Notes

Nate Richman (nate@nrichman.dev)

February 12, 2020

E-Step We want to calculate:

$$Z_{i,j}^t = \frac{P(X_i|Z_{i,j}=1, p^{t-1})}{\sum_{k=1}^m P(X_i|Z_{i,k}=1, p^{t-1})}$$

In log space this is:

$$\begin{aligned}\log(Z_{i,j}^t) &= \log\left(\frac{P(X_i|Z_{i,j}=1, p^{t-1})}{\sum_{k=1}^m P(X_i|Z_{i,k}=1, p^{t-1})}\right) \\ &= \log\left(P(X_i|Z_{i,j}=1, p^{t-1})\right) - \log\left(\sum_{k=1}^m P(X_i|Z_{i,k}=1, p^{t-1})\right)\end{aligned}$$

The rightmost term is a log of sums which we can compute recursively with:

$$\log(x+y) = \log x + \log\left(1 + \exp(\log(y) - \log(x))\right)$$

Which is in log space:

$$\log(x+y) = x' + \log(1 + \exp(y' - x'))$$

Which works because I've already calculated each term in log-space

M-Step We want to compute $p_{c,k}$ where it is defined as:

$$p_{c,k}^t = \frac{n_{c,k} + d_{c,k}}{\sum_{b\in\{A,C,G,T\}} (n_{b,k} + d_{b,k})}$$

For $k \neq 0$, $n_{c,k}$ is the sum over the sequences and all positions j where c appears as the kth column in the motif. If this is written as a loop over all characters (j) in each sequence (j), we have to think about what possible positions the character at position j. For example the first character in a sequence can only be the first column of a motif, the second character in the sequence could be the first or second character in the motif, etc.

So looping over all possible characters we find an array of what columns it could be in if it were part of the motif, and add the $Z_{i,j}$ value to its total where i is the sequence id and j is the starting position that would correspond to the character being in the kth position of that motif.

Likelihood function The M-step computes the joint likelihood function $P(X, Z|\theta) = \prod_i P(X_i, Z_i|\theta)$, but we want $P(X|\theta)$ which we can get by marginilizing over all starting points (j:

$$\begin{aligned} P(X|\theta) &= \prod_i^N \sum_{j=1}^{L-W+1} P(X_i, Z_{i,j} = 1|\theta) \\ &= \prod_i^N \sum_{j=1}^{L-W+1} P(X_i|Z_{i,j} = 1, \theta) P(Z_{i,j} = 1|\theta) \end{aligned}$$

Assuming that the probability of a start being at a certain position is uniform, we have the $P(Z_{i,j} = 1|\theta) = \frac{1}{m} = \frac{1}{L-W+1}$:

$$\begin{aligned} P(X|\theta) &= \prod_i^N \sum_{j=1}^{L-W+1} P(X_i|Z_{i,j} = 1, \theta) \frac{1}{L-W+1} \\ &= \prod_i^N \frac{1}{L-W+1} \sum_{j=1}^{L-W+1} P(X_i|Z_{i,j} = 1, \theta) \end{aligned}$$

And we are given that:

$$P(X_i|Z_{i,j} = 1, \theta) = \prod_{k=1}^{j-1} p_{c_{k,0}} \prod_{k=j}^{j+W-1} p_{c_{k,k-j+1}} \prod_{k=j+W}^{L} p_{c_{k,0}}$$

Putting it all together we have that:

$$P(X|p) = \prod_i^N \left[\frac{1}{L-W+1} \sum_{j=1}^{L-W+1} \left(\prod_{k=1}^{j-1} p_{c_{k,0}} \prod_{k=j}^{j+W-1} p_{c_{k,k-j+1}} \prod_{k=j+W}^{L} p_{c_{k,0}} \right) \right]$$

My code already calculates the inner sum so if I want to do it in log space I have:

$$\log P(X|p) = \sum_{i=1}^{N} \log \ \mathbf{sum} - \log (L-W+1)$$