Quá trình quyết định Markov

Các mô hình ngẫu nhiên và ứng dụng

Ngô Quốc Trần Hiếu
Nguyễn Đức Minh
Lê Thị Duyên

Ngày 4 tháng 1 năm 2020

Mục lục

Quá trình quyết định Markov

Định nghĩa 1

Cho X là một quá trình mô tả hệ thống (system description process) với không gian trạng thái E và cho D là một quá trình quyết định (decision process) với không gian hành động A. Quá trình (X, D) là *Quá trình Quyết định Markov* nếu, với mọi $j \in E$ và $n = 0, 1, \ldots$ ta đều có

$$P(X_{n+1} = j | X_0, D_0, \ldots, X_n, D_n) = P(X_{n+1} = j | X_0, D_0).$$

Hơn nữa, với mỗi $k \in A$, cho f_k là một véc tơ chi phí, P là ma trận Markov. Khi đó:

$$P(X_{n+1} = j | X_n = i, D_n = k) = P_k(i, j)$$

và chi phí $f_k(i)$ phát sinh mỗi khi $X_n = i$ và $D_n = k$.

Ví dụ 1

Cho quá trình Markov (X,D) có $E = \{1, 2, 3, 4\}, A = \{1, 2\}$

$$f_1 = (100, 125, 150, 500)^T,$$
$$f_2 = (300, 325, 350, 600)^T,$$
$$P_1 = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0.0 \\ 0.0 & 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.2 & 0.7 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{bmatrix},$$
$$P_2 = \begin{bmatrix} 0.6 & 0.3 & 0.1 & 0.0 \\ 0.75 & 0.1 & 0.1 & 0.05 \\ 0.8 & 0.2 & 0.0 & 0.0 \\ 0.9 & 0.1 & 0.0 & 0.0 \end{bmatrix}.$$

Chính sách

Định nghĩa 2

Một *chính sách (policy)* là một tập quy tắc, sử dụng thông tin hiện tại, thông tin quá khứ, và/hoặc ngẫu nhiên chỉ định hành động nào được thực hiện tại mỗi thời điểm. Tập tất cả các chính sách được biểu diễn bởi $\mathcal{M}$.

Ví dụ chính sách

- Chính sách 1. Luôn chọn hành động 1, không phụ thuộc vào trạng thái của X, tức $D_n \equiv 1$ với mọi n
- Chính sách 2. Nếu X_n ở trạng thái a hoặc b, cho $D_n = 1$; nếu X_n ở trạng thái c hoặc d, cho $D_n = 2$.
- Chính sách 3. Nếu X_n ở trạng thái a hoặc b, cho $D_n = 1$; nếu X_n ở trạng thái c, tung một đồng xu và cho $D_n = 1$ nếu đồng xu xấp, cho $D_n = 2$ nếu đồng xu ngửa; nếu X_n ở trạng thái d thì cho $D_n = 2$.
- Chính sách 4. Cho $D_n \equiv 1$ nếu $n = 0$ và 1. Với $n \geq 2$, nếu $X_n > X_{n-1}$ và $X_{n-1} = a$, cho $D_n = 1$. Nếu $X_n > X_{n-1}, X_{n-2} = b$ và $D_{n-1} = 2$ cho $D_n = 1$, ngược lại, $D_n = 2$.

Kỳ vọng tổng chi phí có chiết khấu

Cho α là tỷ lệ chiết khẩu sao cho 1 tại thời điểm $n = 1$ có giá trị hiện tại bằng α tại thời điểm $n = 0$. (Trong kinh tế, thông thường $\alpha = \frac{1}{r+1}$ với r là lãi suất). Kỳ vọng tổng chi phí có chiết khấu cho một quá trình quyết định Markov được cho bởi công thức

$$E(\sum_{n=0}^{\infty} \alpha^n f_{D_n}(X_n)).$$

Kỳ vọng tổng chi phí có chiết khấu

Ký hiệu $E_d[.]$ là kỳ vọng tổng chi phí có chiết khấu nếu sử dụng chính sách $d, d \in \mathcal{M}$. Đặt

$$v_d^\alpha = E_d(\sum_{n=0}^{\infty} \alpha^n f_{D_n}(X_n)|X_0 = i)$$

với mọi $i \in E$ và $0 < \alpha < 1$. Từ đó nảy sinh bài toán tìm $d^\alpha \in \mathcal{M}$ sao cho

$$v_{d^\alpha}^\alpha(i) = v^\alpha(i) = \min_{d \in \mathcal{M}} v_d^\alpha(i) \ \forall i \in E. \qquad (1)$$

Chi phí trung bình dài hạn

Chí phí trung bình dài hạn của một quá trình quyết định Markov khi áp dụng chính sách $d \in \mathcal{M}$ được đưa ra bởi công thức

$$\varphi_d = \lim_{m \to \infty} \frac{1}{m} \sum_{n=0}^{m-1} f_{D_n}(X_n).$$

Từ đó, nảy sinh bài toán tối ưu: Tìm $d^* \in \mathcal{M}$ sao cho

$$\varphi^* = \varphi_{d^*} = \min_{d \in \mathcal{M}} \varphi_d \tag{2}$$

Chính sách Cố định (Stationary Policies)

Định nghĩa 3

Một *hàm hành động* là một véc tơ ánh xạ từ không gian trạng thái vào không gian hành động.

Ví dụ: $a_1 = (1, 1, 1, 1), a_2 = (1, 1, 2, 2)$

Định nghĩa 4

Một *chính sách cố định* là một chính sách có thể được biểu diễn bằng một hàm hành động. Chính sách cố định được biểu diễn bằng hàm hành động a thực hiện hành động $a(i)$ tại thời điểm n nếu $X_n = i$, độc lập với các trạng thái trước, hành động trước, và thời điểm n.

Ví dụ: Chính sách 1 được biểu diễn bằng hàm hành động a_1 và chính sách 2 được biểu diễn bằng hàm hành động a_2.

Chính sách Cố định (Stationary Policies)

- Một quá trình quyết định Markov theo chính sách cố định luôn là một xích Markov.
- Một quá trình quyết định Markov theo chính sách cố định được định nghĩa bằng hàm hành động a là xích Markov có ma trận xác xuất chuyển và véc tơ chi phí là

$$P^a(i,j) = P_{a(i)}(i,j) \ \forall i,j \in E, \tag{3}$$

$$f^a(i) = f_{a(i)}(i) \ \forall i \in E. \tag{4}$$

Chính sách Cố định (Stationary Policies)

Tính chất 1

Nếu không gian trạng thái E hữu hạn, tồn tại một chính sách cố định là nghiệm của Bài toán (1.1). Thêm nữa, nếu mỗi chính sách cố định sinh ra một xích Markov không giảm, thì tồn tại một chính sách cố định là nghiệm của Bài toán (1.2). (Chính sách tối ưu phụ thuộc vào chiết khấu và có thể khác nhau đối với hai Bài toán (1.1) và (1.2.)

Chú ý. Xích Markov không giảm là xích chỉ có một lớp liên thông.

Các thuật toán chi phí có chiết khấu

Tính chất 2

Cho v^α là hàm giá trị tối ưu của Bài toán (1.1) với $0 < \alpha < 1$. Hàm v^α thỏa mãn, với mọi $i \in E$, ta đều có

$$v^\alpha(i) = \min_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i,j) v^\alpha(j)\}. \qquad (5)$$

Hơn nữa, nó là hàm duy nhất thỏa mãn tính chất này.

Tính chất 2 cung cấp cho ta một cách để kiểm tra xem một hàm có phải giá trị tối ưu của Bài toán (1) hay không

Các thuật toán chi phí có chiết khấu

Tính chất 3

Cho v^α là hàm giá trị tối ưu của Bài toán (1.1) với $0 < \alpha < 1$. Định nghĩa một hàm hành động, với mỗi $i \in E$, ta có

$$a(i) = \text{argmin}_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i,j) v^\alpha(j)\} \qquad (6)$$

Chính sách cố định được định nghĩa bằng hàm hành động a là chính sách tối ưu.

Cải thiện giá trị cho chi phí có chiết khấu

Thuật toán 1

- Bước 1. Cho $\alpha < 1$, chọn một giá trị dương đủ nhỏ ϵ, đặt $n = 0$, và đặt $v_0(i) = 0$ với mỗi $i \in E$. (Ta đặt $v_0 = 0$ để tiện tính toán, ta có thể chọn v_0 bất kỳ).
- Bước 2. Với mỗi $i \in E$, xác định $v_{n+1}(i)$ như sau

$$v_{n+1}(i) = \min_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i,j) v_n(j)\}.$$

- Bước 3. Tính δ

$$\delta = max\{|v_{n+1}(i) - v_n(i)|\}.$$

- Bước 4. Nếu $\delta < epsilon$, đặt $v^\alpha = v_{n+1}$ và dừng thuật toán; ngược lại, tăng n thêm 1 và quay lại Bước 2.

Cải tiến chính sách cho chi phí có chiết khấu

Thuật toán 2

- Bước 1. Cho $\alpha < 1$, đặt $n = 0$, định nghĩa hàm hành động a_0

$$a_0(i) = argmin_{k \in A} f_k(i)$$

với mọi $i \in A$.

- Bước 2. Xác định ma trận P và véc tơ f

$$f(i) = f_{a_n(i)}(i)$$
$$P(i,j) = P_{a_n(i)}(i,j)$$

với mỗi $i, j \in E$.

Cải tiến chính sách cho chi phí có chiết khấu

Thuật toán 2

- Bước 3. Tính véc tơ v

$$v = (I - \alpha P)^{-1} f.$$

- Bước 4. Xác định hàm hành động a_{n+1} như sau

$$a_{n+1}(i) = \text{argmin}_{k \in A}\{f_k(i) + \alpha \sum_{j \in E} P_k(i,j) v(j)\}$$

- Bước 5. Nếu $a_{n+1} = a_n$, đặt $v^\alpha = v, a^\alpha = a_n$ và dừng thuật toán, ngược lại tăng n lên 1 và quay lại Bước 2.

Quy hoạch tuyến tính cho chi phí có chiết khấu

Thuật toán 3

Nghiệm tối ưu của bài toán sau đây chính là nghiệm tối ưu v^α *của Bài toán (1) với* $0 < \alpha < 1$

$$\max \sum_{i \in E} u(i)$$

với điều kiện

$$u(i) \leq f_k(i) + \alpha \sum_{j \in E} P_k(i,j) u(j) \text{ với mọi } i \in E, k \in A.$$

Các thuật toán chi phí trung bình dài hạn

Tính chất 4

Giả sử rằng mọi chính sách cố định đều sinh ra mỗi xích Markov với tập không giảm. Khi đó, tồn tại một đại lượng vô hướng φ^* và một véc tơ h thỏa mãn, với mọi $i \in E$,

$$\varphi^* + h(i) = \min_{k \in A}\{f_k(i) + \sum_{j \in E} P_k(i,j)h(j)\}. \qquad (7)$$

Đại lượng vô hướng φ^* là giá trị tối ưu của Bài toán (1.2), và hàm hành động tối ưu là

$$a(i) = \operatorname{argmin}_{k \in A}\{f_k(i) + \sum_{j \in E} P_k(i,j)h(j)\}$$

Véc tơ h là duy nhất sai khác một hằng số.

Các thuật toán chi phí trung bình dài hạn

Từ Tính chất 4, ta có

$$\varphi^* + h(i) = f_{a(i)}(i) + \sum_{j \in E} P_{a(i)}(i,j)h(j) \tag{8}$$

$$\Leftrightarrow \varphi^* + h(i) = f_{a(i)}(i) + (P_{a(i)}h)(i). \tag{9}$$

Tính chất 5

Cho v^α là giá trị tối ưu của Bài toán (1),φ^* là giá trị tối ưu của Bài toán (2) giả sử mọi trạng thái cố định đều sinh ra xích Markov với một tập không giảm. Khi đó

$$\lim_{\alpha \to 1}(1-\alpha)v^\alpha(i) = \varphi^*$$

với mọi $i \in E$.

Các thuật toán chi phí trung bình dài hạn

Tính chất 6

Cho v^α là giá trị tối ưu của Bài toán (1), φ^* là giá trị tối ưu của Bài toán (2), h là véc tơ được xác định ở Tính chất (4). Khi đó

$$\lim_{\alpha \to 1}[v^\alpha(i) - v^\alpha(j)] = h(i) - h(j)$$

với mọi $i, j \in E$.

Cải tiến chính sách cho chi phí trung bình

Thuật toán 4

- Bước 1. Cho $n = 0$, ký hiệu trạng thái đầu tiên trong không gian trạng thái là 1, ta xác định hàm hành động a_0

$$a_0(i) = \text{argmin}_{k \in A} f_k(i)$$

với mỗi $i \in E$.

- Bước 2. Xác định ma trận P và véc tơ f như sau

$$f(i) = f_{a_n(i)}(i)$$
$$P(i,j) = P_{a_n(i)}(i,j)$$

với mỗi $i, j \in E$.

Cải tiến chính sách cho chi phí trung bình

Thuật toán 4

- Bước 3. Xác định giá trị φ và h bằng việc giải hệ

$$\varphi + h = f + Ph$$

trong đó $h(1) = 0$.

- Bước 4. Xác định hàm hành động a_{n+1} như sau

$$a_{n+1}(i) = \operatorname{argmin}_{k \in A}\{f_k(i) + \sum_{j \in E} P_k(i,j)h(j)\}$$

với mỗi $i \in E$.

Cải tiến chính sách cho chi phí trung bình

Thuật toán 4

- Bước 5. Nếu $a_{n+1} = a_n$, đặt $\varphi^* = \varphi, a^* = a_n$, và dừng thuật toán, ngược lại tăng n thêm một và quay lại Bước 2.

Quy hoạch tuyến tính cho chi phí trung bình

Thuật toán 5

Giá trị tối ưu của bài toán quy hoạch tuyến tính dưới đây là giá trị tối ưu của bài toán (2)

$$\min \varphi = \sum_{i \in E} \sum_{k \in A} x(i,k) f_k(i)$$

với điều kiện

$$\sum_{k \in A} x(j,k) = \sum_{i \in E} \sum_{k \in A} x(i,k) P_k(i,j) \text{ với mọi } j \in E$$

$$\sum_{i \in E} \sum_{k \in A} x(i,k) = 1$$

$$x(i,k) \geq 0 \text{ với mọi } i \in E \text{ và } k \in A$$

Chính sách tối ưu là chọn hành động k cho trạng thái i sao cho $x(i,k) > 0$

Ứng dụng

- Giải một loạt bài toán tối ưu hóa thông qua quy hoạch động và học tăng cường.
- Áp dụng trong rất nhiều các lĩnh vực khác nhau, bao gồm robot, điều khiển tự động,kinh tế, và chế tạo
- Quá trình quyết định Markov thời gian liên tục được ứng dụng trong các hệ thống xếp hàng, các quá trình dịch bệnh và các quá trình dân số.

Tài liệu tham khảo

1. Richard M. Feldman - Ciriac Valdez-Flores, *Applied Probability and Stochastic Processes, Second Edition* , Springer-Verlag Berlin Heidelberg, 2010

Hết

Cảm ơn cô và các bạn đã theo dõi