Implementing Baum-Welch Algorithm

Assignment 3

Caitlin Ross
Computer Science Department, Rensselaer Polytechnic Institute
rossc3@rpi.edu

ABSTRACT

Hidden Markov models (HMM) can be used to find CpG islands in an unannotated sequence. Here we build a HMM using the Baum-Welch algorithm to estimate the transition and emission probabilities. We have been given 10 DNA sequences that were generated using a two state HMM. We implement the Baum-Welch algorithm and run the resulting program on these 10 sequences. We find that the algorithm does appear to correctly estimate the transition and probability matrices.

1. PROBLEM STATEMENT

Hidden Markov models (HMM) can be used to find CpG islands in an unannotated sequence. The Markov chain models that were explored in the last homework can be used to create a single model that can find CpG sequences of variable length. Here we use a two state HMM. For the HMM, we need to find the transition and emission probabilities. The transition probabilities are the probabilities of either remaining in state or switching to the other state . The emission probabilities are the probability of seeing a given symbol in that state. The problem here is to implement the Baum-Welch algorithm in order to estimate the transition and emission probabilities of the given data set.

2. METHODS

This section describes the program developed to solve this problem and the math needed.

For our HMM we need to determine the transition probabilities a_{kl}, which is the probability of transition from state k to state l and is defined as

$$a_{kl} = P(\pi_i = l|\pi_{i-1} = k) \tag{1}$$

We also need to find the emission probabilities of each symbol b when seen in state k, which is defined as

$$e_k(b) = P(x_i = b|\pi_i = k) \tag{2}$$

The Baum-Welch depends on two other algorithms to assist in finding these probabilities. The forward algorithm uses dynamic programming to find the probability that some state path could have produced the given sequence. We find a matrix f, where each row represents a state. We initialize $f_0(0) = 1$ and $f_1(0) = 0$. Then for each state l, we have that

$$f_l(i) = e_l(x_i) \sum_k f_k(i-1)a_{kl} \tag{3}$$

for $i = 1...L$ where L is the length of the sequence. The termination condition is then

$$P(x) = \sum_k f_k(L)a_{k0} \tag{4}$$

where a_{k0} represents a transition from state k into the end state.

The backward algorithm is similar but finds the posterior probabilities starting from the end of the sequence. In this case, we find a matrix b, again where each row represents a state. We initialize $b_k(L) = a_{k0}$ for each state k. For each state k, we have that

$$b_k(i) = \sum_l a_{kl}e_l(x_{i+1})b_l(i+1) \tag{5}$$

and that the termination condition is

$$P(x) = \sum_l a_{0l}e_l(x_1)b_l(1) \tag{6}$$

where a_{0l} is the probability to start in state l.

For the Baum-Welch, we start off with random numbers generated for the emission probabilities, normalizing the values for each state. For each sequence in the assign3.fasta file provided, we calculate $f_k(i)$ and $b_k(i)$ and add those contributions to the matrices A and E. A_{kl} represents the expected number of times that a_{kl} is used and is formally defined as

$$A_{kl} = \sum_j \frac{1}{P(x^j)} \sum_i f_k^j(i)a_{kl}e_l(x_{i+1}^j)b_l^j(i+1) \tag{7}$$

$E_k(b)$ is the expected number of times that letter b appears in state k and is defined as

$$E_k(b) = \sum_j \frac{1}{P(x^j)} \sum_{\{i|x_i^j=b\}} f_k^j(i)b_k^j(i) \tag{8}$$

Once that is done for each sequence in the data set, the new

Table 1: Transition Matrix

	State 1	State 2
State 1	0.93885	0.06115
State 2	0.10666	0.89334

Table 2: Emission Matrix

	A	C	G	T
State 1	0.40383	0.08516	0.07901	0.43199
State 2	0.12853	0.39166	0.35570	0.12411

model parameters can be calculated by

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \tag{9}$$

and

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \tag{10}$$

Finally the log likelihood of the model can be calculated by

$$\sum_{j \in sequences} \log P(x_j) \tag{11}$$

where $P(x_j)$ is the value returned from the termination step of the forward algorithm for sequence j.

The program uses the file `assign3.fasta`, which contains 10 DNA sequences that are each 300 bases long. We know that the sequences were generated using a two state HMM, where one state is G-C rich and the other is A-T rich. The program reads in this file and runs it through the Baum-Welch algorithm just described.

3. RESULTS

After running the program, we end with a log likelihodd of -4109.0229. The estimated transition and emission matrices are shown in Tables 1 and 2, respectively. As seen in the table, the HMM has correctly estimated that one state is A-T rich, and the other is G-C rich. State 1 appears to be the A-T rich state and State 2 appears to be G-C rich.

4. PROBLEMS ENCOUNTERED

I did encounter some problems understanding the notation. I followed the book closely, but some parts were confusing. For instance, to calculate f, that starts with a 0 index, but the book assumes that the sequence is initially indexed at 1, so I encountered a lot of indexing issues. I might still have some slight error in my code, but I appear to be close as the matrices do tend to go to expected values.