CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Apr 6

- Recap posterior decoding and mean
- Parameter estimation
- Baum-Welch algorithm (EM for HMM)

Recap: Forward-Backward, posterior decoding, and posterior mean

Goal: compute the posterior probability of being in state *k* at step *i*

- ***Posterior probability***: probability of an "unknown" given observed (known) data

$$P(z_i = k|\vec{x}) = \frac{P(\vec{x}, z_i = k)}{P(\vec{x})}$$

Goal: compute the posterior probability of being in state *k* at step *i*

- ***Posterior probability***: probability of an "unknown" given observed (known) data

$$P(z_i = k|\vec{x}) = \frac{P(\vec{x}, z_i = k)}{P(\vec{x})}$$

- Aside: we can rewrite the numerator to include a prior (in a Bayesian setting)

Likelihood of data given an unknown

$$= \frac{P(z_i = k) \cdot P(\vec{x}|z_i = k)}{P(\vec{x})}$$

Prior

For us we will us *P*(***x***) as the (total) likelihood

Goal: compute the posterior probability of being in state *k* at step *i*

Focus on the numerator

- ***Posterior probability***: probability of an "unknown" given observed (known) data

$$P(z_i = k|\vec{x}) = \frac{P(\vec{x}, z_i = k)}{P(\vec{x})}$$

Likelihood of data given an unknown

- Aside: we can rewrite the numerator to include a prior (in a Bayesian setting)

$$= \frac{P(z_i = k) \cdot P(\vec{x}|z_i = k)}{P(\vec{x})}$$

Prior

For us we will us *P*(**x**) as the (total) likelihood

Goal: compute the posterior probability of being in state *k* at step *i*

$$P(\vec{x}, z_i = k) = P(x_1, \cdots x_i, z_i = k, x_{i+1}, \cdots x_L)$$

Break up emitted sequence around step *i*

Goal: compute the posterior probability of being in state *k* at step *i*

$$P(\vec{x}, z_i = k) = P(x_1, \cdots x_i, z_i = k, x_{i+1}, \cdots x_L)$$

Break up emitted sequence around step *i*

$$= P(x_1, \cdots x_i, z_i = k) \cdot P(x_{i+1}, \cdots x_L | x_1, \cdots x_i, z_i = k)$$

Use conditional probability

Goal: compute the posterior probability of being in state *k* at step *i*

$$P(\vec{x}, z_i = k) = P(x_1, \cdots x_i, z_i = k, x_{i+1}, \cdots x_L)$$

Break up emitted sequence around step *i*

$$= P(x_1, \cdots x_i, z_i = k) \cdot P(x_{i+1}, \cdots x_L | x_1, \cdots x_i, z_i = k)$$

Use conditional probability

$$= P(x_1, \cdots x_i, z_i = k) \cdot P(x_{i+1}, \cdots x_L | z_i = k)$$

Use Markov property

Goal: compute the posterior probability of being in state *k* at step *i*

$$P(\vec{x}, z_i = k) = P(x_1, \cdots x_i, z_i = k, x_{i+1}, \cdots x_L)$$

Break up emitted sequence around step *i*

$$= P(x_1, \cdots x_i, z_i = k) \cdot P(x_{i+1}, \cdots x_L | x_1, \cdots x_i, z_i = k)$$

Use conditional probability

$$= P(x_1, \cdots x_i, z_i = k) \cdot P(x_{i+1}, \cdots x_L | z_i = k)$$

Use Markov property

$$= f_k(i) \cdot b_k(i)$$

Define these two pieces as the forward and backward probabilities

Forward

Backward

Posterior decoding and posterior mean

Now we can complete the posterior probability:

$$P(z_i = k|\vec{x}) = \frac{f_k(i) \cdot b_k(i)}{P(\vec{x})}$$

Add last column of the forward DP table to get *P*(***x***):

$$P(\vec{x}) = \sum_k f_k(L)$$

Posterior decoding and posterior mean

Now we can complete the posterior probability:

$$P(z_i = k|\vec{x}) = \frac{f_k(i) \cdot b_k(i)}{P(\vec{x})}$$

Add last column of the forward DP table to get *P*(***x***):

$$P(\vec{x}) = \sum_k f_k(L)$$

<u>Posterior decoding</u>:

- Create a *K*x*L* table for the posterior probabilities
- Take the max over each column to find the posterior decoding

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

Posterior decoding and posterior mean

Now we can complete the posterior probability:

$$P(z_i = k|\vec{x}) = \frac{f_k(i) \cdot b_k(i)}{P(\vec{x})}$$

Add last column of the forward DP table to get *P(**x**)*:

$$P(\vec{x}) = \sum_k f_k(L)$$

Posterior decoding:

- Create a *K*x*L* table for the posterior probabilities
- Take the max over each column to find the posterior decoding

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

$$\bar{g}_i = \sum P(z_i = k|\vec{x}) \cdot g(k)$$

Posterior mean:

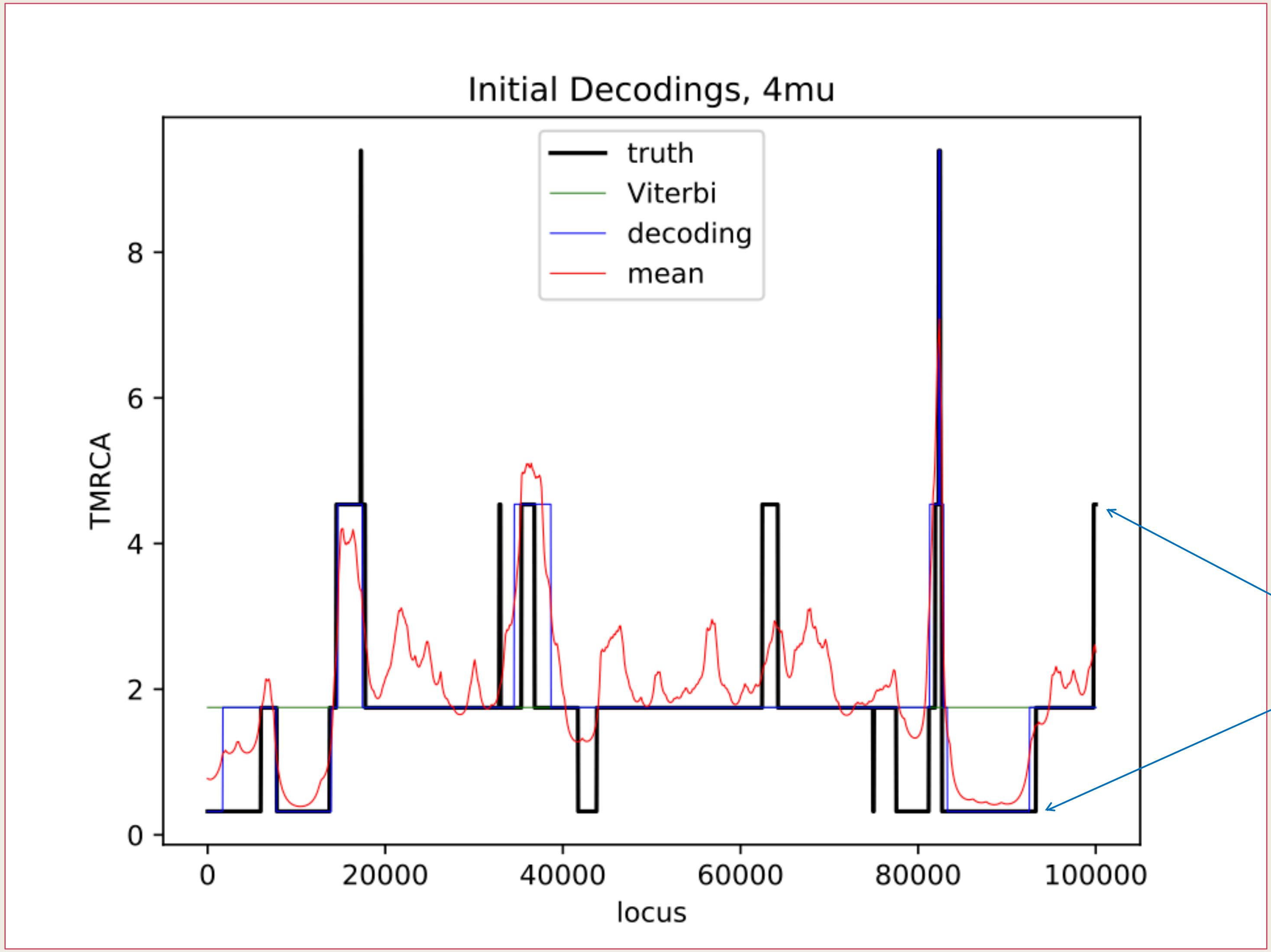
- The sum of each column in the posterior probability table should be 1
- Multiply these probabilities by the value of each state *g(k)* to get the mean

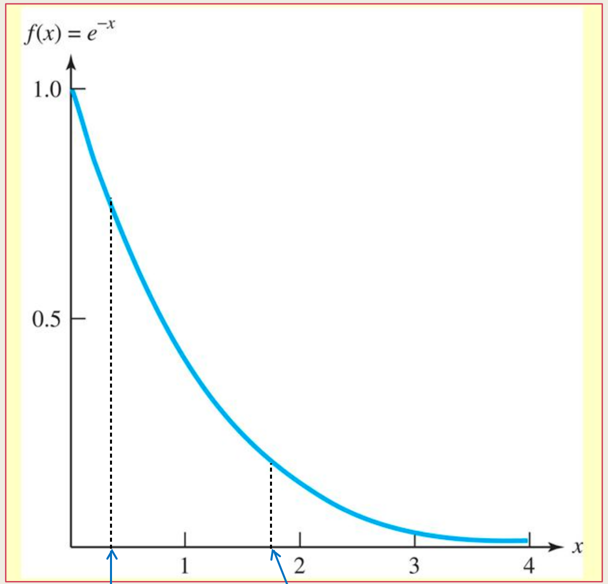
Summary (Part 1 of Lab 8)

- Now we have 3 different ways to estimate the hidden states:
- 1) Viterbi traceback (from the max of last column)
- 2) Posterior decoding (max of posterior probability table from Forward-Backward algorithm)
- 3) Posterior mean (weighted average over each column of the posterior probability table)

- Note: the posterior mean already reflects the “meaning” of each hidden state. For the other two, we need to transform the state sequence (indices of hidden state) using a 1-1 mapping from state index to state value

Lab 8 example: T_{mrca} for n=2





g(0) = 0.32, g(1) = 1.75
g(2) = 4.54, g(3) = 9.40

Exponential distribution image: Modified from SlidePlayer

Parameter Estimation for HMMs

Case 1

know: $\boxed{\vec{x}, \vec{z}, K, B}$

don't know: parameters

↳ transition probs

↳ emission probs

↳ (initial probs)

$K=2, B=2$

$\vec{z} = [1 \to 0 \to 0 \to 1 \to 0 \to 1 \to 1 \to 0 \to 0 \to 0]$

$\vec{x} = [0 \; 1 \; 0 \; 0 \; 1 \; 0 \; 1 \; 1 \; 1 \; 0]$

A_{kl} = # of transitions from $k \to l$

$E_k(b)$ = # of times we emit b from state k

es

estimate probabilities

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{\boxed{kl'}}}, \qquad e_k(b) = \frac{E_k(b)}{\sum_{b'} e_k(b')}$$

$$k \longrightarrow l \qquad\qquad k \downarrow b$$

Handout 24

end

A	0	1
start 0	3	2
1	3	1

emission

E	0	1
hidden state 0	2	4
1	3	1

a	0	1
0	3/5	2/5
1	3/4	1/4

e	0	1
0	2/6	4/6
1	3/4	1/4

goal output

(4)

what if both 0? → denominator is 0!

pseudo counts

$A_{k\ell} = \#\{k \to \ell\} + 1$

$E_k(b) = \#\{k \xrightarrow[\text{emit}]{} b\} + 1$

E				
k	1	1	1	1

→

e				
k	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

(5) what about π?

$$\pi_0 = \frac{\#\text{ of state } 0}{\underbrace{\#\text{ of state } 0 + \#\text{ of state } 1}_{}}$$

L

$$= \frac{6}{10}, \quad \pi_1 = \frac{4}{10}$$

look at z_1 → $\pi_0 = 0$

$\pi_1 = 1$

↳ for many separate sequences

$\vec{z}$ N → AV → Y

$\vec{x}$ Cat is runnin

B=2

K=4

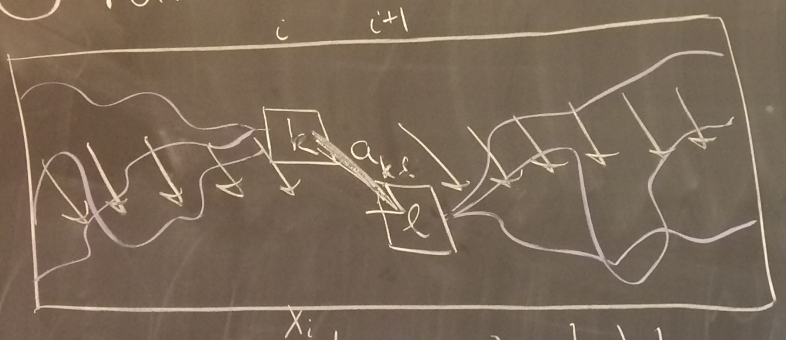
Case 2:

Only know $\vec{x}$

& K & B

1. Set a & e matrices arbitrarily or based on prior

2. run forward / backward



post prob table

ning.

transitions

$$P(z_i=k, z_{i+1}=l \mid \vec{x})$$

$$= \frac{f_k(i)\, a_{kl}\, e_l(x_{i+1})\, b_l(i+1)}{P(\vec{x})}$$

$$\boxed{A_{kl} = \sum_i P(z_i=k, z_{i+1}=l \mid \vec{x})}$$

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

emis

E

emissions

$$\boxed{E_k(b) = \sum_{\{i : x_i = b\}} \frac{f_k(i)\, b_k(i)}{P(\bar{x})}}$$

only want the steps where I observe b