Bioinformatics Research Center
Aarhus University

Advanced Statistical Methods in Bioinformatics
Asger Hobolth
September 4, 2018

Handout 3

Lectures in Week 37

Monday, September 10:

Introduction to Hidden Markov Models (EG Section 12.1) and the Viterbi algoritm (EG Section 12.2.2).

Applications of HMMs in biological sequence analysis: Read about Gene Finding in EG Chapter 12.3.4.

Wednesday, September 12:

The EM algorithm.

Exercises in Week 37

1. EG Problem 12.1 page 428.

2. Show the recursion equation (12.11) in EG on page 414.

3. Write a computer program in R that simulates from a hidden Markov model. What should the input be? What should the output be?

   Use your program to simulate a sequence of heads and tails from the dishonest casino.

4. Decode the sequence from the previous exercise using the Viterbi algorithm. Preferably you should implement your own version of the Viterbi algorithm; otherwise download, understand, test and try my implementation and example of the Viterbi algorithm (viterbi.R and viterbiEx.R; to be found on the homepage).

5. Recall the multinomial distribution in EG Section 2.4.2. Suppose $(x_1, \ldots, x_k)$ is a random sample from a multinomial distribution with n trials (i.e. x_j is non-negative and $\sum_{j=1}^k x_j = n$) and k possible outcomes with probabilities $(p_1, \ldots, p_k)$ (i.e. $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$).

   Show that the maximum likelihood estimate is $\hat{p}_j = x_j/n$, i.e. show that

   $$\prod_{j=1}^k p_j^{x_j} \leq \prod_{j=1}^k \left(\frac{x_j}{n}\right)^{x_j},$$

   with equality exactly when $p_j = x_j/n$, $j = 1, \ldots, n$.
   Hint. There are many ways to show this result. Perhaps use Google to find a proof that you like?