Zachodniopomorski
Uniwersytet Technologiczny
w Szczecinie

Wydział
Informatyki

PROCESY DECYZYJNE MARKOWA

Karol Działowski

nr albumu: 39259
przedmiot: Uczenie ze wzmocnieniem

Szczecin, 7 grudnia 2020

Spis treści

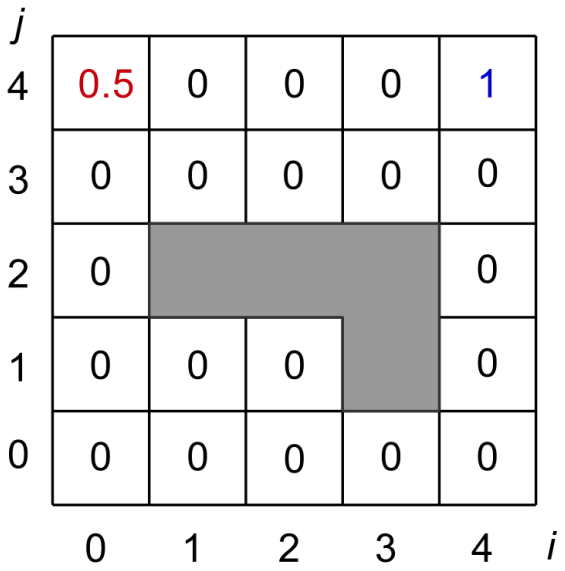
1 Cel laboratorium

Celem laboratorium nr 1 było zapoznanie się z procesami decyzyjnymi Markowa, a konkretnie wyznaczanie funkcji wartości $V(x)$, wyznaczanie funkcji wartości akcji $Q(x, a)$ oraz wyznaczanie strategi zachłannych dla funkcji wartości akcji wybranych strategii.

2 Zadanie 2

W zadaniu 2 dany jest proces deterministyczny, gdzie występują:

stany $X = \{(i,j)\}, \quad i,j = 0..4$,
akcje $A = \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$,
a stany $(0,4)$ i $(4,4)$ są absorbujące.



Rysunek 1: Przedstawienie problemu zadanie 2

Problem przedstawia obrazek 1.
Funkcja wzmocnień przedstawiona za pomocą poniższego wzoru:

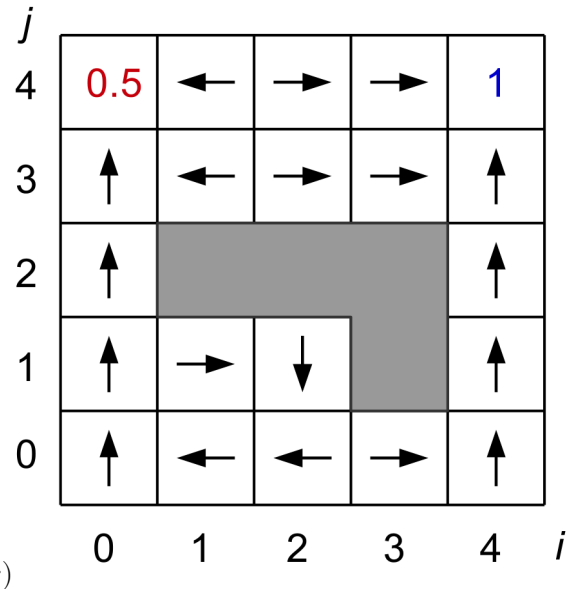
$$\rho((i,j),a) = \begin{cases} 1 \text{ gdy } x = (3,4) \text{ i } a = \rightarrow \\ 1 \text{ gdy } x = (4,3) \text{ i } a = \uparrow \\ 0.5 \text{ gdy } x = (0,3) \text{ i } a = \uparrow \\ 0.5 \text{ gdy } x = (1,4) \text{ i } a = \leftarrow \\ 0 \text{ w innych przypadkach.} \end{cases} \tag{1}$$

Funkcje przejścia dla pozostałych stanów:

$$\begin{aligned} \delta((i,j),\uparrow) &= (i, \text{ if dozwolone } (j+1): \quad j+1 \text{ else } : j) \\ \delta((i,j),\downarrow) &= (i, \text{ if dozwolone } (j-1): \quad j-1 \text{ else } : j) \\ \delta((i,j),\leftarrow) &= (\text{ if dozwolone } (i-1): \quad i-1 \text{ else } : i,j) \\ \delta((i,j),\rightarrow) &= (\text{ if dozwolone } (i+1): \quad i+1 \text{ else } : i,j) \end{aligned} \tag{2}$$

Dana jest strategia $\pi_1(x)$ zdefiniowana na rysunku (2). Celem zadania było

1. Wyznaczenie funkcji wartości dla strategii $\pi_1(x)$
2. Czy istnieją od niej strategie lepsze?
3. Jeśli tak, zaproponuj przykładową.



Rysunek 2: Strategia $\pi_1(x)$

2.1 Funkcje wartości

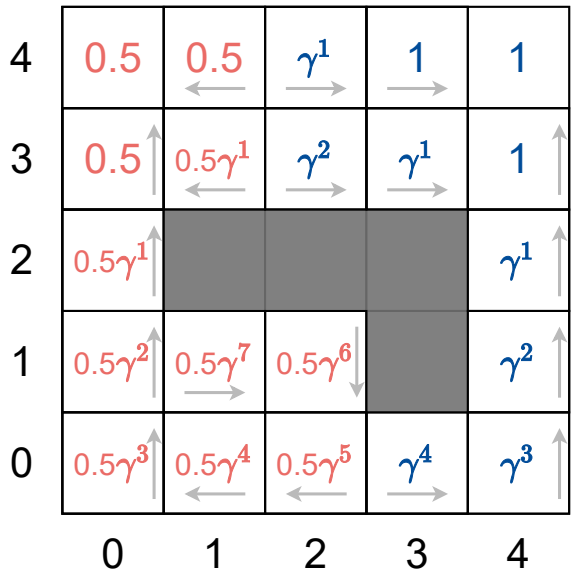
Dla procesu decyzyjnego Markowa, funkcja wartości ze względu na strategię π jest dla każdego stanu określona następująco:

$$V^{\pi}(x) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid x_0 = x\right] \tag{3}$$

Przykład wyznaczenia funkcji wartości dla stanu $x = (3, 0)$:

$$V^{\pi_1}(x) = 0 + 0 \cdot \gamma + 0 \cdot \gamma^2 + 0 \cdot \gamma^3 + 1 \cdot \gamma^4 = \gamma^4. \tag{4}$$

Wyznaczono funkcje wartości przedstawiono na rysunku (3).



Rysunek 3: Funkcja wartości dla strategii $\pi_1(x)$

2.2 Propozycja lepszej strategii

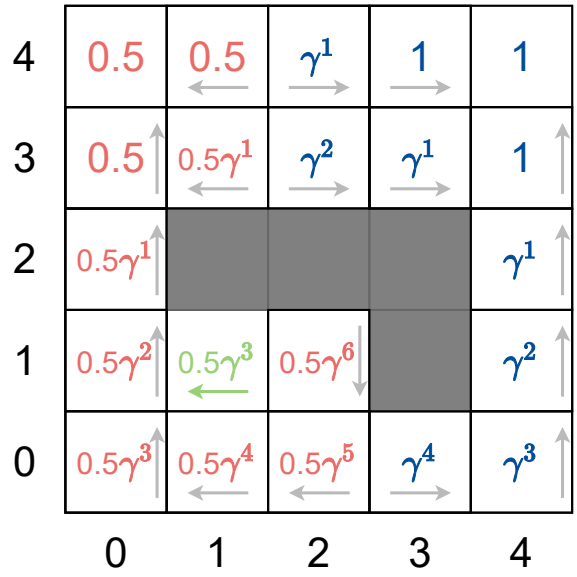
Zaproponowano zmianę akcji w stanie $(1, 1)$ z $\rightarrow$ na $\leftarrow$ tworząc strategię π_2.

Dla tego stanu nastąpi zmiana funkcji wartości. W reszcie stanów funkcje wartości akcji będą identyczne, bo zmiana tego stanu nie wpływa na inne.

Stan	Funkcja wartości π_1	Funkcja wartości π_2
(1, 1)	$0.5\gamma^7$	$0.5\gamma^3$

Tabela 1: Zmiana funkcji wartości

Zaproponowaną strategię przedstawiono na poniższym rysunku.



Rysunek 4: Zaproponowana strategia $\pi_2(x)$.
Na zielono zaznaczono stany i funkcje wartości które ulegają zmianie.

3 Zadanie 3

Dla danego procesu deterministycznego (rysunek 5) dane są:

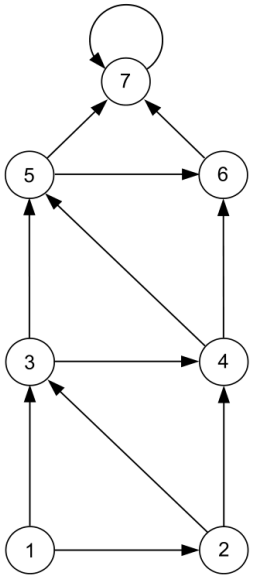
- Stany: $X = \{1, 2, 3, 4, 5, 6, 7\}$
- Akcje: $A = \{1, 2\}$
- Stan 7 jest absorbujący

Funkcja przejścia:

$$\delta(x, a) = \begin{cases} x + 1 & \text{dla } (x = 1, \dots, 6 \text{ i } a = 1) \\ & \text{lub } (x = 6 \text{ i } a = 2) \\ x + 2 & \text{dla } x = 1, \dots, 5 \text{ i } a = 2 \\ x & \text{dla } x = 7 \text{ i } (a = 1 \text{ lub } a = 2) \end{cases} \tag{5}$$

Funkcja wzmocnień:

$$\rho(x,a) = \begin{cases} -\frac{(a+1)^2}{4} & \text{dla } x \neq 7 \\ 0 & \text{dla } x = 7 \end{cases} \tag{6}$$



Rysunek 5: Przedstawienie modelu dla zadania 3

Zadanie miało na celu:

- Wyznaczenie funkcji wartości $V(x)$ i funkcję wartości akcji $A(x,a)$ dla:
  - strategii $\pi_1(x) = 1$
  - strategii $\pi_2(x) = 2$
- Określ która strategia jest lepsza
- Zbadaj wpływ współczynnika dyskontowania γ na jakość obu strategii. Porównaj ich jakość dla $\gamma = 1$, $\gamma = 0.9$, $\gamma = 0.7$.
- Wyznacz strategię zachłanną dla funkcji wartości akcji określonej dla strategii $\pi_2(x)$ przy współczynniku dyskontowania $\gamma = 0.9$
- Sprawdź czy znaleziona strategia zachłanna jest lepsza od strategii $\pi_2(x)$

3.1 Funkcja wartości i funkcja wartości akcji

Dla procesu decyzyjnego Markowa, funkcja wartości i funkcja wartości akcji ze względu na strategię π jest dla każdego stanu określona następująco:

$$V^{\pi}(x) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid x_0 = x\right], \quad Q^{\pi}(x,a) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid x_0 = x, a_0 = a\right] \tag{7}$$

3.1.1 Strategia 1

Funkcja wartości:

$$
\begin{aligned}
V^{\pi_1}(1) &= -\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 - \gamma^4 - \gamma^5 \\
V^{\pi_1}(2) &= -\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 - \gamma^4 \\
V^{\pi_1}(3) &= -\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 \\
V^{\pi_1}(4) &= -\gamma^0 - \gamma^1 - \gamma^2 \\
V^{\pi_1}(5) &= -\gamma^0 - \gamma^1 \\
V^{\pi_1}(6) &= -\gamma^0 \\
V^{\pi_1}(7) &= 0
\end{aligned}
$$

Funkcje wartości akcji:

$$
\begin{aligned}
Q^{\pi_1}(1,1) &= -\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 - \gamma^4 - \gamma^5 & Q^{\pi_1}(1,2) &= -\frac{9}{4}\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 - \gamma^4 \\
Q^{\pi_1}(2,1) &= -\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 - \gamma^4 & Q^{\pi_1}(2,2) &= -\frac{9}{4}\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 \\
Q^{\pi_1}(3,1) &= -\gamma^0 - \gamma^1 - \gamma^2 - \gamma^3 & Q^{\pi_1}(3,2) &= -\frac{9}{4}\gamma^0 - \gamma^1 - \gamma^2 \\
Q^{\pi_1}(4,1) &= -\gamma^0 - \gamma^1 - \gamma^2 & Q^{\pi_1}(4,2) &= -\frac{9}{4}\gamma^0 - \gamma^1 \\
Q^{\pi_1}(5,1) &= -\gamma^0 - \gamma^1 & Q^{\pi_1}(5,2) &= -\frac{9}{4}\gamma^0 \\
Q^{\pi_1}(6,1) &= -\gamma^0 & Q^{\pi_1}(6,2) &= -\frac{9}{4}\gamma^0 \\
Q^{\pi_1}(7,1) &= 0 & Q^{\pi_1}(7,2) &= 0
\end{aligned}
$$

3.1.2 Strategia 2

Funkcja wartości:

$$
\begin{aligned}
V^{\pi_2}(1) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 - \frac{9}{4}\gamma^2 \\
V^{\pi_2}(2) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 - \frac{9}{4}\gamma^2 \\
V^{\pi_2}(3) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 \\
V^{\pi_2}(4) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 \\
V^{\pi_2}(5) &= -\frac{9}{4}\gamma^0 \\
V^{\pi_2}(6) &= -\frac{9}{4}\gamma^0 \\
V^{\pi_2}(7) &= 0
\end{aligned}
$$

Funkcje wartości akcji:

$$
\begin{aligned}
Q^{\pi_2}(1,1) &= -\gamma^0 - \frac{9}{4}\gamma^1 - -\frac{9}{4}\gamma^2 - -\frac{9}{4}\gamma^3 \\
Q^{\pi_2}(2,1) &= -\gamma^0 - \frac{9}{4}\gamma^1 - \frac{9}{4}\gamma^2 \\
Q^{\pi_2}(3,1) &= -\gamma^0 - \frac{9}{4}\gamma^1 - \frac{9}{4}\gamma^2 \\
Q^{\pi_2}(4,1) &= -\gamma^0 - \frac{9}{4}\gamma^1 \\
Q^{\pi_2}(5,1) &= -\gamma^0 - \frac{9}{4}\gamma^1 \\
Q^{\pi_2}(6,1) &= -\gamma^0 \\
Q^{\pi_2}(7,1) &= 0
\end{aligned}
\qquad
\begin{aligned}
Q^{\pi_2}(1,2) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 - \frac{9}{4}\gamma^2 \\
Q^{\pi_2}(2,2) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 - \frac{9}{4}\gamma^2 \\
Q^{\pi_2}(3,2) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 \\
Q^{\pi_2}(4,2) &= -\frac{9}{4}\gamma^0 - \frac{9}{4}\gamma^1 \\
Q^{\pi_2}(5,2) &= -\frac{9}{4}\gamma^0 \\
Q^{\pi_2}(6,2) &= -\frac{9}{4}\gamma^0 \\
Q^{\pi_2}(7,2) &= 0
\end{aligned}
$$

3.2 Porównanie strategii – wpływ współczynników dyskontowania

Aby porównać strategie należy posiadać wartość γ. Sprawdzono jakość strategii dla trzech wartości $\gamma = 1$, $\gamma = 0.9$, $\gamma = 0.7$.

	$\gamma = 1$		$\gamma = 0.9$		$\gamma = 0.7$	
x	$V^{\pi_1}(x)$	$V^{\pi_2}(x)$	$V^{\pi_1}(x)$	$V^{\pi_2}(x)$	$V^{\pi_1}(x)$	$V^{\pi_2}(x)$
1	-6	-6.75	-4.68559	-6.0975	-2.9411	-4.9275
2	-5	-6.75	-4.0951	-6.0975	-2.7731	-4.9275
3	-4	-4.5	-3.439	-4.275	-2.533	-3.825
4	-3	-4.5	-2.71	-4.275	-2.19	-3.825
5	-2	-2.25	-1.9	-2.25	-1.7	-2.25
6	-1	-2.25	-1	-2.25	-1	-2.25
7	0	0	0	0	0	0

Tabela 2: Porównanie strategii dla różnych współczynników dyskontowania γ

Na podstawie zebranych danych można stwierdzić, że dla współczynników dyskontowania $\gamma \in \{1, 0.9, 0.7\}$ strategia π_1 jest lepsza od strategii π_2, ponieważ wartości funkcji wartości są większe lub równe w dla każdego stanu x.

3.3 Strategia zachłanna

Dla współczynnika dyskontowania $\gamma - 0.9$ wyznaczamy funkcje wartości akcji akcji dla strategii $\pi_2(x)$. Lepszą strategię w danej akcji zaznaczono podkreśleniem.

x	$Q^{x_2}(x, 1)$	$Q^{x_2}(x, 2)$
1	-6.48775	<u>-6.0975</u>
2	<u>-4.8475</u>	-6.0975
3	-4.8475	<u>-4.275</u>
4	<u>-3.025</u>	-4.275
5	-3.025	<u>-2.25</u>
6	<u>-1</u>	-2.25
7	0	0

Tabela 3: Wyznaczenie strategii zachłannej

Wyznaczona strategia zachłanna opisywana jest sekwencją akcji: 2, 1, 2, 1, 2, 1, *. Przykładowo:

$$\pi_{zach}(x) = \begin{cases} 1 & \text{dla } x(\text{mod}2) = 0 \\ 2 & \text{dla } x(\text{mod}2) = 1 \end{cases} \tag{8}$$

Porównano wyznaczoną strategię zachłanną z strategią $\pi_2(x)$.

x	$V^{\pi_{\text{zach}}}(x)$	$V^{\pi_2}(x)$
1	-6.0975	-6.0975
2	-4.8475	-6.0975
3	-4.275	-4.275
4	-3.025	-4.275
5	-2.25	-2.25
6	-1	-2.25
7	0	0

Tabela 4: Porównanie strategii π_{zach} i $\pi_2(x)$ dla współczynnika dyskontowania $\gamma = 0.9$

Strategia zachłanna π_{zach} wyznaczona nad strategią π_2 jest strategią lepszą.