Extra-credit

EN 600.438/638

April 21, 2016

Due date: April 29, 2016 by midnight
Submission: Please submit your assignments by email to en438.spring2016@gmail.com, including both code and written work. Problem sets can be submitted in Word, PDF, or plain text.
Reminders: You may discuss problems in small groups but must complete your own code. List group members on your submission. **You have 7 late days on this assignment.**
Programming component: You must submit all source code in addition to answering the questions below. Please include scripts for each sub-problem that will, when run using the data specified, produce the answer you provide. If you pre-process the data to enable easier loading or any other changes, please also provide the processed data along with a script and description of the processing applied. If your code does not run and is not documented, we will not be assigning any credit. Partial credits will be assigned in case code is well documented.

Data Location: Data for this assignment can be found on piazza, folder:**extra_credit**

Exercise 1. (50 points) Motif Finding
DNA binding proteins perform regulatory functions by binding to pattern in DNA sequence. The pattern of sequence that proteins can recognize can are called motifs. Motifs in general can be defined as short DNA sequence of biological importance.
EM for Motif Finding
In a typical case you do not know what a motif looks like. You also do not know the start site of a motif in a given sequence of DNA. We can use Expectation Maximization for motif identification. Here the hidden state is the start position of the motif in each training sequence.
How do you represent a motif? Let us assume that the motif has fixed length W. A motif can be represented as a Position Weight Matrix, P where the entry P_{ck} gives the probability of observing a nucleotide c at position k, where $k = 1, 2, 3, ..., W$ and $c = A, T, C, G$. B_c represents the background probability for each nucleotide.

Z_{ij} is an indicator variable that represents if the motif start site is j for sequence i. Given a sequence of length L and motif of size W, $j = \{1, 2, 3, 4, ..., L-W+1\}$. $\gamma(Z_{ij})$ represents the probability that the motif starts at position j in sequence i, i.e. $\gamma(Z_{ij}) = p(Z_{ij} = 1|X_i, P)$.

Given a specific length of motif W and training set of sequences set initial values for p:

1. **E Step:** Estimate $\gamma(Z_{ij})$ from P

$$\begin{aligned}\gamma(Z_{ij}^{(t)}) &= p(Z_{ij}=1|X_i,P^{(t)}) \\ &= \frac{p(X_i|Z_{ij}=1,P^{(t)})p(Z_{ij}=1)}{\sum_{k=1}^{L-W+1} p(X_i|Z_{ik}=1,P^{(t)})p(Z_{ik}=1)}\end{aligned}$$

We assume that it is equally likely that a motif will start at any position. Therefore, we can write the above equation as:

$$\gamma(Z_{ij}^{(t)}) = \frac{p(X_i|Z_{ij}=1,P^{(t)})}{\sum_{k=1}^{L-W+1} p(X_i|Z_{ik}=1,P^{(t)})}$$

To compute $p(X_i|Z_{ij}=1,P)$,

$$p(X_i|Z_{ij}=1,P) = \prod_{k=1}^{j-1} B_{c_k} \prod_{k=j}^{j+W-1} P_{c_k,k-j+1} \prod_{k=j+W}^{L} B_{c_k}$$

Here, X_i refers to i^{th} sequence, Z_{ij} is 1 if motif starts at position j in sequence i and c_k is character c at position k.

2. **M Step:** Estimate P_{ck} and B_c using $\gamma(Z)$

$$P_{ck} = \frac{n_{ck}+1}{\sum_b (n_{bk})+4}$$

where,

$$n_{ck} = \sum_i \sum_{j|X_{i,j+k-1}=c} Z_{ij}$$

$$B_c = \frac{g_c+1}{\sum_b (g_b)+4}$$

where,

$$g_c = m_c - \sum_{j=1}^{W} n_{c,j}$$

where m_c is the total number of times c appears in data. 1 and 4 in the above equations in the M step are pseudocounts to avoid situation where the numerator and/or denominator are 0.

3. Compute log likelihood of datagiven by:

$$\sum_i \sum_j \log\, p(X_i|Z_{ij}=1,P)$$

4. Repeat until convergence, i.e. absolute change in log likelihood is $< \epsilon$. For this exercise set $\epsilon = 0.001$.

A (30 points) Programming
Implement a function *findmotif(file_name, motif_width, iterations)* using the EM algorithm described above. It should be able to take the following inputs:

   (a) file name of sequence data

   (b) width of motif, i.e. W from above

   (c) max number of iterations to run EM (set default to 100)

   It should return:

   (a) P_{ck}

   Set background probability, i.e. $P_{c,0} = 0.25$ for all nucleotides.

B (10 points) Test your function using sequences in file *shortmotif.txt*.

   (a) Set motif width to 5

   (b) Max number of iterations to 50

   Report:

   (a) The string that corresponds to most likely nucleotide at each position in the motif

   (b) Position Weight Matrix in a tab-delimited file

C (10 points) Test your function using sequences in file *longmotif.txt*.

   (a) Set motif width to 8

   (b) Max number of iterations to 50

   Report:

   (a) The string that corresponds to most likely nucleotide at each position in the motif

   (b) Position Weight Matrix in a tab-delimited file