Chuỗi Markov và ứng dụng dự đoán từ tiếp theo

Báo cáo giữa kỳ

Nhóm 9:
21120511-Lê Nguyễn
21120355-Nguyễn Anh Tú
21120312-Phan Nguyên Phương
21120143-Vũ Minh Thư

Trường Đại học Khoa học Tự nhiên VNU-HCM
Khoa Công nghệ thông tin
Môn: Phương pháp toán cho Trí tuệ nhân tạo

Mục lục

Mục lục

Mở đầu

Định nghĩa

Một **quá trình ngẫu nhiên** là một chuỗi biến ngẫu nhiên $(X_t)_{t\geq 0}$ trên một không gian mẫu Ω, được đánh số bởi thời gian t và các biến ngẫu nhiên có giá trị nằm trong một tập hợp S, gọi là **không gian trạng thái**.

Ví dụ: Xét X_n là vị trí của một con cua tại thời điểm $t = n$ (di chuyển ngang, xem như di chuyển trên trục x). Con cua có xác suất 0.67 để di chuyển sang trái, có xác suất 0.33 để di chuyển sang phải, tại thời điểm đầu tiên, con cua đứng yên nên $X_0 = 0$, tại thời điểm tiếp theo, con cua sẽ di chuyển sang trái nên $X_1 = -1$, nhưng thời điểm tiếp theo nữa, con cua lại di chuyển sang phải nên $X_2 = 0$, và cứ ngẫu nhiên như thế.

Mở đầu

- Trong phạm vi này, ta chỉ xét quá trình ngẫu nhiên **thời gian rời rạc**, nghĩa là t có giá trị từ $0, 1, 2, \dots$ hay nói cách khác, $t \in E$ với $E \subseteq \mathbb{N}$.
- Để một quá trình ngẫu nhiên trở thành một chuỗi Markov (ta chỉ xét chuỗi Markov thời gian rời rạc), ta cần thoả 3 điều kiện dưới đây:
  - Là quá trình ngẫu nhiên thời gian rời rạc.
  - Không gian trạng thái S là một tập hợp đếm được (ta sẽ tập trung xét S là hữu hạn) và X_i là biến ngẫu nhiên rời rạc.
  - Có **tính chất Markov**.
- Ta có thể hiểu tính chất Markov đơn giản như sau "Xác suất của trạng thái tiếp theo của quá trình ngẫu nhiên chỉ phụ thuộc vào hiện tại mà không phụ thuộc vào quá khứ".

Định nghĩa

Định nghĩa

Một quá trình ngẫu nhiên thoả mãn **Tính chất Markov** nếu:

$$\begin{aligned} &\mathbb{P}(X_{n+1} = s \mid X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) \\ =&\mathbb{P}(X_{n+1} = s \mid X_n = x_n) \end{aligned}$$

với mọi $n \geq 0$ và với mọi $s, x_0, x_1, \ldots, x_n \in S$. Và ta gọi quá trình ngẫu nhiên trên là một **Chuỗi Markov**.

Định nghĩa

Định nghĩa

Một chuỗi Markov được nói là **đồng nhất** nếu:

$$\mathbb{P}(X_{n+1} = y \mid X_n = x) = \mathbb{P}(X_1 = y \mid X_0 = x)$$

với mọi $x, y \in S$ và với mọi $n \geq 0$.

Định nghĩa

- Từ đây trở về sau, ta sẽ chỉ dùng đến chuỗi Markov đồng nhất.
- Ngoài ra ta sẽ kí hiệu P_{xy} cho xác suất có điều kiện $P(X_1 = y \mid X_0 = x)$ $\forall x, y \in S$ và gọi là **xác suất chuyển tiếp 1-bước**.
- Và cuối cùng điều quan trọng nhất đối với một chuỗi Markov có không gian trạng thái hữu hạn hay $S = \{x_0, x_1, ..., x_n\}$, đó là **phân phối ban đầu**, đặt là π_0 và đặt $\pi_0(x_i) = \mathbb{P}(X_0 = x_i)$ với $x_i \in S$. Lúc này ta có:

$$\pi_0 = \begin{bmatrix} \pi_0(x_0) & \pi_0(x_1) & \dots & \pi_0(x_n) \end{bmatrix}$$

- Do $S = \{x_0, x_1, ..., x_n\}$ là một tập hợp hữu hạn nên thông thường ta đưa về dạng ma trận gồm các xác suất chuyển tiếp 1-bước mà chuỗi Markov có thể có, kí hiệu là **P** và gọi ma trận đó là **ma trận chuyển tiếp** hay **ma trận Markov**, ta có:

$$\mathbf{P} = \begin{bmatrix} P_{x_0x_0} & P_{x_0x_1} & \dots & P_{x_0x_n} \\ P_{x_1x_0} & P_{x_1x_1} & \dots & P_{x_1x_n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{x_nx_0} & P_{x_nx_1} & \dots & P_{x_nx_n} \end{bmatrix}$$

Định nghĩa

Một ma trận vuông được gọi là **ma trận ngẫu nhiên** nếu thoả mãn 2 điều kiện sau:

(a) Tổng các phần tử của mỗi dòng đều là 1.

(b) Mọi phần tử đều không âm.

Ta có thể thấy ma trận chuyển tiếp **P** luôn luôn là một ma trận ngẫu nhiên.

Định lý

Định lý

Cho một chuỗi Markov $(X_n)_{n\geq 0}$ với một không gian trạng thái S hữu hạn. Đặt thời gian i là thời điểm hiện tại, khi đó thời điểm quá khứ và tương lai của chuỗi Markov độc lập lẫn nhau. Nghĩa là, với mọi $n > i$ và với mọi $x_0, x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n \in S$, ta có:

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_n = x_n \mid X_i = x_i)\\ =&\mathbb{P}(X_0 = x_0, \ldots, x_{i-1} = x_{i-1} \mid X_i = x_i)\mathbb{P}(X_{i+1} = x_{i+1}, \ldots, X_n = x_n \mid X_i = x_i)\end{aligned}$$

Chứng minh

- Đặt $F = (X_0 = x_0, \ldots, X_{i-1} = x_{i-1})$, $E = (X_{i+1} = x_{i+1}, \ldots, X_n = x_n)$ và $G = (X_i = x_i)$, ta có:

$$\begin{aligned}
\mathbb{P}(E \cap F \mid G) &= \frac{\mathbb{P}(E \cap F \cap G)}{\mathbb{P}(G)} \\
&= \frac{\mathbb{P}(E \cap F \cap G)}{\mathbb{P}(F \cap G)} \frac{\mathbb{P}(F \cap G)}{\mathbb{P}(G)} \\
&= \mathbb{P}(E \mid F \cap G)\mathbb{P}(F \mid G) \\
&= \mathbb{P}(E \mid G)\mathbb{P}(F \mid G) \quad \text{(áp dụng tính chất của chuỗi Markov)}
\end{aligned}$$

- Vậy:

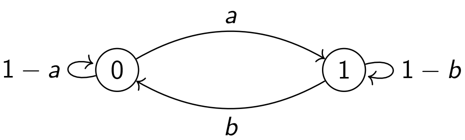
$$\begin{aligned}
&\mathbb{P}(X_0 = x_0, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_n = x_n \mid X_i = x_i) \\
=&\mathbb{P}(X_0 = x_0, \ldots, x_{i-1} = x_{i-1} \mid X_i = x_i)\mathbb{P}(X_{i+1} = x_{i+1}, \ldots, X_n = x_n \mid X_i = x_i)
\end{aligned}$$

Định lý

- Ngoài ra để có thể hình dung tốt hơn Chuỗi Markov, ta có thể biểu diễn nó thành một đồ thị.
- Xét ví dụ một chuỗi Markov chỉ có 2 trạng thái, hay nói cách khác S chỉ có 2 phần tử. Xét $S = \{0, 1\}$, ta có $P_{01} = a$ và $P_{10} = b$. Khi đó ma trận chuyển tiếp **P** sẽ là:

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

- Đồ thị biểu diễn cho chuỗi Markov là một đồ thị $G = (V, E)$ (với $V = S$) *có hướng* và *có trọng số* với các trọng số chính là các xác suất chuyển tiếp từ trạng thái này sang trạng thái khác hay từ đỉnh này sang đỉnh khác.



Hình: Đồ thị G biểu diễn của chuỗi Markov 2 trạng thái

Mục lục

Lũy thừa ma trận

- Xét một chuỗi Markov có $S = \{x_1, x_2, ..., x_n\}$ và một phân phối ban đầu π_0.
- Để tính toán được π_1, áp dụng công thức xác suất đầy đủ, ta có:

$$\begin{aligned}\pi_1(x_i) &= \mathbb{P}(X_1 = x_i) \\ &= \sum_{j=0}^{n} \mathbb{P}(X_0 = x_j)\mathbb{P}(X_1 = x_i \mid X_0 = x_j) \\ &= \sum_{j=0}^{n} \pi_0(x_j) P_{x_j x_i}\end{aligned}$$

- Ta có thể thấy $\pi_1(x_i)$ chính là tích vô hướng giữa vector dòng π_0 và cột thứ i của ma trận $\mathbf{P}$. Vậy ta có:

$$\pi_1 = \begin{bmatrix}\pi_1(x_1) & \pi_1(x_2) & \cdots & \pi_1(x_n)\end{bmatrix} = \pi_0\mathbf{P}$$

- Tiếp theo để tính được $\pi_2(x_i)$ ta tiếp tục áp dụng công thức xác suất đầy đủ:

$$\begin{aligned}
\pi_2(x_i) &= \mathbb{P}(X_2 = x_i) \\
&= \sum_{j=0}^{n} \mathbb{P}(X_1 = x_j)\mathbb{P}(X_2 = x_i \mid X_1 = x_j) \\
&= \sum_{j=0}^{n} \pi_1(x_j)\mathbb{P}(X_1 = x_i \mid X_0 = x_j) \\
&= \sum_{j=0}^{n} \pi_1(x_j) P_{x_j x_i} \\
&= \sum_{j=0}^{n} \left(\sum_{k=0}^{n} \pi_0(x_k) P_{x_k x_j} \right) P_{x_j x_i} \\
&= \sum_{k=0}^{n} \pi_0(x_k) \left(\sum_{j=0}^{n} P_{x_k x_j} P_{x_j x_i} \right)
\end{aligned}$$

Lũy thừa ma trận

- Ta có thể thấy $\pi_2(x_j)$ chính là tích vô hướng giữa vector dòng π_0 và cột thứ i của ma trận $\mathbf{P}^2$. Vậy ta có:

$$\pi_2 = \begin{bmatrix} \pi_2(x_1) & \pi_2(x_2) & \dots & \pi_2(x_n) \end{bmatrix} = \pi_0 \mathbf{P}^2$$

- Ta gọi $\mathbb{P}(X_2 = y \mid X_0 = x)$ là *xác suất chuyển tiếp 2-bước*. Tương tự với $\mathbb{P}(X_n = y \mid X_0 = x)$ ta gọi là *xác suất chuyển tiếp n-bước*.
- Dựa vào xác suất chuyển tiếp 2-bước và 1-bước, ta có thể doán được:

$$\pi_n = \pi_0 \mathbf{P}^n \tag{1}$$

Chứng minh

Ta sẽ dùng quy nạp để chứng minh. Đầu tiên ta đã chứng minh được phương trình (1) đúng với $n = 1$ và $n = 2$, giả sử $n = k$ đúng, nghĩa là:

$$\pi_k = \pi_0 \mathbf{P}^k$$

Xét $n = k + 1$, ta có:

$$\begin{aligned}
\pi_{k+1}(x_i) &= \mathbb{P}(X_{k+1} = x_i) \\
&= \sum_{j=0}^{n} \mathbb{P}(X_k = x_j)\mathbb{P}(X_{k+1} = x_i \mid X_k = x_j) \\
&= \sum_{j=0}^{n} \mathbb{P}(X_k = x_j)\mathbb{P}(X_1 = x_i \mid X_0 = x_j) = \sum_{j=0}^{n} \pi_k(x_j) P_{x_j x_i}
\end{aligned}$$

Ta có thể thấy $\pi_{k+1}(x_i)$ là tích vô hướng giữa vector dòng π_k và cột thứ i của ma trận chuyển tiếp $\mathbf{P}$, do đó:

$$\pi_{k+1} = \pi_k \mathbf{P} = \pi_0 \mathbf{P}^k \mathbf{P} = \pi_0 \mathbf{P}^{k+1}$$

Vậy theo quy nạp, phương trình (1) đúng với mọi $n \geq 1$.

Lũy thừa ma trận

Định nghĩa

Cho $(X_n)_{n\geq 0}$ là một chuỗi Markov với ma trận chuyển tiếp $\mathbf{P}$. Với mọi $n \geq 1$, ta gọi ma trận $\mathbf{P}^n$ là **ma trận chuyển tiếp n-bước**. Các phần tử của ma trận $\mathbf{P}^n$ được gọi là **xác suất chuyển tiếp n-bước**. Với mọi $x, y \in S$ ta kí hiệu xác suất chuyển tiếp n bước từ x tới y là P^n_{xy}.

- Để dễ dàng hơn trong việc tính toán ma trận $\mathbf{P}^n$ ta có thể dùng phương pháp chéo hoá.
- Dùng tính chất của luỹ thừa, ta có:

$$\pi_{n+m} = \pi_0 \mathbf{P}^{n+m} = \pi_0 \mathbf{P}^n \mathbf{P}^m$$

- Ngoài ra ta có:

$$\mathbb{P}(X_n = y \mid X_0 = x) = P^n_{xy} \Rightarrow \mathbb{P}(X_{m+n} = y \mid X_m = x) = P^n_{xy}$$

- Các trị riêng của ma trận Markov có một số tính chất đặc biệt, các tính chất này giúp chúng ta có thể giải thích các tính chất của chuỗi Markov bằng ngôn ngữ của đại số tuyến tính một cách trực quan và dễ dàng hơn.

Định lý

Ma trận markov luôn có ít nhất một trị riêng bằng 1.

Chứng minh

- Với ma trận Markov $\mathbf{P}$ có kích thước $n \times n$, xét phương trình:

$$\mathbf{P}\mathbf{x} = \mathbf{x} \tag{2}$$

$$(\mathbf{P} - I)\mathbf{x} = 0. \tag{3}$$

- Với b_{ij} là phẩn tử hàng i cột j của $\mathbf{P} - I$, $k \leq n$, ta có:

$$\sum_{j=1}^{n} b_{kj} = \sum_{j=1}^{n} P_{kj} - 1 = 1 - 1 = 0. \tag{4}$$

- Suy ra, với $\mathbf{u}_i$ là các vector cột của $(\mathbf{P} - \lambda I)$, ta có:

$$\sum_{i=1}^{n} \mathbf{u}_i = 0. \tag{5}$$

- Suy ra các các vector cột của $(\mathbf{P} - \lambda I)$ phụ thuộc tuyến tính, từ đó suy ra $r(\mathbf{P}) < n$, hay phương trình (2) luôn có nghiệm.
- Vậy, vì $\lambda = 1$ thoả $\mathbf{P}\mathbf{x} = 1\mathbf{x}$ nên 1 là trị riêng của $\mathbf{P}$.

Trị riêng của ma trận

Định lý

Các trị riêng của ma trận Markov luôn bé hơn hoặc bằng 1.

Chứng minh

- Với ma trận Markov **P** có kích thước $n \times n$, λ là 1 trị riêng của **P**, ta có:

$$\mathbf{P}\mathbf{x} = \lambda \mathbf{x}$$

- Xét hàng thứ k của cả 2 vế, ta có:

$$\sum_{j=1}^{n} P_{kj} x_j = \lambda x_k$$

- Đặt phần tử x_m thoả:

$$|x_m| = \max(|x_1|, |x_2|, ..., |x_n|)$$

Chứng minh: (tt)

- Lúc này, ta có:

$$|\lambda x_m| = \left|\sum_{j=1}^{n} P_{mj}x_j\right| \le \sum_{j=1}^{n} |P_{mj}x_j| \le \sum_{j=1}^{n} |P_{mj}x_m|$$
$$= |x_m|\sum_{j=1}^{n} |P_{mj}| = |x_m|.1$$

- Suy ra $\lambda \le 1$

Ví dụ tiêu biểu

Bài toán

Theo khảo sát của sinh viên đối với ba quán cafe A, B, C, ta biết rằng ban đầu, hai quán A, B chưa mở nên 100% khách đều đến C và:

- Trong những sinh viên đến quán A, sẽ có 20% người tiếp tục đến A, có 60% người sang B và có 20% người sang C.
- Trong những sinh viên đến quán B, sẽ có 40% người tiếp tục đến B, có 10% người sang A và có 50% người sang C.
- Trong những sinh viên đến quán C, sẽ có 10% người tiếp tục đến C, có 70% người sang A và có 20% người sang B.

(a) Hãy tìm xem tỉ lệ phần trăm người đến quán A, B, C sau 3 tuần.

(b) Hãy tìm xem tỉ lệ sinh viên đến quán B ở tuần thứ 5 biết rằng tuần thứ 2 ở quán C và tuần thứ 3 ở quán A.

(c) Hãy tìm xem tỉ lệ sinh viên đến quán B ở tuần thứ 2 và đến quán A ở tuần thứ 4 biết rằng đến quán B ở tuần thứ 3.

Ví dụ tiêu biểu

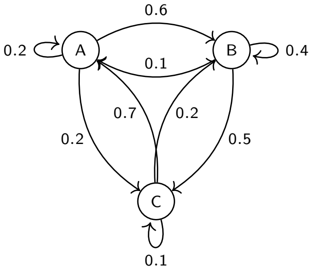
Thời gian chúng ta xét sẽ là tuần. Tiếp theo ta có không gian trạng thái $S = \{A, B, C\}$. Đặt X_t là biến ngẫu nhiên đại diện cho quán cafe mà sinh viên đến ở tuần thứ t và có phân phối ban đầu π_0 là

$$\pi_0 = \begin{bmatrix} \mathbb{P}(X_0 = A) & \mathbb{P}(X_0 = B) & \mathbb{P}(X_0 = C) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

Dựa theo đề bài ta có ma trận chuyển tiếp **P** là:

$$\mathbf{P} = \begin{bmatrix} P_{AA} & P_{AB} & P_{AC} \\ P_{BA} & P_{BB} & P_{BC} \\ P_{CA} & P_{CB} & P_{CC} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix}$$

Ta biểu diễn chuỗi Markov trên thành đồ thị như dưới đây:



Hình: Đồ thị G biểu diễn của chuỗi Markov của bài toán

Ví dụ tiêu biểu

Tính toán luỹ thừa 2 và 3 của ma trận chuyển tiếp **P**, ta được:

$$\mathbf{P}^2 = \begin{bmatrix} 0.24 & 0.4 & 0.36 \\ 0.41 & 0.32 & 0.27 \\ 0.23 & 0.52 & 0.25 \end{bmatrix} \quad \text{và} \quad \mathbf{P}^3 = \begin{bmatrix} 0.34 & 0.376 & 0.284 \\ 0.303 & 0.428 & 0.269 \\ 0.273 & 0.396 & 0.331 \end{bmatrix}$$

(a) Ta có tỉ lệ sinh viên đến quán A, B, C ở tuần thứ 3 chính là π_3, do đó:

$$\pi_3 = \pi_0 \mathbf{P}^3 = \begin{bmatrix} 0.273 & 0.396 & 0.331 \end{bmatrix}$$

(b) Tỉ lệ sinh viên đến quán B ở tuần thứ 5 biết rằng tuần thứ 2 ở quán C và tuần thứ 3 ở quán A chính là:

$$\begin{aligned} \mathbb{P}(X_5 = B \mid X_2 = C, X_3 = A) &= \mathbb{P}(X_5 = B \mid X_3 = A) \\ &= \mathbb{P}(X_2 = B \mid X_0 = A) = P^2_{AB} = 0.4 \end{aligned}$$

(c) Tỉ lệ sinh viên đến quán B ở tuần thứ 2 và đến quán A ở tuần thứ 4 biết rằng đến quán B ở tuần thứ 3 chính là:

$$\begin{aligned} \mathbb{P}(X_2 = B, X_4 = A \mid X_3 = B) &= \mathbb{P}(X_4 = A \mid X_3 = B)\mathbb{P}(X_2 = B \mid X_3 = B) \\ &= \mathbb{P}(X_1 = A \mid X_0 = B)\frac{\mathbb{P}(X_3 = B \mid X_2 = B)\mathbb{P}(X_2 = B)}{\mathbb{P}(X_3 = B)} \\ &= P_{BA}P_{BB}\frac{\pi_2(B)}{\pi_3(B)} \approx 0.052525 \end{aligned}$$

Ví dụ tiêu biểu

Bài toán

Một khảo sát được thực hiện trên 100 người dùng điện thoại. Ban đầu có 60 người dùng điện thoại sử dụng hệ điều hành Android, 40 người dùng điện thoại sử dụng hệ điều hành IOS. Sau mỗi quý, sẽ có 12 người chuyển từ Android sang IOS, 4 người chuyển từ IOS sang Android. Hãy tìm xem sau 50 quý thì tỉ lệ người dùng ở hai hệ điều hành sẽ như nào, ngoài ra sau 100 quý hay 200 quý hay 500 quý thì có thay đổi gì không ?

Lời giải

Thời gian chúng ta xét sẽ là quý. Tiếp theo ta có không gian trạng thái $S = \{A, I\}$ với A viết tắt cho Android và I cho IOS. Đặt X_t là biến ngẫu nhiên cho hệ điều hành mà người dùng sử dụng ở quý thứ t và có phân phối ban đầu π_0 là:

$$\pi_0 = \begin{bmatrix} \mathbb{P}(X_0 = A) & \mathbb{P}(X_0 = I) \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$$

Dựa theo đề bài ta có ma trận chuyển tiếp **P** là:

$$\mathbf{P} = \begin{bmatrix} P_{AA} & P_{AI} \\ P_{IA} & P_{II} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix}$$

Ví dụ tiêu biểu

Lời giải: (tt)

Khi đó tỉ lệ người dùng sau 50 quý hay nói cách khác là phân phối tại $t = 50$ là:

$$\pi_{50} = \pi_0 \mathbf{P}^{50} = \begin{bmatrix} 0.333 & 0.6667 \end{bmatrix}$$

Tiếp theo tỉ lệ người dùng sau 100 quý sẽ là:

$$\pi_{100} = \pi_0 \mathbf{P}^{100} = \begin{bmatrix} 0.33333333 & 0.66666667 \end{bmatrix}$$

Tỉ lệ người dùng sau 500 quý sẽ là:

$$\pi_{500} = \pi_0 \mathbf{P}^{500} = \begin{bmatrix} 0.33333333 & 0.66666667 \end{bmatrix}$$

Ta có thể thấy, kể từ 100 quý trở về sau, phân phối sẽ không thay đổi nữa. Điều này làm ta có thể nghĩ đến việc phân phối của chuỗi Markov có "giới hạn" và phân phối giới hạn đó được gọi là **phân phối bất động** (stationary distribution), kí hiệu là π:

$$\pi_n = \pi_0 \mathbf{P}^{\mathbf{n}} \xrightarrow{n \to \infty} \pi$$

Trong trường hợp của bài toán này phân phối bất động chính là:

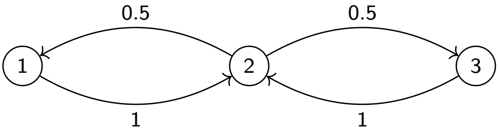
$$\pi = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

Sự hội tụ của chuỗi Markov dưới góc nhìn ma trận

Định nghĩa

Một chuỗi Markov được gọi là **chuỗi chính tắc** nếu tồn tại một số $n \geq 1$ sao cho các phần tử của $\mathbf{P}^n$ đều dương (lớn hơn 0).

- Xét chuỗi Markov sau:



Hình: Ví dụ chuỗi Markov

- Dựa vào đồ thị ta có ma trận chuyển tiếp:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix}$$

Định lý

Nếu chuỗi markov là một chuỗi chính tắc và ma trận chuyển tiếp **P** chéo hoá được thì tồn tại phân phối bất động π. Đặt e là vector riêng ứng với trị riêng bằng 1 của ma trận $\mathbf{P}^T$, khi đó $\pi = e^T$.

Chứng minh

Vì chúng ta vẫn thường thao tác với không gian cột, nên bài chứng minh sẽ xét với ma trận chuyển vị từ hàng thành cột để dễ dàng hình dung hơn.

- Xét ma trận Markov $\mathbf{P}$ chéo hoá được thành $\mathbf{P} = UDU^{-1}$, ta có $\mathbf{P}^T$ cũng chéo hoá được, chứng minh như sau:

$$\mathbf{P}^T = (UDU^{-1})^T = (U^{-1})^T D^T U^T = (U^T)^{-1} D^T U^T \tag{6}$$

- Đặt $(U^{-1})^T = X$, đồng thời lại có $D^T = D$, từ đó ta được:

$$\mathbf{P}^T = (U^{-1})^T D^T U^T = XDX^{-1} \tag{7}$$

- Suy ra $\mathbf{P}^T$ chéo hoá được. Xét ma trận $\mathbf{P}^T$, X là ma trận có các cột là các vector riêng, cột đầu tiên là vector riêng ứng với $\lambda_1 = 1$, ta được:

$$\mathbf{P}^T = XDX^{-1} \tag{8}$$

Chứng minh: (tt)

- Với $\mathbf{x}_i$ là các cột của X, λ_i là các trị riêng tương ứng với $\mathbf{x}_i$, ta được

$$X = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{bmatrix}, D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{9}$$

- Lúc này, $\mathbf{x}_i$ là các cột của X, phân phối ban đầu π_0 sẽ được biểu diễn thành tổ hợp tuyến tính của x_i như sau (vì P^T chéo hoá được nên ta sẽ có các vector x_i là cớ sở của không gian R^{nxn}):

$$\pi_0^T = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \dots + c_n\mathbf{x}_n \tag{10}$$

$$\pi_0^T = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_n \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \tag{11}$$

$$\mathbf{c} = X^{-1}\pi_0^T \tag{12}$$

Chứng minh: (tt)

- Vector trạng thái sau thời gian $k+1$:

$$\pi_{k+1} = \pi_0 \mathbf{P}^k \tag{13}$$
$$\Leftrightarrow \quad \pi_{k+1}^T = (\mathbf{P}^T)^k \pi_0^T \tag{14}$$
$$\Leftrightarrow \quad \pi_{k+1}^T = XD^k X^{-1} \pi_0^T \tag{15}$$
$$\Leftrightarrow \quad \pi_{k+1}^T = XD^k \mathbf{c} \tag{16}$$

- Ta có:

$$XD^k = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{bmatrix} \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} \tag{17}$$

Chứng minh: (tt)

- Sử dụng phương pháp nhân hai ma trận (ma trận này nhân với từng cột ma trận kia), ta được:

$$XD^k = \begin{bmatrix} \lambda_1^k \mathbf{x}_1 & \lambda_2^k \mathbf{x}_2 & \dots & \lambda_n^k \mathbf{x}_n \end{bmatrix} \tag{18}$$

- Cùng với quy ước $\lambda_1 = 1$, ta suy ra:

$$\pi_{k+1}^T = XD^k C = \begin{bmatrix} \lambda_1^k \mathbf{x}_1 & \lambda_2^k \mathbf{x}_2 & \dots & \lambda_n^k \mathbf{x}_n \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \tag{19}$$

$$= c_1 \mathbf{x}_1 + c_2 \lambda_2^k \mathbf{x}_2 + \dots + c_n \lambda_n^k \mathbf{x}_n \tag{20}$$

Sự hội tụ của chuỗi Markov dưới góc nhìn ma trận

Chứng minh: (tt)

- Như đã chứng minh ở trên, các trị riêng khác 1 và -1 đều có trị tuyệt đối nhỏ hơn 1, chính vì thế khi k càng tăng đến số vô cùng lớn, thì π_{k+1}^T sẽ càng tiến gần về $c_1\mathbf{x}_1$. Và đây chính là lí do cho sự "hội tụ" của các phân phối.
- Vậy mới một ma trận Markov chéo hoá được, các phân phối sẽ hội tụ dần về phân phối bất động với phân phối bất động chính là vector riêng tương ứng với trị riêng bằng 1.

Mục lục

Đặt vấn đề

- Trong rất nhiều các công cụ tìm kiếm, cũng như các phần mềm cho ứng dụng tìm kiếm hiện nay, chúng ta có thể dễ dàng nhận ra được một chức năng rất phổ biến, đó chính là đưa ra những đề xuất về từ ngữ tiếp theo dựa vào những từ vừa được gõ. Từ đó chúng ta có bài toán làm thế nào để tìm ra đâu là những từ có xác suất xuất hiện tiếp theo cao nhất.
- Có rất nhiều những phương pháp có thể được áp dụng trong việc giải quyết bài toán trên một cách hiệu quả chẳng hạn như Deep Learning. Tuy nhiên, chúng ta cũng có thể giải quyết bài toán này (một cách đơn giản, song không tối ưu như DL) bằng chuỗi Markov.

Chuẩn bị và xử lý dữ liệu

- Đầu tiên, từ một tập dữ liệu (dataset) gồm những từ, câu đã được tìm kiếm (lấy từ database của công cụ tìm kiếm) hoặc từ những từ đã được gõ (lấy từ database của phần mềm bàn phím điện thoại như LabanKey), chúng ta sẽ loại bỏ các kí tự đặt biệt (dấu câu,..), sau đó phân tách từng câu ra thành cụm lần lượt có 1, 2 và 3 từ, sau đó xây dựng ma trận xác suất (sẽ được trình bày ở phần sau).
- Để dễ hình dung cách hoạt động của thuật toán, trong quy mô bài báo cáo này, chúng tôi xin được minh hoạt bằng 1 dataset lấy từ một vài câu comment trên Facebook như sau:

"Tin này là tin chuẩn chưa anh?"
"Này là tin chuẩn em ạ!"
"Đừng spam tin này nữa em nhé!"
"Tin này chưa chuẩn em nhé!"
"Anh sẽ chặn các bạn spam tin này."

Dự đoán từ tiếp theo

- Phân tách thành các cụm 1 từ, ta được 16 cụm sau: 'tin', 'này', 'là', 'chuẩn', 'chưa', 'anh', 'em', 'ạ', 'đừng', 'spam', 'nữa', 'nhé', 'sẽ', 'chặn', 'các', 'bạn'.
- Phân tách thành các cụm 2 từ, ta được 20 cụm sau: 'tin này', 'này là', 'là tin', 'tin chuẩn', 'chuẩn chưa', 'chưa anh', 'chuẩn em', 'em ạ', 'đừng spam', 'spam tin', 'này nữa', 'nữa em', 'em nhé', 'này chưa', 'chưa chuẩn', 'anh sẽ', 'sẽ chặn', 'chặn các', 'các bạn', 'bạn spam'.
- Phân tách thành các cụm 3 từ, ta được 21 cụm sau: 'tin này là', 'này là tin', 'là tin chuẩn', 'tin chuẩn chưa', 'chuẩn chưa anh', 'tin chuẩn em', 'chuẩn em ạ', 'đừng spam tin', 'spam tin này', 'tin này nữa', 'này nữa em', 'nữa em nhé', 'tin này chưa', 'này chưa chuẩn', 'chưa chuẩn em', 'chuẩn em nhé', 'anh sẽ chặn', 'sẽ chặn các', 'chặn các bạn', 'các bạn spam', 'bạn spam tin'.

Xây dựng ma trận chuyển tiếp

- Từ những cụm từ chúng ta đã phân tách ở phần trên, chúng ta có thể xây dựng các ma trận chuyển trạng thái biểu diễn xác suất những từ sẽ xuất hiện tiếp theo dựa vào 1 hoặc 2 hoặc 3 từ trước đó.
- Ma trận trạng thái sẽ được định nghĩa như sau: phần tử hàng i cột j sẽ biểu diễn xác suất cho sự xuất hiện của trạng thái tiếp theo j tương ứng với trạng thái hiện tại i. Xác suất đó sẽ được tính bằng số lần xuất hiện của trạng j sau trạng thái i chia cho tổng số lần xuất hiện của các trạng thái.

Tiến hành dự đoán từ tiếp theo

- Đầu tiên, với từ ngữ đầu tiên được nhập (từ bàn phím), ta có thể thành lập 1 vector phân phối các trạng thái hiện tại. Vì mới chỉ có 1 từ nên chúng ta sử dụng ma trận phân phối với các trạng thái hiện tại là 1 từ, ví dụ cụ thể như sau:
  - Với từ đầu tiên được nhập là "tin", thứ tự tương ứng với thứ tự các trạng thái trong ma trận chuyển tiếp ở trên, ta được vector phân phối:

$$\pi_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Từ đó, ta có thể xác định phân phối tiếp theo:

$$\pi_1 = \pi_0 P_1 = \begin{bmatrix} 0 & 2/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Vậy, ta có thể đề xuất những từ có xác suất cao nhất:
  Từ "này" với xác suất $\frac{2}{3}$.
  Từ "chuẩn" với xác suất $\frac{1}{3}$.

Dự đoán từ tiếp theo

- Nếu người dùng đã viết được ít nhất 2 từ, để tăng độ chính xác khi đề xuất từ cũng như đảm bảo một phần yếu tố ngữ nghĩa của câu, ta sử dụng ma trận chuyển tiếp của trạng thái 2 từ và ma trận chuyển trạng thái 3 từ. Từ đó, ta có thể xây dựng thuật toán như sau:
  - Với những câu chỉ có một từ, sử dụng ma trận chuyển tiếp với trạng thái 1 từ.
  - Với những câu có nhiều hơn 2 từ, lần lượt xét 3 từ cuối của câu xem đã tồn tại thái là 3 từ đấy chưa, nếu có, sử dụng ma trận chuyển tiếp với trạng thái 3 từ, nếu chưa thì chuyển sang xét 2 từ cuối và ma trận chuyển tiếp với trạng thái 2 từ, và cứ như thế cho đến khi còn 1 từ.
  - Sau khi hoàn thành, tiếp tục cập nhật các trạng thái và xác suất mới vào dataset.

Một số ví dụ:

- Với cụm từ đã gõ "Anh cho em hỏi tin này là", ta sẽ xét 3 từ cuối là "tin này là", nhận thấy đây là 1 trạng thái trong ma trận chuyển tiếp, áp dụng mô hình ở trên.

→ Từ đó, ta đề xuất từ "tin" cho người dùng với xác suất bằng 1.

- Với cụm từ đã gõ "Đây có phải tin chuẩn", ta sẽ xét 3 từ cuối là "phải tin chuẩn", nhận thấy đây không là 1 trạng thái trong ma trận chuyển tiếp, ta xét 2 từ cuối là "tin chuẩn", nhận thấy đây là một trạng thái của ma trận chuyển tiếp 2 từ, áp dụng mô hình.

→ Từ đó, ta đề xuất từ "tin" cho người dùng với xác suất bằng 0.5, hoặc từ "em" với xác suất bằng 0.5.