Module 4: Lesson 1

Reinforcement Learning: Introduction to Markov Processes

WORLDQUANT UNIVERSITY

Outline

- What is Reinforcement Learning?
- Markov chains: Definition and stationary distribution
- Estimating Markov transition probabilities from transition data
- Absorbing Markov chains

What is Reinforcement Learning?

Machine learning (ML) provides automated methods that can detect patterns in data and use them to achieve some tasks.

Within ML methods, Reinforcement learning (RL) is the task of learning how a decision-maker should take sequences of actions in order to optimize the cumulative rewards.

The general RL problem is formalized as a discrete time stochastic control problem.

At each time step t, the environment is represented by:

- A state realization $s_t \in \mathcal{S}$, where $\mathcal{S}$ represents all the possible situations where the agent can fall.
- Given s_t, the agent must take an action $a_t \in \mathcal{A}$.
- The action a_t leads to (i) a reward $r_t \in \mathcal{R}$ for the agent, and (ii) a transition to state $s_{t+1} \in \mathcal{S}$.

Beginning at time $t = 0$, the RL problem boils down to finding a sequence of actions $\{a_0, a_1, ..., a_t, ...\}$ that maximizes the cumulative rewards for the agent.

Reinforcement Learning and the Markov property

We are going to consider the stochastic control process with the Markovian property.

A discrete-time stochastic control process is Markovian (i.e., it has the Markov property) if

- $\mathbb{P}(s_{t+1}|s_t, a_t) = \mathbb{P}(s_{t+1}|s_t, a_t, ..., s_0, a_0)$.
- $\mathbb{P}(r_t|s_t, a_t) = \mathbb{P}(r_t|s_t, a_t, ..., s_0, a_0)$.

The Markov property means that the future of the process only depends on the current observation, and the agent has no interest in looking at the full history.

Before delving further into Reinforcement Learning, we are going to focus on the properties of random processes with the Markov property and some specific finance applications.

Markov chains

If the state space consists of countably many states, the Markov process is called a **Markov chain**.

We say that a Markov chain is **homogeneous** if the probabilities of transition across states are independent of t. Thus, we can specify the **transition matrix** $P(p_{ij})$, where it must satisfy that $\sum_j p_{ij} = 1$ for all i.

How do the further-ahead realizations of the chain also depend on current realizations? Letting p_{ij}^n denote the $(i,j)^{th}$ position of matrix P^n:

$$\mathbb{P}(s_{t+m+n} = s_j | s_t = s_i) = p_{ij}^{m+n} = \sum_{k=1}^{N} p_{kj}^n p_{ik}^m \tag{1}$$

If the memory of the past dies out with increasing n, P^n should converge to a limit as $n \Rightarrow \infty$ and each column j should converge toward the same value π.

Assuming convergence, this means that the limiting or **stationary distribution** as $n \Rightarrow \infty$ satisfies:

$$\pi = \pi P; \text{ where } \pi = \{\pi_1, ..., \pi_N\}, \pi_k \geq 0 \text{ and } \sum_k \pi_k = 1 \tag{2}$$

Estimating transition probabilities from transition data

Suppose that we observe the realizations of a Markov chain and wish to estimate the probabilities of the transition matrix P. If n_{ij} is the number of times that we observe a change from state i to state j and N is the total number of states, then the estimated transition probabilities can be computed as:

$$\widehat{p}_{ij} = \frac{n_{ij}}{\sum_{k=1}^{N} n_{ik}} \tag{3}$$

One can show that this expression corresponds to the Maximum Likelihood Estimator for p_{ij}.

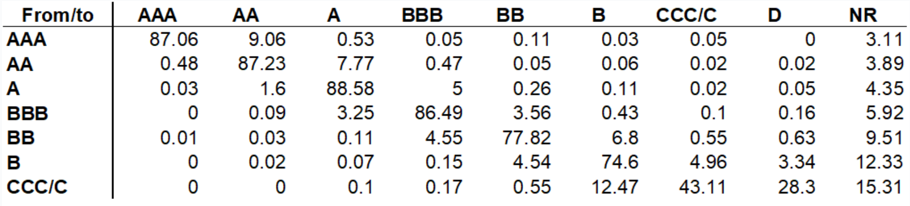
Absorbing states in a Markov chain

If a Markov chain is homogeneous, we say that state i is **absorbing** if $p_{ii} = 1$. That is, once the Markov chain enters an absorbing state, it stays there forever.

Absorbing Markov chains play a prominent role in finance, for example, to model credit ratings.

Consider a homogeneous Markov chain S_t defined on the state space $\mathcal{S} = \{1, 2, ..., D-1, D\}$ with transition matrix P. State 1 represents the highest credit class, state 2 the second highest, and so on, while state $D-1$ refers to the lowest credit class, and state D represents default.

The table below shows the actual transition probabilities for credit ratings published by Standard & Poor's (NR stands for "rating withdrawn")



From/to	AAA	AA	A	BBB	BB	B	CCC/C	D	NR
AAA	87.06	9.06	0.53	0.05	0.11	0.03	0.05	0	3.11
AA	0.48	87.23	7.77	0.47	0.05	0.06	0.02	0.02	3.89
A	0.03	1.6	88.58	5	0.26	0.11	0.02	0.05	4.35
BBB	0	0.09	3.25	86.49	3.56	0.43	0.1	0.16	5.92
BB	0.01	0.03	0.11	4.55	77.82	6.8	0.55	0.63	9.51
B	0	0.02	0.07	0.15	4.54	74.6	4.96	3.34	12.33
CCC/C	0	0	0.1	0.17	0.55	12.47	43.11	28.3	15.31

Summary of Lesson 1

In Lesson 1, we have looked at:

- The Reinforcement Learning setup
- Markov processes and their basic properties

⇒ **References for this lesson:**

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018. (Check the introductory chapters)

Polyanskiy, Y. ”Fundamentals of Probability.” Fall 2018. Massachusetts Institute of Technology: MIT OpenCouseWare, https://ocw.mit.edu/. (Check the Lecture Notes on Markov processes)

TO DO NEXT: Now, please go to the associated Jupyter notebook for this lesson to understand the properties of Markov processes and their application to credit ratings.

In the next lesson, we will take a closer look at further applications of Markov processes.