Implementation of EM algorithm in HMM training

Jens Lagergren

April 22, 2009

EM Algorithms: An expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on hidden variables. EM is a 2 step iterative method. In the first step(E-step) it computes an expectation of the log likelihood with respect to the current estimate of the distribution for the hidden variables. In second step (M-step), it computes the parameters which maximize the expected log likelihood. In each cycle, it uses the aproximations from previous step to aproximate the solution progressively[wikipedia.com]. In BioInformatics, EM is commonly used to generate HMM's from selected genetic sequences.

Training Hidden Markov Models using EM

HMM profiles are usually trained with a set of sequences that are known to belong to a single family. Once the HMM has been trained and optimized to identify these known members, it is then used to clasify unknown genes. Namely, this requires us to construct a HMM, M, that generates a collection of sequences F with the maximum probability:

$$\max_{\text{HMMs}M} \prod_{x \in F} \Pr[x|M] \tag{1}$$

But this can be computationally hard. Therefore we break down the problem into following sub-problems

- Finding the right structure, i.e. determining the number of layers. (We will not consider this sub-problem at the moment)
- Finding optimal parameters for a given structure.

To solve the latter problem we now let

- $A_{\pi\pi'}$ be the expected number of $\pi \to \pi'$ transitions.
- A_π be the expected number of visits to π.
- $G_{\pi,\sigma}$ be the expected number of times σ is generated from state π.
- θ be the set of parameters(hidden variables to be optimized) composed of:
  - $e_\pi(\sigma)$ is the probability to generate the symbol σ in state π.
  - $a_{\pi\pi'}$ is the transition probability from state π to π'.

We also continue with the notation $X^i := x_1, \ldots, x_i$ and $\Pi^i = \pi_1, \ldots, \pi_i$, from previous notes.

Expectation Maximization

We will use the EM algorithm to optimize the parameters to create the best fitting HMM from a given set of sequences F. From the initial HMM M and the initial parameter set θ we can generate a new optimized parameter set: θ', by iterating this procedure the aproximation will eventually converge with the local maximum. We calculate this new set θ' as follows

$$\begin{aligned} a'_{\pi\pi'} &= \frac{\sum\limits_{x \in F} E[A_{\pi\pi'} \mid x, \theta]}{\sum\limits_{x \in F} E[A_{\pi} \mid X, \theta]} \\ e'_{\pi}(\sigma) &= \frac{\sum\limits_{x \in F} E[G_{\pi,\sigma} \mid X, \theta]}{\sum\limits_{x \in F} E[A_{\pi} \mid X, \theta]} \end{aligned}$$

We can compute each nominator and denominator separately:

1. $E[A_\pi|X,\theta]$ (denominator)
2. $E[G_{\pi,\sigma}|X,\theta]$ (nominator for $e'_\pi(\sigma)$)
3. $E[A_{\pi\pi'}|X,\theta]$ (nominator for $a'_{\pi\pi'}$)

1. Calculating the denominator $E[A_\pi|X,\theta]$:
Notice that:

$$E[A_\pi|X,\theta] = \sum_{\pi'} E[A_{\pi\pi'}|X,\theta]$$

Therefore $E[A_\pi|X,\theta]$ is easily computed once $E[A_{\pi,\pi'}|X,\theta]$ is resolved in point 3.

2. Calculating the nominator:$E[G_{\pi,\sigma}|X,\theta]$

$$\begin{aligned} E[G_{\pi,\sigma}|X,\theta] &= \sum_{i, x_i = \sigma} Pr[\pi_i = \pi|X,\theta] \\ &= \sum_{i, x_i = \sigma} \frac{Pr[\pi_i = \pi, X|\theta]}{Pr[x|\theta]} \\ &= \sum_{i, x_i = \sigma} \sum_{\pi'} \frac{Pr[\pi_{i-1} = \pi', \pi_i = \pi, X|\theta]}{Pr[x|\theta]} \end{aligned}$$

3. Calculating $E[A_{\pi\pi'}|X,\theta]$

$$\begin{aligned} E[A_{\pi\pi'}|X,\theta] &= \sum_i \Pr[\pi_i = \pi, \pi_{i+1} = \pi'|X,\theta] \\ &= \frac{\sum_i \Pr[\pi_i = \pi, \pi_{i+1} = \pi', X|\theta]}{\Pr[X|\theta]} \end{aligned}$$

so it is enough to be able to compute $\Pr[\pi_i = \pi, \pi_{i+1} = \pi', X|\theta]$

$$
\begin{aligned}
& \Pr[\pi_i = \pi, \pi_{i+1} = \pi', X|\theta] \\
= \; & \Pr[\pi_i = \pi, \pi_{i+1} = \pi', X|X^i, \pi_i = \pi, \theta] \Pr[X^i, \pi_i = \pi|\theta] \\
= \; & \underbrace{\Pr[x_{i+1}, \ldots, x_n, \pi_{i+1} = \pi'|\pi_i = \pi, \theta]}_{\text{The Markov property!}} \underbrace{\Pr[X^i, \pi_i = \pi|\theta]}_{f_\pi(i)} \\
= \; & \Pr[x_{i+1}, \ldots, x_n|\pi_{i+1} = \pi', \pi_i = \pi, \theta] \underbrace{\Pr[\pi_{i+1} = \pi'|\pi_i = \pi, \theta]}_{a_{\pi\pi'}} f_\pi(i) \\
= \; & \underbrace{\Pr[x_{i+2}, \ldots, x_n|\pi_{i+1} = \pi', \theta]}_{b_\pi(i+1)} \underbrace{\Pr[x_{i+1}|\pi_{i+1} = \pi']}_{e_{\pi'}(x_{i+1})} a_{\pi\pi'} f_\pi(i) \\
= \; & b_\pi(i+1) e_{\pi'}(x_{i+1}) a_{\pi\pi'} f_\pi(i)
\end{aligned}
$$

... where
1. $b_\pi(i+1)$ is the "backward" variable defined as:

$$b_\pi(i) = \Pr[x_{i+1}, \ldots, x_n|\pi_i = \pi, \theta]$$

which can be computed using dynamic programming in a similar way as $f_\pi(i)$ was computed in lecture 5.
2. $f_\pi(i)$ is the "forward" variable defined in lecture 5.
3. $e_{\pi'}(x_{i+1})$ and $a_{\pi\pi'}$ were already calculated in the last cycle.

Stopping the iteration

When to stop? Notice that for θ' given by $a'_{\pi\pi'}$ and $e'_\pi(\sigma)$ either

- θ' are the locally optimal parameters, in which case you may stop the algorithm.
- or $\prod_{x \in F} \Pr[x|\theta'] > \prod_{x \in F}[x|\theta]$, which means a solution has not been reached yet.

Time analysis of EM alorithm in HMM Training Problem

A full time analysis of the complete EM algorithm would need to analyze the number of iterations required for convergence which may vary in different cases, nevertheless one may estimate the time required to calculate some of its components, Ie:

Claim 1. $\max_{\pi_1, \ldots, \pi_n}$ *$Pr[x_1, \ldots, x_n, \pi_0, \ldots, \pi_n]$ can be computed in time $O(|Q|^2 l)$.*

Proof: Let

$$v_\pi(i) = \max_{\pi_1,\ldots,\pi_{i-1}} \Pr[x_1,\ldots,x_i,\pi_0,\ldots,\pi_{i-1},\pi_i=\pi]$$

and

$$v_\pi(0) = \begin{cases} 1 & \text{when } \pi = q_{start} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned}
v_\pi(i) &= \max_{\pi_1,\ldots,\pi_{i-1}} \Pr[x_1,\ldots,x_i,\pi_0,\ldots,\pi_{i-1},\pi_i=\pi] \\
&= \max_{\pi'} \max_{\pi_1,\ldots,\pi_{i-1}} \Pr[X^i,\Pi^{i-2},\pi_{i-1}=\pi',\pi_i=\pi] \\
&= \max_{\pi'} \max_{\pi_1,\ldots,\pi_{i-1}} \Pr[X^i,\Pi^{i-2},\pi_{i-1}=\pi',\pi_i=\pi | X^{i-1},\Pi^{i-2},\pi_0,\pi_{i-1}=\pi'] \\
&\quad \Pr[X^{i-1},\Pi^{i-2},\pi_0,\pi_{i-1}=\pi'] \\
&= \max_{\pi'} \max_{\pi_1,\ldots,\pi_{i-1}} \underbrace{\Pr[x_i,\pi_i=\pi|\pi_{i-1}=\pi_{i-1}]}_{\text{The Markov property!}} \\
&\quad \Pr[X^{i-1},\Pi^{i-2},\pi_0,\pi_{i-1}=\pi'] \\
&= \max_{\pi'} \max_{\pi_1,\ldots,\pi_{i-1}} \Pr[x_i,\pi_i=\pi]\Pr[\pi_i=\pi|\pi_{i-1}=\pi_{i-1}] \\
&\quad \Pr[X^{i-1},\Pi^{i-2},\pi_0,\pi_{i-1}=\pi'] \\
&= \underbrace{\Pr[x_i,\pi_i=\pi]}_{e_\pi(x_i)} \max_{\pi'} \left(\underbrace{\Pr[\pi_i=\pi|\pi_{i-1}=\pi_{i-1}]}_{a_{\pi'\pi}} \underbrace{\max_{\pi_1,\ldots,\pi_{i-1}} \Pr[X^{i-1},\Pi^{i-2},\pi_0,\pi_{i-1}=\pi']}_{v_{\pi'}(i-1)} \right) \\
&= e_\pi(x_i) \max_{\pi'} a_{\pi'\pi} v_{\pi'}(i-1)
\end{aligned}$$

That is,

$$v_\pi(i) = \max_{\pi'} v_{\pi'}(i-1) a_{\pi'\pi} e_\pi(x_i)$$

Since there are $|Q|n$ subproblems $V_\pi(i)$ and each can be computed in time $O(|Q|)$ this gives a recursion that can be computed in time $O(|Q|^2 n)$. ■