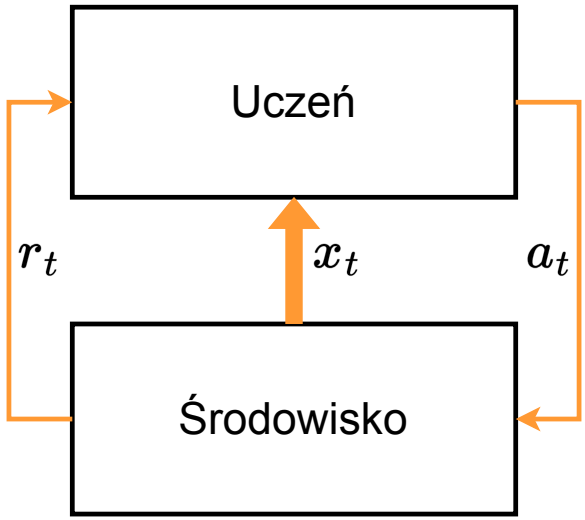
Uczenie maszynowe:
wykład 13

Paweł Cichosz

Scenariusz

W kroku t:

1. obserwuj aktualny stan x_t;
2. wybierz akcję a_t dla stanu x_t;
3. wykonaj akcję a_t;
4. obserwuj nagrodę r_t i następny stan x_{t+1};
5. *ucz-się*(x_t, a_t, r_t, x_{t+1}).



Zadanie ucznia

- Długookresowa maksymalizacja nagród, zazwyczaj:

$$\mathbf{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$$

  gdzie $0 < \gamma \leq 1$ – współczynnik dyskontowania.
- Wymagane uwzględnienie opóźnionych efektów wykonanych akcji.

Zadania epizodyczne

- Seria prób (epizodów) o skończonej liczbie kroków, w każdej z których jakość działania jest oceniana niezależnie.

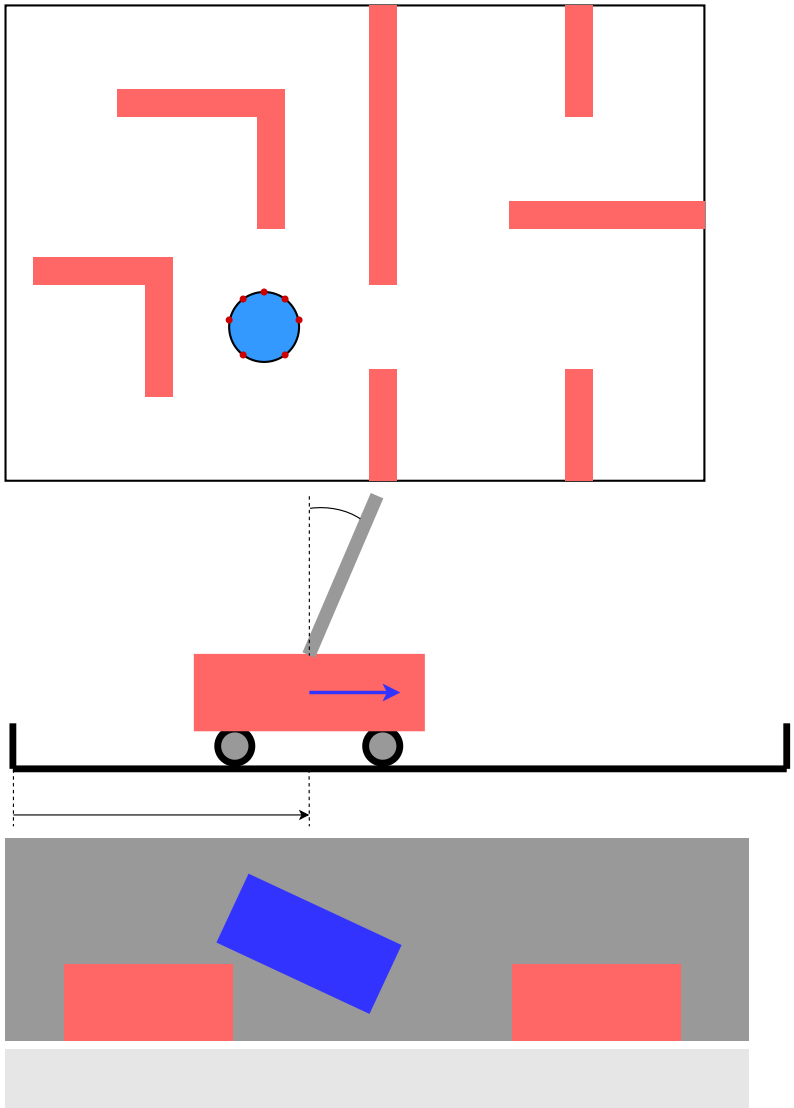
Do-sukcesu: próba kończy się po osiągnięciu sukcesu, należy dążyć do niego jak najszybciej.

- Wartość nagrody na końcu próby dodatnia (lub 0), wartość nagrody we wcześniejszych krokach 0 (lub ujemna).

Do-porażki: próba kończy się po odniesieniu porażki, należy ją opóźniać jak najdłużej.

- Wartość nagrody na końcu próby ujemna (lub 0), wartość nagrody we wcześniejszych krokach 0 (lub dodatnia).

Przykłady zadań praktycznych



- Mobilny robot.
- Odwrócone wahadło.
- Parkowanie samochodu.
- Sterowanie systemem wind.
- Gra w planszową grę dwuosobową.

Proces decyzyjny Markowa

$\langle X, A, \delta, \varrho \rangle$

- X – skończony zbiór stanów,
- A – skończony zbiór akcji,
- δ – funkcja przejść,
- ϱ – funkcja nagród.

Funkcja przejść: $\delta(x, a)$ – zmienna losowa oznaczająca stan osiągany po wykonaniu akcji a w stanie x:

$$P_{xy}(a) = P(\delta(x, a) = y)$$

Funkcja nagród: $\varrho(x, a)$ – zmienna losowa oznaczająca nagrodę po wykonaniu akcji a w stanie x:

$$R(x, a) = \mathbf{E}[\varrho(x, a)]$$

Własność Markowa: nagrody i przejścia zależą tylko od aktualnego stanu i akcji, a nie od historii.

Strategie i funkcje wartości

Strategia: $\pi : X \to A$ (stacjonarna, deterministyczna).

Funkcja wartości ze względu na strategię π:

$$V^{\pi}(x) = \mathbf{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \;\middle|\; x_0 = x\right]$$

gdzie $\mathbf{E}_{\pi}$ – wartość oczekiwana pod warunkiem posługiwania się strategią π.

Funkcja wartości akcji ze względu na strategię π:

$$Q^{\pi}(x,a) = \mathbf{E}_{\pi}\left[\varrho(x,a) + \sum_{t=1}^{\infty} \gamma^t r_t \;\middle|\; x_0 = x, a_0 = a\right]$$

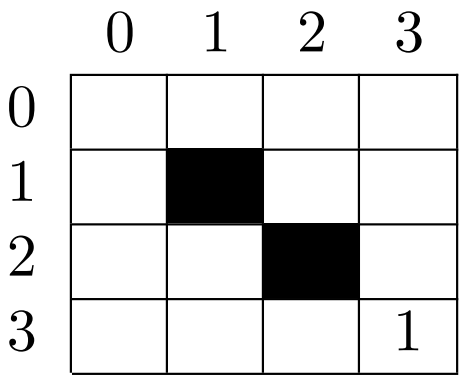
Optymalność

- Strategia π_1 jest lepsza niż strategia π_2 jeśli:

$$(\forall x \in X)\ V^{\pi_1}(x) \geq V^{\pi_2}(x)$$
$$(\exists x \in X)\ V^{\pi_1}(x) > V^{\pi_2}(x)$$

- Strategia jest optymalna jeśli nie istnieje strategia od niej lepsza.
- Istnieje co najmniej jedna strategia optymalna (może być więcej).
- Wszystkie strategie optymalne mają tę samą optymalną funkcję wartości V^* i optymalną funkcję wartości akcji Q^*.

Przykład: siatka 4×4



- Każda komórka oznacza stan.
- Możliwe jest wykonywanie akcji $\leftarrow$, $\rightarrow$, $\uparrow$ i $\downarrow$, z których każda powoduje:
  - z prawdopodobieństwem p przejście do sąsiedniej komórki w wybranym kierunku, chyba że ruch jest niemożliwy z powodu blokady lub osiągnięcia granicy środowiska,
  - z prawdopodobieństwem $1 - p$ pozostanie w dotychczasowej komórce.
- Nagroda 0 niezależnie od akcji w każdej komórce z wyjątkiem komórki w prawym dolnym rogu, której nie można opuścić (wykonanie dowolnej akcji nie powoduje przejścia do innego stanu) i w której zawsze otrzymuje się nagrodę 1.
- Dla dowolnej strategii π możemy wyznaczyć V^π lub Q^π wprost z definicji, jeśli efekty akcji są deterministyczne ($p = 1$).
- Możemy również wyznaczyć V^* lub Q^* (łatwo zauważyć, jaka strategia jest optymalna).

Równania Bellmana

- Rekurencyjne zależności dla funkcji wartości i wartości akcji:

  dla V^π:

$$V^\pi(x) = R(x, \pi(x)) + \gamma \sum_y P_{xy}(\pi(x)) V^\pi(y)$$

  dla Q^π:

$$Q^\pi(x, a) = R(x, a) + \gamma \sum_y P_{xy}(a) Q^\pi(y, \pi(y))$$

  dla V^*:

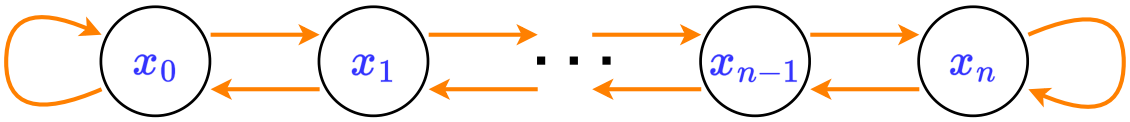
$$V^*(x) = \max_a \left[R(x, a) + \gamma \sum_y P_{xy}(a) V^*(y) \right]$$

  dla Q^*:

$$Q^*(x, a) = R(x, a) + \gamma \sum_y P_{xy}(a) \max_{a'} Q^*(y, a')$$

- Umożliwiają wyznaczenie odpowiednich funkcji (np. przez rozwiązanie układu równań).

Przykład: „$x_0 \to \cdots \to x_n$”



Stany: $\{x_0, x_1, \ldots, x_n\}$.
Akcje: $\{\leftarrow, \rightarrow\}$.
Nagrody: 1 w stanie x_n,
0 w pozostałych.

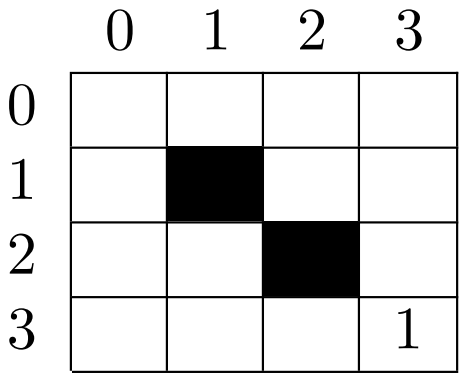
Przejścia: z pr. p przejście do kolejnego stanu w kierunku określonym przez akcję (z wyjątkiem skrajnych), z pr. $1-p$ pozostanie w aktualnym stanie.

- Oczywiście strategia π, która w każdym stanie rekomenduje akcję $\rightarrow$, jest optymalna dla $\gamma > 0$.
- Używając równania Bellmana można wyznaczyć wartość każdego stanu dla tej strategii:

$$V^\pi(x_n) = 1 + \gamma V^\pi(x_n)$$
$$V^\pi(x_{n-1}) = \gamma(pV^\pi(x_n) + (1-p)V^\pi(x_{n-1}))$$
$$\ldots$$

- Co zmieni się, jeśli z prawdopodobieństwem $1-p$ akcja powoduje przejście w kierunku przeciwnym do wybranego?

Przykład: siatka 4×4



- Dla dowolnej strategii π można wyznaczyć V^π lub Q^π używając odpowiedniego równania Bellmana, np. dla strategii „zawsze w prawo”:

$$V^\pi(32) = 0 + \gamma(pV^\pi(33) + (1-p)V^\pi(32))$$
$$V^\pi(23) = 0 + \gamma V^\pi(23)$$
$$V^\pi(33) = 1 + \gamma V^\pi(33)$$
$$\dots$$

- Łatwo zauważyć, jaka strategia jest optymalna.
- Co zmieni się, jeśli z prawdopodobieństwem $1-p$ akcja powoduje przejście w kierunku przeciwnym do wybranego?