Quantitative Biology II
Lecture 1: Hidden Markov Models

February 26, 2015
Plan for Today

• Conditional independence and Markov models
• Hidden Markov Models (HMMs)
• Viterbi Algorithm
• Forward Algorithm
• Forward/Backward Algorithm
A Markov Model (Chain)

• Suppose $Z = (Z_1, \ldots, Z_L)$ is a sequence of cloudy ($Z_i = 0$) or sunny ($Z_i = 1$) days.

• We could assume days are iid with probability $\theta$ of sun but cloudy and sunny days occur in *runs*.

• We can capture the correlation between successive days by assuming a first-order Markov model:

$$P(Z_1, \ldots, Z_L) = P(Z_1)P(Z_2|Z_1)P(Z_3|Z_2) \cdots P(Z_L|Z_{L-1})$$

instead of complete independence:

$$P(Z_1, \ldots, Z_L) = P(Z_1) \cdots P(Z_L)$$
Factorization

• A joint distribution can always be factored as:
  \[ P(Z_1, \ldots, Z_L) = P(Z_1)P(Z_2|Z_1)P(Z_3|Z_1, Z_2) \cdots P(Z_L|Z_1, \ldots, Z_{L-1}) \]

• A first-order Markov model assumes:
  \[ P(Z_i|Z_1, \ldots, Z_{i-1}) = P(Z_i|Z_{i-1}) \]

• This is a step closer to the truth than complete independence, which amounts to:
  \[ P(Z_i|Z_1, \ldots, Z_{i-1}) = P(Z_i) \]

• Essentially, we assume conditional independence of \( Z_{i-1} \) and \( Z_{i+1} \) given \( Z_i \)

\[
\begin{array}{c c c c c}
\circ & \rightarrow & \circ & \cdots & \circ \\
Z_1 & Z_2 & & Z_L \\
\end{array}
\]
Markov Chain As A Sequence Generator

\[ P(Z_i = 1 | Z_{i-1} = 0) \]
\[ P(Z_i = 0 | Z_{i-1} = 0) \]
\[ P(Z_i = 1 | Z_{i-1} = 1) \]
\[ P(Z_i = 0 | Z_{i-1} = 1) \]

\[ P(Z_1 = 0) \]
\[ P(Z_1 = 1) \]

100000000000000001100000000000111000000000000..
Three Views

1. $P(z) = P(z_1) \prod_{i=2}^{L} a_{z_{i-1},z_i}$

   where $a_{c,d} = P(z_i = d | z_{i-1} = c)$

2. 

   $Z_1 \rightarrow Z_2 \rightarrow \cdots \rightarrow Z_L$

3. 

   $P(z_1 = 0)$ $a_{0,0}$ $a_{0,1}$ $a_{1,1}$

   $P(z_1 = 1)$ $a_{1,0}$
Process Interpretation

- Let’s add an end state and cap the sequence with $z_0 = B$, $z_{L+1} = E$, e.g. $z = B000011000E$

- This is a probabilistic machine that generates sequences of any length. It is a stochastic finite state machine and defines a grammar.

- We can now simply say: $P(z) = \prod_{i=0}^{L} a_{z_i, z_{i+1}}$

$P(z)$ is a probability distribution over all sequences (for given alphabet).
This Markov Chain Is:

- **Discrete.** It proceeds in discrete steps. Continuous-time or -space models are possible.

- **Finite.** It has a finite number of states. Infinite state spaces are possible.

- **Homogeneous.** The state transition probabilities do not depend on \( i \). In some cases they do.

- **First-order.** \( Z_i \) depends only on \( Z_{i-1} \). In an \( N \)th order model \( Z_i \) depends on \( Z_{i-N},...,Z_{i-1} \).
DNA Sequences as Markov Chains

\[ A = \begin{pmatrix}
  a_{A,A} & a_{A,C} & a_{A,G} & a_{A,T} \\
  a_{C,A} & a_{C,C} & a_{C,G} & a_{C,T} \\
  a_{G,A} & a_{G,C} & a_{G,G} & a_{G,T} \\
  a_{T,A} & a_{T,C} & a_{T,G} & a_{T,T}
\end{pmatrix} \]
Duration Distribution

• Suppose a state has a self transition. What is the distribution over “run lengths”?

• If $Y$ is the number of times state $X$ is visited, conditional on starting in $X$, then:

\[ P(Y = y) = (1 - a)^{y-1}a, \quad z \in \{1, 2, 3, \ldots \} \]

\[ = \text{Geom}(y; a) \]

• This means that: \[ E[Y] = \frac{1}{a} \]
Distribution After $k$ Steps

• The Markov matrix $\mathbf{A}$ defines $P(z_i = d \mid z_{i-k} = c)$ for all $c$, $d$ and $k=1$ step. Let $\mathbf{A}_k$ be the equivalent matrix for $k \geq 0$ ($\mathbf{A}_1 = \mathbf{A}$ and $\mathbf{A}_0 = \mathbf{I}$)?

• Notice that:

$$P(z_i = d \mid z_{i-2} = c) = \sum_b P(z_i = d \mid z_{i-1} = b)P(z_{i-1} = b \mid z_{i-2} = c)$$

$$a'_{c,d} = \sum_b a_{c,b}a_{b,d}$$

$$\mathbf{A}_2 = \mathbf{A} \times \mathbf{A} = \mathbf{A}^2$$

• Similarly,

$$P(z_i = d \mid z_{i-k} = c) = \sum_b P(z_i = d \mid z_{i-1} = b)P(z_{i-1} = b \mid z_{i-k} = c)$$

$$\mathbf{A}_k = \mathbf{A} \times \mathbf{A}_{k-1} \implies \mathbf{A}_k = \mathbf{A}^k$$
A Mathematical Trick

• In practice, we diagonalize $A$:

$$A = U \Lambda U^{-1}$$

where $U$ is a matrix of eigenvectors, $U^{-1}$ its inverse, and $\Lambda$ is a diagonal matrix of eigenvalues.

• Now,

$$A^k = (U \Lambda U^{-1})(U \Lambda U^{-1}) \cdots (U \Lambda U^{-1})(U \Lambda U^{-1})$$

$$= U \Lambda (U^{-1}U) \Lambda (U^{-1} \cdots U) \Lambda (U^{-1}U) \Lambda U^{-1}$$

$$= U \Lambda \Lambda I \cdots \Lambda I \Lambda U^{-1}$$

$$= U \Lambda^k U^{-1}$$

where:

$$\Lambda^k = \begin{pmatrix} 
\lambda_1^k & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \lambda_n^k 
\end{pmatrix}$$
Stationary Distribution

• Note that:

\[ P(z_k = d) = \sum_c P(z_1 = c)P(z_k = d | z_1 = c) \]

• Therefore, if \( \mathbf{v}_k \) is a (row) vector representing \( P(z_k = d) \) for all \( d \), and if \( \mathbf{b} \) is a (row) vector of “begin” probabilities, then:

\[ \mathbf{v}_k = \mathbf{b} \mathbf{A}^{k-1} \]

• The *stationary distribution* of the Markov chain is:

\[ \lim_{k \to \infty} \mathbf{v}_k \]

• It is given by the eigenvector associated with the largest eigenvalue of \( \mathbf{A} \)
A Hidden Markov Model

• Let $X = (X_1, ..., X_L)$ indicate whether AS bikes on day $i$ ($X_i = 1$) or not ($X_i = 0$)

• Suppose AS bikes on day $i$ with probability $\theta_0 = 0.25$ if it is cloudy ($Z_i = 0$) and with probability $\theta_1 = 0.75$ if it is sunny ($Z_i = 1$)

• Further suppose the $Z_i$s are hidden; we see only $X = (X_1, ..., X_L)$

• This hidden Markov model is a mixture model in which the $Z_i$s are correlated

• We call $Z = (Z_1, ..., Z_L)$ the path
Example

\[ P(x_i = 1|z_i = 0) = 0.25 \]
\[ P(x_i = 1|z_i = 1) = 0.75 \]

\[
\begin{array}{cccccccccccccccc}
X & = & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0
\end{array}
\]
Example

\[ P(x_i = 1|z_i = 0) = 0.25 \]
\[ P(x_i = 1|z_i = 1) = 0.75 \]

\[ Z = 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \]

\[ X = 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0 \]
Parameters of the Model

• Transition parameters: \( a_{s_1, s_2} \) for all \( s_1 \in S \cup \{B\}, \ s_2 \in S \cup \{E\} \)

• Emission parameters: \( e_{s, x} \) for all \( s \in S, \ x \in A \)

• The transition parameters define conditional distributions for state \( s_2 \) at position \( i \) given state \( s_1 \) at position \( i - 1 \): \( a_{s_1, s_2} = P(Z_i = s_2 | Z_{i-1} = s_1) \)

• The emission parameters define conditional distributions over observation \( x \) given state \( s \), both at position \( i \): \( e_{s, x} = P(X_i = x | Z_i = s) \)

• The observations can be anything!
Joint Probability

• \( X_i \)s are *conditionally independent* given \( Z_i \)s:

\[
P(x, z) = P(z)P(x|z) = P(z) \prod_{i=1}^{L} P(x_i|z_i)
\]

• \( Z_i \)s form a Markov chain:

\[
P(z) = P(z_1) \prod_{i=1}^{L} P(z_{i+1}|z_i)
\]

• Therefore,

\[
P(x, z) = P(z_1) \prod_{i=1}^{L} P(x_i|z_i)P(z_{i+1}|z_i)
\]

\[
= a_{B, z_1} \prod_{i=1}^{L} e_{z_i, x_i} a_{z_i, z_{i+1}}
\]
Distribution Over Seqs

• The likelihood of the model is:

\[
P(x) = \sum_z P(x, z) = \sum_z a_{B,z_1} \prod_{i=1}^L e_{z_i,x_i,a_{z_i},z_{i+1}}
\]

• \(P(x)\) is a probability distribution over sequences of length \(L\):

\[
\sum_x P(x) = \sum_x \sum_z a_{B,z_1} \prod_{i=1}^L e_{z_i,x_i,a_{z_i},z_{i+1}}
\]

\[
= \sum_{x_1} \cdots \sum_{x_L} \sum_{z_1} \cdots \sum_{z_L} a_{B,z_1} \prod_{i=1}^L e_{z_i,x_i,a_{z_i},z_{i+1}}
\]

\[
= \sum_{z_1} a_{B,z_1} \sum_{z_2} a_{z_1,z_2} \cdots \sum_{z_L} a_{z_L,E} \sum_{x_1} a_{z_1,E} \cdots \sum_{x_L} e_{z_L,x_L}
\]

\[
= 1 \cdot 1 \cdots 1 \cdot 1 \cdots 1 = 1
\]
Key Questions

- Given the model (parameter values) and a sequence $x$, what is the most likely path?
  $$
  \hat{z} = \arg \max_z P(z|x) = \arg \max_z P(x, z)
  $$

- What is the likelihood of the sequence?
  $$
  P(x) = \sum_z P(x, z)
  $$

- Other questions (later)
  - What is the posterior probability of $Z_i$ given $x$?
  - What is the maximum likelihood estimate of all parameters?
Example (Again)

\[ P(x_i = 1|z_i = 0) = 0.25 \]
\[ P(x_i = 1|z_i = 1) = 0.75 \]

\[
\begin{align*}
Z &= \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \\
X &= 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0
\end{align*}
\]
Example (Again)

\[ P(x_i = 1 | z_i = 0) = 0.25 \]
\[ P(x_i = 1 | z_i = 1) = 0.75 \]

\[ Z = 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \]

\[ X = 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \]
Graph Interpretation of Most Likely Path

\[ \hat{z} = \arg \max_{z} P(x, z) = \arg \max_{z} a_{B, z_1} \prod_{i=1}^{L} e_{z_i, x_i} a_{z_i, z_{i+1}} \]
Graph Interpretation of Probability of $\mathbf{x}$

$$P(\mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} a_{B, z_1} \prod_{i=1}^{L} e_{z_i, x_i} a_{z_i, z_{i+1}}$$
Viterbi Algorithm

• Let $v_{i,j}$ be the weight of the most likely path for $(x_1, ..., x_i)$ that ends in state $j$

• Base case: $v_{0,B} = 1$, $v_{i,B} = 0$ for $i > 0$

• Recurrence: $v_{i,j} = e_{x_i,j} \max_k v_{i-1,k} a_{k,j}$

• Termination: $P(x, \hat{z}) = \max_k v_{L,k} a_{k,E}$

Keep back-pointers for traceback, as in alignment (see Durbin)
Forward Algorithm

- Let $f_{i,j}$ be the (marginal) probability of $(x_1, \ldots, x_i)$ and $z_i = j$: $f_{i,j} = P(x_1, \ldots, x_i, z_i = j)$

- Base case: $f_{0,B} = 1$, $f_{i,B} = 0$ for $i > 0$

- Recurrence: $f_{i,j} = e_{x_i,j} \sum_k f_{i-1,k} a_{k,j}$

- Termination: $P(x) = \sum_k f_{L,k} a_{k,E}$
Why HMMs Are Cool

- Extremely general and flexible models for sequence modeling
- Efficient tools for *parsing* sequences
- Also proper probability models: allow maximum likelihood parameter estimation, likelihood ratio tests, etc.
- Inherently *modular*, accommodating of complexity
- In many cases, strike an ideal balance between simplicity and expressiveness
Some Applications In Bioinformatics

Krogh, Mian & Haussler, 1994
HMMs Generalize Motif Models
Krogh et al., 1994
Posterior Distributions

• In most applications, we want to make inferences about the path

• We know how to find the most likely path,
  \[ \hat{z} = \arg\max_z P(x, z) \]
  but what if many paths have high likelihood?

• From a Bayesian perspective, we want the full posterior distribution over paths,
  \[ P(z|x) = \frac{P(x, z)}{P(x)} \]

• We can calculate this for any given path, but how to characterize \( m^L \) paths?
Marginal Posteriors

- Consider the marginal posterior distribution over states for a particular position $i$:
  
  $$P(z_i|x) = \sum_{z_{\bar{i}}} P(z_i, z_{\bar{i}}|x)$$

- Even if computed for all $i$, doesn’t fully describe joint posterior, but provides a window into it

- Can be computed efficiently by dynamic programming
Graph Interpretation

\[ P(z_4 = 1|x) = \frac{P(x, z_4 = 1)}{P(x)} \]
Forward Algorithm

\[ f_{4,1} = P(x_1, \ldots, x_4, z_4 = 1) \]
Backward Algorithm

\[ b_{4,1} = P(x_5, \ldots, x_L | z_4 = 1) \]
Forward/Backward

\[ P(z_4 = 1|x) = \frac{P(x_1, \ldots, x_4, z_4 = 1)P(x_5, \ldots, x_L|z_4 = 1)}{P(x)} = \frac{f_{4,1}b_{4,1}}{P(x)} \]
Backward Algorithm

- Let $b_{i,j}$ be the (marginal) probability of $(x_{i+1}, \ldots, x_L)$ given $z_i = j$: $b_{i,j} = P(x_{i+1}, \ldots, x_L | z_i = j)$

- Base case: $b_{L,j} = a_{j,E}$ for all states $j$

- Recurrence: $b_{i,j} = \sum_k a_{j,k} e_{x_{i+1},k} b_{i+1,k}$

- Termination: $P(x) = \sum_k a_{B,k} e_{x_1,k} b_{1,k}$
Real-world Use
Other Uses of Marginal Posteriors

- Posterior expected values:

\[ G(U, V | x) = \sum_{i \in U} \sum_{j \in V} P(z_i = j | x) g(i, j) \]

- Posterior decoding:

\[ \hat{z}_i = \arg\max_k P(z_i = k | x) \]

- Warning: may not be a good path!

- Heuristic solution: feed through Viterbi
Underflow

• Problem: large products of probabilities are very small numbers

• Solution: take logs

\[ P(x, z) = a_{B, z_1} \prod_{i=1}^{L} e_{z_i, x_i} a_{z_i, z_{i+1}} \]

\[ \log P(x, z) = \log a_{B, z_1} + \sum_{i=1}^{L} \log e_{z_i, x_i} + \log a_{z_i, z_{i+1}} \]

• Viterbi is as before but with sums, e.g.,

\[ V_{i, j} = \log e_{x_i, j} + \max_k (V_{i-1, k} + \log a_{k, j}) \]

• Take logs of all quantities once, at beginning
Forward is Messier

\[ f_{i,j} = e_{x_{i,j}} \sum_k f_{i-1,k} a_{k,j} \]

\[ F_{i,j} = \log e_{x_{i,j}} + \log \left[ \sum_k \exp \left( F_{i-1,k} + \log a_{k,j} \right) \right] \]

\[ = \log e_{x_{i,j}} + \log \left[ \sum_k \exp(l_k) \right] \quad \text{[where } l_k = F_{i-1,k} + \log a_{k,j} \text{]} \]

\[ = \log e_{x_{i,j}} + \log \left[ \exp(l_{\max}) \left( 1 + \sum_{\text{other } k} \frac{\exp(l_k)}{\exp(l_{\max})} \right) \right] \]

\[ = \log e_{x_{i,j}} + l_{\max} + \log \left[ 1 + \sum_{\text{other } k} \exp(l_k - l_{\max}) \right] \]
Duration Modeling

• As discussed, a state with a self transition induces a geometric distribution over feature lengths

• A similar phenomenon occurs with any cycle in the state-transition graph

• This is effectively a prior—it can be overridden by the data, but it will influence predictions

• If features have other length distributions, may lead to suboptimal performance
Non-geometric Lengths

\[ P(\text{length } l) = (1 - \mu)^{l-1} \mu \]

\[ P(\text{length } l) = \binom{l - 1}{n - 1} (1 - \mu)^{l-n} \mu^n \]
Other Issues

• Training
• Generalized HMMs
• Higher-order emission probabilities
• Silent states, tying of states, and other application-specific issues
• ...

That’s All

• Best textbook on HMMs: “Biological Sequence Analysis” by Durbin, Eddy, Krogh, and Mitchison

• See also [Rabiner 1989]

• Phylo-HMMs: see [Siepel & Haussler, 2005]