Assignment 1 solution

Contents

1 Part 1

- $\psi_{k,p}^{(l)}$: a matrix of parameters of size $4 \times P$, where P is a hyperparameter representing the length of a motif (e.g. $P = 6$ or $P = 8$ might be common choices). $l = 0$ or $l = 1$.
- $X_{i,j,p}$: The base at position p of the subsequence starting at position j of sequence i
- $X_{i,j,p,k}$: A dummy indicator for which $X_{i,j,p,k} = 1$ iff $X_{i,j,p} = k$, otherwise 0.
- $C_i \in \{1, ..., L - P + 1\}$: a latent variable indicating the position of the motif in sequence i
- $C_{i,j}$: A dummy indicator for which $C_{i,j} = 1$ iff $C_i = j$, otherwise 0.
- $P(C_i = j|\boldsymbol{\theta}) = \lambda_j$
- $P(X_{i,j,p} = k|C_i, \boldsymbol{\theta}) = \psi_{p,k}^{(C_{i,j})}$

The complete log likelihood is therefore:

$$\ln P(X, C|\boldsymbol{\theta}) = \sum_i \sum_j C_{i,j} \ln P(C_i = j|\boldsymbol{\theta}) + \sum_i \sum_j \sum_p \sum_k X_{i,j,p,k} \ln \left[P(X_{i,j,p} = k|C_i = 1, \boldsymbol{\theta})\right]^{C_{i,j}} \left[P(X_{i,j,p} = k|C_i = 0, \boldsymbol{\theta})\right]^{1-C_{i,j}} \tag{1}$$

$$= \sum_i \sum_j C_{i,j} \ln \lambda_j + \sum_i \sum_j \sum_p \sum_k X_{i,j,p,k} \left[C_{i,j} \ln \psi_{p,k}^{(1)} + (1 - C_{i,j}) \ln \psi_{p,k}^{(0)}\right] \tag{2}$$

2 Part 2