again!
- subproblem: define $v_k(i)$ to be the probability of the most probable path accounting for the first $i$ characters of $x$ and ending in state $k$

  - we want to compute $v_N(L)$, the probability of the most probable path accounting for all of the sequence and ending in the end state
  - can define recursively
  - can use DP to find $v_N(L)$ efficiently
Finding the Most Probable Path: The Viterbi Algorithm

• initialization:

\[ \nu_0(0) = 1 \]

\[ \nu_k(0) = 0, \quad \text{for } k \text{ that are not silent states} \]
The Viterbi Algorithm

• recursion for emitting states \((i = 1\ldots L)\):

\[
v_l(i) = e_l(x_i) \max_k [v_k(i - 1)a_{kl}]
\]

\[
ptr_l(i) = \arg \max_k [v_k(i - 1)a_{kl}]
\]

keep track of most probable path

• recursion for silent states:

\[
v_l(i) = \max_k [v_k(i)a_{kl}]
\]

\[
ptr_l(i) = \arg \max_k [v_k(i)a_{kl}]
\]
The Viterbi Algorithm

- termination:

$$\Pr(x, \pi^*) = \max_k \left( v_k(L) a_{kN} \right)$$

$$\pi_L^* = \arg \max_k \left( v_k(L) a_{kN} \right)$$

- traceback: follow pointers back starting at $\pi_L^*$
Forward & Viterbi Algorithms

• Forward/Viterbi algorithms effectively consider all possible paths for a sequence
  – Forward to find probability of a sequence
  – Viterbi to find most probable path
• consider a sequence of length 4…
Using HMM to detect CpG islands

- Recall the 8-state HMM for our CpG island
- Apply the Viterbi algorithm to a DNA sequence on this HMM
- Contiguous assignments of ‘+’ states will correspond to CpG islands
Summary

- Hidden Markov models are extensions to Markov chains enabling us to model and segment sequence data.
- HMMs are defined by a set of states and emission characters, transition probabilities and emission probabilities.
- We have examined two questions for HMMs:
  - Computing the probability of a sequence of observed characters given an HMM (Forward algorithm).
  - Computing the most likely sequence of states (or path) for a sequence of observed characters.
Next time

• How likely is a given sequence?
• What is the most probable “path” for generating a given sequence?
• How can we learn the HMM parameters given a set of sequences?