Reinforcement Learning

Introduction to Reinforcement Learning

Nguyễn Đăng Trị

School of Engineering and Technology
Hue University

Ngày 1 tháng 3 năm 2024

Overview

1. Markov chains
2. Chapman-Kolmogorov equations
3. Mô phỏng MC

Markov chains

- Ask Gemini:

  Chuỗi Markov, hay còn gọi là xích Markov, là một mô hình toán học mô tả một dãy các biến cố ngẫu nhiên, trong đó xác suất của mỗi biến cố chỉ phụ thuộc vào trạng thái của biến cố trước đó. Nói cách khác, chuỗi Markov ghi nhớ quá khứ để dự đoán tương lai

- Ask chatGPT:

  Markov chain là một mô hình xác suất thống kê mà mô tả sự di chuyển của một hệ thống qua các trạng thái khác nhau trong một chuỗi thời gian rời rạc. Đặc điểm chính của mô hình này là tính chất Markov, có nghĩa là trạng thái tương lai chỉ phụ thuộc vào trạng thái hiện tại và không phụ thuộc vào lịch sử trạng thái trước đó.

Markov chain

- Đặt $n = 1, 2, 3, \ldots$ thời gian
- Gọi X_n là trạng thái của hệ thống tại thời gian n
- N là tập hợp các trạng thái có thể của hệ thống
- $i = 0, 1, 2, \ldots$ với $i \in N$ là các trạng thái có thể xảy ra trong hệ thống

Markov chain

- Ta có $X_{\mathbb{N}}$ được gọi là quá trình ngẫu nhiên
- Và, $\mathbf{X}_n = [X_n, X_{n-1}, \ldots, X_0]^T$ là "**lịch sử**" của quá trình X_N
- Quá trình X_N được gọi là Markov chain (MC) khi và chỉ khi $\forall n \geq 1, i, j, x \in \mathbb{N}^n$

$$P(X_{n+1} = j | X_n = i, \mathbf{X}_{n-1} = x) = P(X_{n+1} = j | X_n = i) = P_{ij} \tag{1}$$

- **Trạng thái tương lai** chỉ phụ thuộc vào **trạng thái hiện tại** và **không phụ thuộc trạng thái trong quá khứ**

Markov chain

- **"Quá khứ"** không liên quan đến **"tương lai"**
- $\mathbf{X}_{n-1}$ không liên quan đến X_{n+1}
- P_{ij} được gọi là xác suất chuyển từ trạng thái i sang trạng thái j

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i) = P_{ij} \qquad (2)$$

- Tuy nhiên .....

Markov chain

- $$P(X_{n+m} = j | X_n = i, \mathbf{X}_{n-1} = x) = P(X_{n+m} = j | X_n = i) \quad (3)$$
- X_{n+m} phụ thuộc vào X_{n+m-1}
- X_{n+m-1} lại phụ thuộc vào X_{n+m-2}
- X_{n+m-2} lại phụ thuộc vào X_{n+m-3}
- ...
- X_{n+1} lại phụ thuộc vào X_n

Hừm, có gì đó sai sai...

Markov chain

- P_{ij} là xác suất chuyển từ trạng thái i sang trạng thái j
- Vậy

$$P_{ij} \geq 0,$$
$$\sum_{j=1}^{\infty} P_{ij} = 1. \tag{4}$$

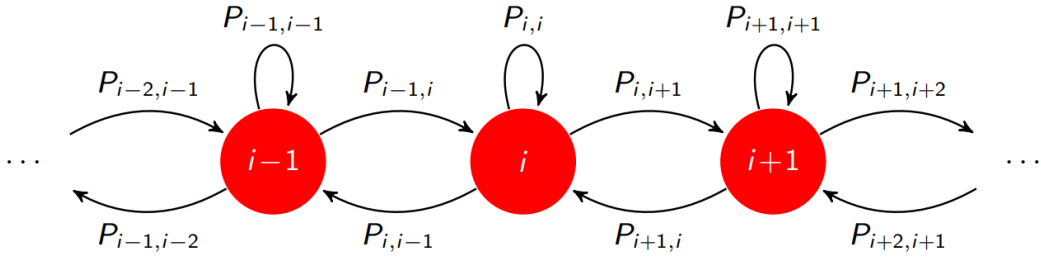
Tại sao thế? vì sao thế? Why?

Biểu diễn dạng ma trận

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & P_{02} & \dots & P_{0j} & \dots \\ P_{10} & P_{11} & P_{12} & \dots & P_{1j} & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots \\ P_{i0} & P_{i1} & P_{i2} & \dots & P_{ij} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{5}$$

- Là ma trận nhưng không phải ma trận; là số lượng các trạng thái hữu hạn
- Tổng các hàng có gì đặc biệt?

Biểu diễn dạng đồ thị



- Tổng các mũi tên đi ra từ một nốt có gì đặc biệt?

Let's play

Example

Hôm nay tôi buồn: $X_n = 1$
Hôm nay tôi vui: $X_n = 0$
Ngày mai tôi vui hay buồn, chỉ phụ thuộc vào hôm nay

- Xác xuất chuyển trạng thái

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix} \tag{6}$$

- Vẽ đồ thị thể hiện toàn bộ quá trình trên

Let's play

Example

Túy quyền và những bước đi. Ông A nhậu quá chén và tìm đường về nhà. Vì **mất quyền điều khiển hệ thống** nên ông bước đi loạng choạng. Giả sử rằng xác suất tại mỗi thời điểm ông A tiến một bước là p và lùi một bước là $1 - p$.

- Vẽ sơ đồ thể hiện quá trình ông A về nhà

Let's play

Example

Túy quyền và những bước đi. Ông A nhậu quá chén và tìm đường về nhà. Vì **mất quyền điều khiển hệ thống** nên ông bước đi loạng choạng. Giả sử rằng xác suất tại mỗi thời điểm ông A tiến một bước là p và lùi một bước là $1 - p$.

- Giả sử rằng tại thời điểm ông A đứng dựa vào tường
- Trên đường về nhà không có vật cản nào cả
- Nếu ông A đâm đầu vào tường thì nằm im không đi nữa
- Nếu ông A về đến nhà thì nằm im không đi nữa

Xác suất chuyển trạng thái qua nhiều bước

- Nhắc lại trạng thái hôm nay chỉ phục thuộc vào trạng hôm qua đó
- Làm sao tính được trạng thái ngày kia?
- $P(X_{m+2} = j | X_m = i) = ?$
- Gọi $P_{ij}^2 = P(X_{m+2} = j | X_m = i)$ *(2) là ký hiệu không phải bình phương

Chapman-Kolmogorov's equation

- Vận dụng luật của tổng xác suất ta có:

$$P_{ij}^2 = \sum_{k=0}^{\infty} P(X_{n+2} = j | X_{n+1} = k, X_n = i) P(X_{n+1} = k | X_n = i) \tag{7}$$

- Quy luật "ngày mai", "hôm nay", "hôm qua":

$$P_{ij}^2 = \sum_{k=0}^{\infty} P(X_{n+2} = j | X_{n+1} = k) P(X_{n+1} = k | X_n = i) \tag{8}$$

$$P_{ij}^2 = \sum_{k=0}^{\infty} P_{kj} P_{ik} \tag{9}$$

Chapman-Kolmogorov's equation

- Vậy $P(ij)^{m+n} = P(X_{n+m} = j|X_0 = i) = ?$

$$P_{ij}^{m+n} = \sum_{k=0}^{\infty} P(X_{m+n} = j|X_m = k, X_0 = i)P(X_m = k|X_0 = i) \tag{10}$$

- Quy luật "ngày mai", "hôm nay", "hôm qua":

$$P_{ij}^{m+n} = \sum_{k=0}^{\infty} P(X_{m+n} = j|X_m = k)P(X_m = k|X_0 = i) \tag{11}$$

$$P_{ij}^{m+n} = \sum_{k=0}^{\infty} P_{kj}^{n} P_{ik}^{m}, \forall i, j, n, m \geq 0 \tag{12}$$

Chapman-Kolmogorov's equation

- Đặt $\mathbf{P}^{(m)}$ với các phần tử là P_{ij}^m
- Đặt $\mathbf{P}^{(n)}$ với các phần tử là P_{ij}^n
- Đặt $\mathbf{P}^{(m+n)}$ với các phần tử là P_{ij}^{m+n}
- Mặt khác

$$\mathbf{P}_{ij}^{m+n} = \sum_{k=0}^{\infty} P_{ik}^m P_{kj}^n = \mathbf{P}_{ik}^m \mathbf{P}_{kj}^n \tag{13}$$

- Có nhìn thấy gì không?

Xác suất chuyển trạng thái N-lần

Theorem

Ma trận xác suất chuyển trạng thái sau n-lần được tính bởi công thức:

$$\mathbf{P}^{(n)} = \mathbf{P}^n \tag{14}$$

Từ giờ n chính là n số mũ không còn là n số lần nữa rồi

Let's play

Example

Hôm nay tôi buồn: $X_n = 1$
Hôm nay tôi vui: $X_n = 0$
Ngày mai tôi vui hay buồn, chỉ phụ thuộc vào hôm nay

- Xác xuất chuyển trạng thái

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix} \tag{15}$$

- Tính xác suất chuyển trạng thái của tôi trong 7 ngày sau, 1 tháng sau, hay 1 năm sau :/

Xác suất không có điều kiện

- Toàn bộ xác suất chuyển trạng thái đều có điều kiện $P_{ij}^n = P(X_n = j|X_0 = i)$
- Mục tiêu tìm $p_j(n) = P(X_n = j)$
- Vậy đặt $p_i(0) = P(X_0 = i)$
- **Chúng ta** có (căn cứ vào luật của tổng xác suất, và định nghĩa):

$$p_j(n) = P(X_n = j) = \sum_{i=0}^{\infty} P(X_n = j|X_0 = i)P(X_0 = i) \\ = \sum_{i=0}^{\infty} P_{ij}^n p_i(0) \tag{16}$$

Ma trận xác suất không có điều kiện

- Đặt $\mathbf{p}(n) = [p_1(n), p_2(n), \ldots]^T$
- Từ (16), ta có:

$$\mathbf{p}(n) = (\mathbf{P}^n)^T \mathbf{p}(0) \tag{17}$$

- Cái này hiểu thì hiểu, mà không hiểu thì hiểu.

Luyện tập

Cho ma trận chuyển trạng thái

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix} \tag{18}$$

- Tính và vẽ đồ thị xác suất "buồn" hay "vui" trong 30 ngày tới.
  Gợi ý: Công thức số (17) + python, numpy, or torch

Mô phỏng trò chơi với MC

- Mỗi lượt chơi, bạn chỉ được đặt cược 1 USD
- Xác suất thắng là p, bạn nhận được 1 USD
- Nếu thua bạn mất 1 USD,xác suất: $q = 1 - p$
- Bạn phải chơi đến khi hết sạch tiền trong túi,
- Hoặc, khi thắng đến một số tiền định trước N, hoặc quá T lượt chơi!
- Biết rằng, số tiền ban đầu bạn có là i
- Viết chương trình tính số tiền còn lại của bạn sau T lượt chơi. Tương ứng với $p = 0.25, 0.5, 0.75, i = 20, T = 1000$

Mô phỏng trò chơi với MC

- Tính xác suất bạn phải ra về tay trắng?
- Tính xác suất bạn ra về với N USD
- Liệu số tiền ban đầu có ảnh hưởng đến kết quả thắng hay thua không?
- Viết chương trình tính xác suất thua trận với $p = 0.1 \sim 0.8$, với step $\epsilon = 0.2$, $p_n = p_{n-1} + \epsilon$, hay $p = \{0.1, 0.32, 0.34, 0.36, \ldots\}$, và vẽ đồ thị.