Sequential Models in Data Science Markov Chains and Endomorphism Reduction

Yirmeyahu Kaminski

INTRODUCTION

This is the fourth and last project you need to submit. In this project, you will train your knowledge on Markov chains, hidden Markov models.

The project is to be done by pairs of students. Each pair must present a pdf file with the theoretical answers and a Jupyter notebook with the Python function. The notebook will be tested as this. No modification should be necessary for making it work.

1 SOME QUESTIONS ON ENDOMORPHISMS

1. Let p be a strictly positive natural number and M be a complex matrix in $M_n(\mathbb{C})$, such that $M^p = I_n$. Let ω be an order p root of unity, which is not an eigenvalue of M. Show that:
$$I_n + \omega M + \omega^2 M^2 + \cdots + \omega^{p-1} M^{p-1} = 0.$$

2. Let E_n be the real linear space of polynomials of degree less or equal to n with real coefficients. Let u be the function defined on E_n by $u(P) = (X - a)(X - b)P' - nXP$, where $a, b \in \mathbb{R}$. Show that $u \in \text{End}(E_n)$. Compute the eigenvalues of u and for each eigenspace a basis. Find the kernel of u and the set of polynomials P such that $u(P) = 1$.

3. Compute the Jordan form of the following matrix:

$$\begin{pmatrix} -8 & 1 & 5 \\ 2 & -3 & -1 \\ -4 & 1 & 1 \end{pmatrix}$$

2 A STANDARD QUESTION ON MARKOV CHAINS...

Consider a Markov chain with states $S = \{0, \cdots, N\}$ and transition probabilities $p_{i,i+1} = p$, $p_{i,i-1} = q$, for $1 \leq i \leq N-1$, where $p+q = 1$, $0 < p < 1$; assume $p_{0,1} = 1$, $p_{N,N-1} = 1$.

1. Draw the graph (i.e. transition diagram).

2. Is the Markov chain irreducible?

3. Find the stationary distribution.

3 THE GAMBLER'S RUIN...

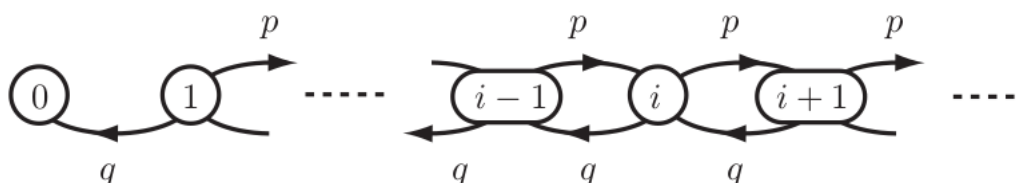Consider the following chain that models the situation of a gambler in some casino:

Figure 3.1: The gambler's Markov chain

In this chain, the transition probabilities are given by:

$$
\begin{aligned}
p_{00} &= 1, \\
p_{i,i-1} &= q, \quad p_{i,i+1} = p \text{ for } i = 1,2,\cdots
\end{aligned}
$$

We assume that both $p > 0$ and $q > 0$.
Compute the ruin probability: $h_i = \mathbb{P}(\{\exists n \geq 0, X_n = 0\}|X_0 = i)$ (use exercise 6 in the second sheet of exercises). Show that even if the casino is fair, that is if $p = q$, the gambler probability to be ruined is 1.

4 HIDDEN MARKOV MODELS

4.1 SYSTEM DESCRIPTION

Welcome to the Emperor palace! The elder daughter of Probabil the Great, beautiful Princess Variance, reached the 2-pi-square age that is considered as a mean of a normal distribution for the age when maidens of the Empire get married. Today, the traditional matrimonial contest will be held in the Palace. The winner will marry the Princess.
The contest has ancient roots. When a princess is born, the Emperor assigns her a special servant. Every day from the day of the princess birth to the day of her 2-pi-square age her

servant must visit four ponds in the Emperor Garden, in accordance to a Markov process, and catch one hundred fishes. The servant must record fish colors (red, blue, or green) and then return the fish back to the ponds. Each pond is strictly maintained and has its own proportion of fish of different colors. Every day, the Emperor mass media announced the results of the fishing.

Ten days before the princess comes of age, her servant will put every caught fish into a transparent numbered jar and send it to the Palace. (It is assumed that taking out 1,000 fishes does not change the statistical properties of the ponds.) In order to win the Princess hand in marriage, a contestant has to guess as accurately as possible from which pond each fish came.

Our data is a sequence of observations O of length $L = 1000$. Every data element is a color of a fish from a finite set of colors (or a finite alphabet of symbols). In our case the set contains three symbols: "red", "blue", and "green". Each fish was taken out of some pond, or we can say "a data point was emitted in some state q". A first-order Markov chain determines the sequence of states, according to the following transition matrix:

$$A = \begin{pmatrix} 0.6 & 0.05 & 0.1 & 0.25 \\ 0.1 & 0.05 & 0.7 & 0.15 \\ 0.05 & 0.7 & 0.1 & 0.15 \\ 0.1 & 0.1 & 0.6 & 0.2 \end{pmatrix}$$

Every state has its own discrete probability distribution for fish color. We shall call this distribution a symbol emission vector in $i-$th state. Collecting all vectors as columns of the matrix, we can get a symbol emission matrix:

$$B = \begin{pmatrix} 0.8 & 0.2 & 0.1 & 0.5 \\ 0.1 & 0.7 & 0.7 & 0.3 \\ 0.1 & 0.1 & 0.2 & 0.2 \end{pmatrix} \begin{matrix} R \\ G \\ B \end{matrix}$$

The first state is randomly chosen from the four possible states according to the prior distribution:

$$\pi = (\ 0.3 \quad 0.1 \quad 0.4 \quad 0.2\ )$$

4.2 Tasks

1. Simulate the discrete system and extract a sequence of 1000 states and observations.

2. Implement the forward algorithm

3. Implement the backward algorithm

4. Compute with both algorithms the probability of the sequence of observations generated by the simulation.

5. Implement the Viterbi algorithm.

6. Compute the most probable sequence of states given the sequence of observations generated by the simulation. Compare it with the sequence of states generated by the simulation.

Good Luck!