Algorithmic Bioinformatics DD2450, spring 2010, Lecture 7-8

Lecturer Jens Lagergren
Several current and previous students
will be acknowledged in a separate document.

May 6, 2010

Chapter 1

Training of HMMs using EM

Recall

$$x^a y^b \text{ where } x + y = 1, \quad 0 \leq x, y, \leq 1$$

is maximized by $x = \frac{a}{a+b}$ and $y = \frac{b}{a+b}$.

In general,

$$\prod_{i=1}^{n} x_i^{a_i} \text{ where } \sum_{i=1}^{n} = 1, \quad 0 \leq x_i \leq 1$$

is maximized by $x_i = \frac{a_i}{\sum_{i=1}^{n} a_i}$.

In other words we are setting probabilities to normalized powers. Note that maximizing $\prod_{i=1}^{n} x_i^{a_i}$ is equivalent to maximizing

$$\log \left(\prod_{i=1}^{n} x_i^{a_i} \right) = \sum_{i=1}^{n} a_i \log x_i.$$

1.1 Framework

We have a fixed set of states Q and an alphabet Σ. Recall that training HMM is an iterative process, which means that we want to improve the transition and emission probabilities iteratively so that in each step the likelihood for a given family F is improved.

- Old parameters will be θ, i.e. $\{e_\pi(\sigma)\}$ and $\{a_{\pi\pi'}\}$.
- New parameters will be θ', i.e. $\{e'_\pi(\sigma)\}$ and $\{a'_{\pi\pi'}\}$.

As before we use the following notation

- $A_{\pi,\pi'}$ = number of transitions from π to π'.
- A_π = number of visits to the state π.
- $G_{\pi,\sigma}$ = number of times σ is generated when visiting π.
- A path (through the model) is denoted z. Note that z is a hidden variable, i.e., z is never observed, although it is generated by the process..
- A generated sequence is denoted x. It is an observable variable.

1.2 The iterative process

One step of the iterative procedure is performed as follows:

$$a'_{\pi\pi'} = \frac{\sum_{x\in F} \mathrm{E}\left[A_{\pi,\pi'}|x,\theta\right]}{\sum_{x\in F} \mathrm{E}\left[A_{\pi}|x,\theta\right]}$$

$$e'_{\pi}(\sigma) = \frac{\sum_{x\in F} \mathrm{E}\left[G_{\pi,\sigma}|x,\theta\right]}{\sum_{x\in F} \mathrm{E}\left[A_{\pi}|x,\theta\right]}$$

Lectures 5 and 6 show how $a'_{\pi,\pi'}$ and $e'_\pi(\sigma)$ can be computed using dynamic programming.

In the iterative procedure, if

$$\prod_{x\in F} \Pr\left[x|\theta'\right] > \prod_{x\in F} \Pr\left[x|\theta\right]$$

i.e., if the new parameters improves the likelihood for generating F, then we set $\theta \leftarrow \theta'$ and continue, otherwise we stop.

Next For ease of notation, we assume $F = \{x\}$, and show that one step is consistent with setting

$$a'_{\pi\pi'} = \frac{\mathrm{E}\left[A_{\pi,\pi'}|x,\theta\right]}{\mathrm{E}\left[A_{\pi}|x,\theta\right]}$$

$$e'_{\pi}(\sigma) = \frac{\mathrm{E}\left[G_{\pi,\sigma}|x,\theta\right]}{\mathrm{E}\left[A_{\pi}|x,\theta\right]}$$

We want to maximize $\Pr\left[x|\theta'\right]$ or equivalently $\log\Pr\left[x|\theta'\right]$. In order to do this, we apply a special case of Jensen's inequality, namely,

$$\log \mathrm{E}\left[f(x)\right] \geq \mathrm{E}\left[\log f(x)\right]$$

where x is a random variable. The inequality holds due to the fact that the logarithm is a concave function.

1.3 General derivation of EM

Let us use the special case of Jensen's inequality for the general derivation of EM. As before, let x be the observed data and z the hidden data. We have

$$\begin{aligned}
\log \Pr[x|\theta'] &= \log \sum_{z \in Q^{|x|}} \Pr[x, z|\theta'] \\
&= \log \sum_z \Pr[z|x,\theta] \frac{\Pr[x,z|\theta']}{\Pr[z|x,\theta]} \\
&= \log E_z \left[\frac{\Pr[x,z|\theta']}{\Pr[z|x,\theta]} \mid x, \theta \right] \\
&\geq^{\text{Jensen}} E_z \left[\log \frac{\Pr[x,z|\theta']}{\Pr[z|x,\theta]} \mid x, \theta \right] \\
&= \sum_z \Pr[z|x,\theta] \log \frac{\Pr[x,z|\theta']}{\Pr[z|x,\theta]} \\
&= \sum_z \Pr[z|x,\theta] \log \Pr[x,z|\theta'] - \sum_z \Pr[z|x,\theta] \log \Pr[z|x,\theta] \\
&= Q(\theta';\theta) - R(\theta;\theta)
\end{aligned}$$

where we define Q and R as

$$Q(\theta';\theta) = \sum_z \Pr[z|x,\theta] \log \Pr[x,z|\theta']$$

$$R(\theta;\theta) = \sum_z \Pr[z|x,\theta] \log \Pr[z|x,\theta]$$

Moreover, as is easy to show, we have that

$$\log \Pr[x|\theta] = Q(\theta;\theta) - R(\theta;\theta)$$

which yields the following implication

$$Q(\theta';\theta) > Q(\theta;\theta) \Rightarrow \log \Pr[x|\theta'] > \log \Pr[x|\theta]$$

which is what we are looking for. The reason for introducing Q is that Q is easy to maximize. What we want to do is maximize $Q(\theta';\theta)$ with respect to the new parameters θ'.

1.3.1 EM-algorithm for HMM training

$$\Pr[x,z|\theta'] = \prod_{\substack{\pi \in Q \\ \sigma \in \Sigma}} e'_\pi(\sigma) G^{x,z}_{\pi,\sigma} \prod_{\pi,\pi' \in Q} a'_{\pi\pi'} A^z_{\pi,\pi'}$$

where

$$G^{x,z}_{\pi,\sigma} = \# \text{ of times } \sigma \text{ is generated in state } \pi \text{ for x and z}$$

$$A^{z}_{\pi,\pi'} = \# \text{ of } \pi \to \pi' \text{ transitions in z}$$

We get

$$\begin{aligned}
Q(\theta';\theta) &= \sum_z \Pr[z|x,\theta] \log \Pr[x,z|\theta'] \\
&= \sum_z \Pr[z|x,\theta] \left(\sum_{\pi,\sigma} G^{x,z}_{\pi,\sigma} \log e'_\pi(\sigma) + \sum_{\pi,\pi'} A^z_{\pi,\pi'} \log a'_{\pi\pi'} \right) \\
&= \sum_{\pi,\sigma} \left(\sum_z \Pr[z|x,\theta] G^{x,z}_{\pi,\sigma} \right) \log e'_\pi(\sigma) + \sum_{\pi,\pi'} \left(\sum_z \Pr[z|x,\theta] A^z_{\pi,\pi'} \right) \log a'_{\pi\pi'} \\
&= \sum_{\pi,\sigma} E[G_{\pi,\sigma}|x,\theta] \log e'_\pi(\sigma) + \sum_{\pi,\pi'} E[A_{\pi,\pi'}|x,\theta] \log a'_{\pi\pi'}
\end{aligned}$$

Note that the first sum only depends on the emission probabilities and that the second sum only depends on the transition probabilities. We have that transitions probabilities from π are dependent and that emission probabilities for π are dependent. All other probabilities are independent.

This means that $Q(\theta$'; $\theta)$ is maximized by our $a'_{\pi\pi'}$ and $e'_\pi(\sigma)$ as before, that is

$$a'_{\pi,\pi'} = \frac{\mathrm{E}\left[A_{\pi,\pi'}|x,\theta\right]}{\mathrm{E}\left[A_\pi|x,\theta\right]}$$

$$e'_\pi(\sigma) = \frac{\mathrm{E}\left[G_{\pi,\sigma}|x,\theta\right]}{\mathrm{E}\left[A_\pi|x,\theta\right]}$$

1.3.2 Computing the required probabilities

We want to compute $E[A_{\pi,\pi'}|x,\theta]$, $E[A_\pi|x,\theta]$ and $E[G_{\pi,\sigma}|x,\theta]$.
First note that

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

so $E[A_\pi|x,\theta]$ is easily computed given $E[A_{\pi,\pi'}|x,\theta]$. But

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

so it is enough to be able to compute

$$\Pr[\pi_i = \pi, \pi_{i+1} = \pi', x|\theta]$$

To do this we introduce the backward variable

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

which can be computed using dynamic programming in a similar way as $f_\pi(i)$ was computed in lecture 5.

Now,

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

This way we can compute $E[A_{\pi,\pi'}|x, \theta]$ and from that $E[A_\pi|x, \theta]$.