Dr. Peter Arndt

Master for AI and Data Science
Universität Düsseldorf

Math for AI
Winter term 2023/24

Exercise sheet 11

Exercise 43 (11 points)

Consider a Markov chain with state space $\{1, \dots, 7\}$ given by the following transition matrix:

$$\begin{pmatrix} 0 & 0 & \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{6} & 0 \\ 0 & 0 & 0 & \frac{2}{5} & 0 & 0 & \frac{3}{5} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} & 0 & 0 & 0 \end{pmatrix}$$

(a) (1 point) Draw the transition diagram.

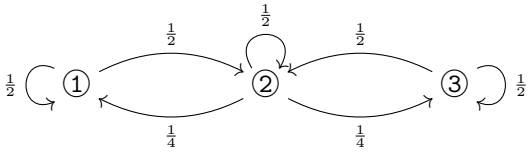
(b) (4 point) Say for each state whether it is transient or recurrent and justify your answer.

(c) (4 points) Determine $\lim_{n\to\infty} p_{ii}(n)$ for all i.

(d) (1 point) Find two different invariant distributions.

Exercise 44 (11 points)

Consider the Markov chain given by the following transition diagram:



(a) (1 point) Write down the transition matrix of the Markov chain

(b) (2 points) Say, with proof, which states are transient and which states are recurrent.

(c) (4 points) Show that there is a unique stationary distribution and compute it.

(d) (1 point) If one starts in state ②, is the probability distribution of the states after n steps equal to the stationary distribution for some n? Justify your answer.

(e) (2 points) (tricky!) If one starts in state ①, is the probability distribution of the states after n steps equal to the stationary distribution for some n? Justify your answer.

Exercise 45 (9 points)

The AI researcher J. Doe thinks that Markov chains are awesome. While thinking about those, Doe switches between two states, E (excited enthusiasm) and C (calm satisfaction). In state E, Doe sits in an armchair smiling (S) with probability $\frac{1}{2}$ and dances through the office (D) with probability $\frac{1}{2}$. In state C, Doe sits in an armchair smiling (S) with probability 1 and dances (D) with probability 0.

From one minute to the next, Doe switches between the states E and C with probability α, and stays in the same state with probability $1 - \alpha$.

At the beginning Doe is enthusiastic.

(a) (3 point) Describe the situation by a hidden Markov model, linking the observed sequence of Doe's behaviour $\{Y_t\}_{t\in\mathbb{N}}$ (where the Y_t take values S, D) and the latent sequence of Doe's states $\{X_t\}_{t\in\mathbb{N}}$ (where the X_t take values E, C). Concretely: Write down a distribution vector for the initial state π, a transition matrix P and an emission matrix B

(b) (7 points) Let $\alpha = \frac{1}{4}$. Suppose, starting at time $t = 1$ you see Doe sit, then dance, then sit again (i.e. $Y_1 = S, Y_2 = D, Y_3 = S$). What is the most likely sequence of states that Doe was in? [i.e. which sequence of states (x_1, x_2, x_3) maximizes $P(X_1 = x_1, X_2 = x_2, X_3 = x_3 \mid Y_1 = S, Y_2 = D, Y_3 = S)$?]

Exercise 46 (9 points)

In a DNA sequence there are places called *introns*, where the DNA is *not* used to produce proteins, and there are places called *exons* where the DNA *is* used for building proteins. In the exercise folder you are given the complete genome of the Blochmannia bacterium. Your goal in this exercise is to guess which places of the genome are introns, and which are exons, using a hidden Markov model.

In the exercise folder you find the genome in the file `BlochmannDNA.txt` in the form of a string consisting of the letters 'a','c','g' and 't' (the file contains a single long line). The file `BlochmannDNA2.txt` contains the same information broken up into many lines, in case you have trouble reading the first file. I got the genome from the National Center for Biotechnology Information genome data base.

Recall from your biology classes long ago in school that the nucleic acids denoted by the letters a,c,g,t encode proteins as specified in the Codon tables. Now suppose that in the Blochmannia bacterium you have observed proteins which make you pretty sure that the positions 405000–405070 and 690040–690160 are exons, and from the lack of corresponding proteins in the bacterium you know for sure that positions 1040–1111, 2060–2150 and 386370–386450 are introns. From the distribution of the letters a,c,g,t in these regions you can estimate the exon/intron distributions, which are the rows of the emission matrix of a hidden Markov model.

This you can use then to guess for other sequences of places in your DNA string whether they are introns or exons.

Your task is to do this using Python's Pomegranate package. It is not too well documented, but there is an example file, showing how you would model the tree ring example given in the extra video.

See the supplied files for more precise instructions.

Deadline: Friday 19th of January, 10:00.
Upload your solution to this link.