

Data Mining and Machine Learning in Bioinformatics

Dr. Holger Fröhlich, Dr. Martin Vogt

Sabyasachi Patjoshi

sabyasachi2k13@gmail.com, martin.vogt@bit.uni-bonn.de

Due: Jul 1, 10:30 (by the end of the lecture)

Exercise Series 9

General: Exercises are to be solved and submitted in fixed groups by at most 3 students. Every member of a group should help solving each task and thus be able to answer questions to each task. No late submissions are accepted. Copying solutions will automatically lead to a point reduction of at least 50%. N – 1 homework assignments and N – 2 programming tasks have to be submitted in total.

A group can gain additional bonus points, if it presents its solution for a particular task during the tutorials. Accordingly, each task has a defined number of points as well as bonus points.

1. We define a HMM with states S1, S2, S3, alphabet $A=\{a,b,c\}$ and transition

probability matrix $P = \begin{bmatrix} 0 & 1/4 & 3/4 \\ 1/2 & 1/4 & 1/4 \\ 0 & 2/3 & 1/3 \end{bmatrix}$. The emission probabilities are given in

matrix $B = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 3/4 & 1/4 & 0 \\ 0 & 1/3 & 2/3 \end{bmatrix}$ (Notice: rows denote states and columns characters).

- a) Make a graphical representation of the HMM model (c.f. lecture slides). (4 points + 1 bonus point)
- b) What is the probability for observing the sequence (a, b, c, b) given the hidden state sequences
- (i) $S1 \rightarrow S3 \rightarrow S3 \rightarrow S2$
 - (ii) $S2 \rightarrow S1 \rightarrow S3 \rightarrow S3?$
- (8 points + 1 bonus point)
2. Manually apply the Viterbi algorithm to find the most probable hidden state sequence for (a, b, c, a) and its likelihood (c.f. lecture slides). (6 points + 1 bonus point)
3. Use either R, Python, or Java to make an as simple as possible implementation of the Viterbi algorithm shown as pseudo-code in the lecture. Use this implementation to calculate the optimal path and its log-likelihood for the observed sequence (a, b, c, b, a) using the HMM model defined in task 1. (8 points + 1 bonus point)