Bioinformatics Research Center
Aarhus University

Advanced Statistical Methods in Bioinformatics
Asger Hobolth
September 12, 2018

Handout 4

Lectures in Week 38

Monday, September 17 and Wednesday, September 19:
EM algorithm: General theory.
EM algorithm for HMMs: Forward and backward algorithms (EG Section 12.2.1 and 12.2.3).
Posterior decoding.

Exercises in Week 38

1. Show the recursion equations EG (12.5) page 412 and (12.7) page 413.
2. EG (12.12), (12.13) og (12.14) are the M-steps in the EM-algorithm for HMMs.
   Show these three equations.
3. EG Problem 12.4 page 429.
4. Show EG equation (12.16).
5. **CG-islands**
   R programs and data for this exercise can be found on the homepage.

   Stretches of DNA with a high CG-content often have specific interest. For example, promotor-regions typically have a high CG-content. In this exercise we will estimate an HMM for finding these so-called 'CG-islands'. Within a CG-islands the content of C and G is particularly high. The distribution of the four nucleotides is

   $$\text{Within}: \; b(A) = 1/8, \; b(G) = 3/8, \; b(C) = 3/8, \; b(T) = 1/8.$$

   Outside CG-islands the four nucleotides appear in equal proportions

   $$\text{Outside}: \; b(A) = 1/4, \; b(G) = 1/4, \; b(C) = 1/4, \; b(T) = 1/4.$$

   We describe the situation with an HMM with two hidden states corresponding to inside and outside a CG-island. The transition matrix between the two hidden states is unknown.

   a) The DNA-sequence to be analysed can be found at the homepage (CpGisland.dat). Download the sequence. You can read and translate the sequence to a numeric vector using the commands

   ```
   ## Read sequence from file
   CpGdat <- readLines("CpGisland.dat")
   ## Split and translate raw sequence from AGCT to 1234
   ## and make the vector numeric
   ObsSeq <-
      as.numeric(strsplit(chartr("AGCT","1234",CpGdat),"")[[1]])
   ```

b) Use the Forward-Backward algorithms and the EM algorithm to estimate the transition matrix. Let the initial distribution be $\pi = (1/2, 1/2)$.

   Preferably you should implement your own versions of the Forward and Backward algorithms; otherwise you can download my implementation of the algorithms from the homepage (HMMexpectations.R).

c) Find CG-islands in the DNA-sequence. How many are there? How big are they? Use both the Viterbi sequence and posterior decoding for investigating the underlying hidden state sequence. Which type of decoding do you prefer?

Hint: The program EMexample.R gives an example of estimation using the EM algorithm. The program uses HMMsim.R for simulating an observed sequence.