Exercises – Lecture 4
Hidden Markov Models

Question 1 *(without R)*
Define the hidden Markov model $(\boldsymbol{\pi}, \mathbf{P}, \mathbf{B})$ with the following parameters:

- three states S_1, S_2, S_3, alphabet $\mathcal{A} = \{1, 2, 3\}$.
- $\mathbf{P} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{pmatrix}$
- $\boldsymbol{\pi} = (1, 0, 0)^T$
- $\mathbf{B} = \begin{pmatrix} 0.5 & 0.5 & 0.0 \\ 0.5 & 0.0 & 0.5 \\ 0.0 & 0.5 & 0.5 \end{pmatrix}$

What are all possible state sequences for the following observed sequences $\mathcal{O}$, and what is $P(\mathcal{O} \,|\, (\boldsymbol{\pi}, \mathbf{P}, \mathbf{B}))$?
(i) $\mathcal{O} = 1, 2, 3$.
(ii) $\mathcal{O} = 1, 3, 1$.

Question 2
Question 2a (without R)
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.

Question 2b (without R)
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 3
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.

Question 3a (without R)
Give the parametrization of the hidden Markov model.

Question 3b (without R)
Calculate the probability of a normally methylated DNA section belonging to a CpG island.

Question 3c (without R)
Calculate the probability of normally methylated DNA section belonging to a CpG island, given that the directly preceeding region is hyper-methylated.

Question 4
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:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_I \\ \pi_E \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} P_{I,I} & P_{I,E} \\ P_{E,I} & P_{E,E} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},$$
$$\text{and} \quad \mathbf{B} = \begin{pmatrix} b_{I,A} & b_{I,C} & b_{I,G} & b_{I,T} \\ b_{E,A} & b_{E,C} & b_{E,G} & b_{E,T} \end{pmatrix} = \begin{pmatrix} 0.49 & 0.01 & 0.49 & 0.01 \\ 0.01 & 0.49 & 0.01 & 0.49 \end{pmatrix}.$$

Question 4a (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.

Question 4b (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?

Question 4c (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.

Question 4d (without R)
The hidden Markov model of Question 4c reduces to a simple Markov chain model. Write down the latter.

Question 5
Consider the same hidden Markov model as in Question 2 now with the following emission matrix:

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

Question 5a (with R)
Generate a long DNA sequence of (say) 100000 base pairs using this model. Save the sequences of states and nucleotides. Verify using this sequence that a random base pair in this sequence has a probability of 0.50 to be an intron.

Question 5b (without R)
Calculate from the model (do not use the sequence) the probabilities of the nucleotides, $P(A), P(C)$, $P(G), P(T)$. Verify your answer using the sequence generated in Question 5a.

Question 5c (without R)
Calculate the conditional probabilities, $P(A|A), P(A|C), \ldots, P(T|T)$ (see lecture notes).

Question 5d (with R)
Suppose we do not know the sequence of Question 3a is generated by a hidden Markov model, and assume a first order Markov chain process instead. Fit the latter model by means of maximum likelihood. Compare its transition matrix to the probabilities calculated in Questions 5c.

Question 5e (with R)
Using the sequence generated in Question 5a, calculate the nucleotide frequencies per state. Compare these to the parameters of the emission matrix. Verify that, given the intron/exon sequence, the parameters of the emission matrix can be estimated straightforwardly.

Question 6
Consider the same hidden Markov model as in Question 4.

Question 6a (without R)
Calculate the likelihood for the sequences `TGC` and `CGATG`.

Question 6b (without R)
Determine the most likely underlying series of states for the sequences `TGC` and `CGATG`.