Exercises – Lecture 4
Hidden Markov Models

Question 1 *(without R)*
Define the hidden Markov model $(\boldsymbol{\pi}, \mathbf{P}, \mathbf{B})$ with the three states S_1, S_2, S_3, alphabet $\mathcal{A} = \{a, b, c\}$, and the following parameters:

$$\boldsymbol{\pi} \;=\; (1,0,0)^{\top}, \quad \mathbf{P} \;=\; \left(\begin{array}{ccc} 0 & 0.5 & 0.5 \\ 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{array} \right), \quad \text{and} \quad \mathbf{B} \;=\; \left(\begin{array}{ccc} 0.5 & 0.5 & 0.0 \\ 0.5 & 0.0 & 0.5 \\ 0.0 & 0.5 & 0.5 \end{array} \right)$$

Let the observed sequences be either *i)* $(y_1, y_2, y_3) = (a, b, c)$ or *ii)* $(y_1, y_2, y_3) = (a, c, a)$. What are all possible underlying state sequences (X_1, X_2, X_3) of these two observed sequences? Moreover, calculate $P((Y_1, Y_2, Y_3) = (y_1, y_2, y_3))$ for both observed sequences.

Question 2 *(without R)*
Consider a binary DNA sequence (comprising of zero's and ones only). The dependence in the observed sequence of zero's and ones is a consequence of the underlying intron-exon structure of the DNA. A hidden Markov model links the observed $(\{Y_t\}_{t=1}^{\infty})$ and latent $(\{X_t\}_{t=1}^{\infty})$ sequences. The parameters of this model $(\boldsymbol{\pi}, \mathbf{P}, \mathbf{B})$ are given by $\boldsymbol{\pi} = (1, 0)^T$,

$$\mathbf{P} = \begin{array}{r} \text{intron} \\ \text{exon} \end{array} \left(\begin{array}{cc} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{array} \right), \quad \text{en} \quad \mathbf{B} = \begin{array}{r} \text{intron} \\ \text{exon} \end{array} \left(\begin{array}{cc} 0.5 & 0.5 \\ 1 & 0.0 \end{array} \right),$$

where $\boldsymbol{\pi}$ is the initial distribution, $\mathbf{P}$ the transition matrix, and $\mathbf{B}$ the emission matrix. In the latter the 1st and 2nd column correspond to the 0 and 1, respectively.

a) Express $P(Y_{t-1} = 0, Y_{t+1} = 1 \,|\, X_t = \text{exon})$ in terms of the parameters of the hidden Markov model. Hereby assume that the latent Markov process is stationary.

b) Let $\alpha = \frac{1}{4}$. Which (x_1, x_2, x_3) maximizes $P((X_1, X_2, X_3) = (x_1, x_2, x_3) \,|\, (Y_1, Y_2, Y_3) = (0, 1, 0))$?

Question 3 *(without R)*

a) Construct an HMM that generates a sequence of the form $\mathtt{A}^{k_1}\mathtt{C}^{k_2}\mathtt{A}^{k_3}\mathtt{C}^{k_4}\ldots$ where, e.g., $\mathtt{A}^{k_1}$ represents a series of length k_1 consisting of only A's. The $k_1, k_2, k_3, \ldots$ are drawn from the set $\{1, 2, 3\}$ with equal probabilities.

b) Construct an HMM that generates a sequence of the form $\mathtt{A}^{k_1}\mathtt{C}^{4-k_1}\mathtt{A}^{k_2}\mathtt{C}^{4-k_2}\ldots$ where, e.g., $\mathtt{A}^{k_1}$ represents a series of length k_1 consisting of only A's. The $k_1, k_2, k_3, \ldots$ are drawn from the set $\{1, 2, 3\}$ with equal probabilities.

Question 4 *(without R)*

Sections of the DNA rich in `C`'s and `G`'s are called CpG islands. No unambiguous definition exist however. CpG islands tend be more prone to methylation. Methylation refers to the addition of a methyl group. Methylation levels may either be below or above normal, referred to as hypo- and hyper-methylation, respectively. Suppose knowledge of the DNA sequence is unavailable, but methylation levels are known. From the latter, try to reconstruct the sequence of CpG islands. Hereto use a hidden Markov model. Hence, model the hidden sequence of CpG islands by a first order Markov chain. Hereto use that 10% of the DNA (which has an infinite length) belongs to a CpG island, and there is a 5% chance to jump to a CpG island from DNA with a low `C`'s and `G`'s frequency. Assume that a CpG island has twice as high a probability to be normally methylated than hyper-methylated, and never hypo-methylated. Outside CpG islands hyper-methylation does not occur and two-third of the DNA is normally methylated.

a) Give the parametrization of the hidden Markov model.

b) Calculate the probability of a DNA section belonging to a CpG island, given that it is normally methylated.

c) Calculate the probability of a DNA section belonging to a CpG island, given that it is normally methylated *and* its directly preceeding region is hyper-methylated.

Question 5

Assume a caricature of the DNA molecule consisting of alternating introns and exons. The nucleotide sequences of introns and exons have different base pair compositions. From the observed sequence of nucleotides we wish to infer the unobserved sequence of introns/exons. Hereto we assume that the nucleotide sequence can be modeled by the following hidden Markov model. Let $\mathcal{S} = \{I, E\}$ be the set of states (corresponding to Intron and Exon) of the hidden Markov chain, $\boldsymbol{\pi}$ the initial state probabilities, $\mathbf{P}$ the transition matrix, $\mathcal{A} = \{A, C, G, T\}$ the emission alphabet, and $\mathbf{B}$ emission matrix, where:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_I \\ \pi_E \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} P_{I,I} & P_{I,E} \\ P_{E,I} & P_{E,E} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},$$

$$\text{and} \quad \mathbf{B} = \begin{pmatrix} b_{I,A} & b_{I,C} & b_{I,G} & b_{I,T} \\ b_{E,A} & b_{E,C} & b_{E,G} & b_{E,T} \end{pmatrix} = \begin{pmatrix} 0.49 & 0.01 & 0.49 & 0.01 \\ 0.01 & 0.49 & 0.01 & 0.49 \end{pmatrix}.$$

a) *(with R)* Generate a DNA sequence of 1000 nucleotides. Save the sequences of states and nucleotides. Report the R-code and the nucleotide sequence. Hint: use the `sample` function and `for`-loop construction.

b) *(with R)* Compare the sequence of states with that of the nucleotides. Do you think it would be possible to infer (with high certainty) the state sequence from the nucleotide sequence alone? Why?

c) *(without R)* Would it still be possible to infer (with high certainty) the state sequence from the nucleotide sequence alone if that sequence is generated with the same π and $\mathbf{P}$, but with the following new emission matrix:

$$\mathbf{B} = \begin{pmatrix} 0.35 & 0.15 & 0.15 & 0.35 \\ 0.35 & 0.15 & 0.15 & 0.35 \end{pmatrix}.$$

Motivate your answer.

d) *(without R)* The hidden Markov model of part *c* reduces to a simple Markov chain model. Write down the latter.

Question 6 *(without R)*

Consider the same hidden Markov model as in Question 5 (the first one!).

a) Calculate the likelihood for the sequences `TGC` and `CGATG`.

b) Determine the most likely underlying series of states for the sequences `TGC` and `CGATG`.