Sztuczna inteligencja. O szukaniu szczęścia w niepewnym świecie (MDP)

Paweł Rychlikowski

Instytut Informatyki UWr

14 maja 2018

Procesy decyzyjne Markowa (MDP)

- Coś pomiędzy grami a zwykłym zadaniem przeszukiwania.
- a jednocześnie krok w stronę uczenia ze wzmocnieniem

... o **szukaniu** szczęścia w **niepewnym** świecie ...

MDP a przeszukiwanie

Standardowe przeszukiwanie

Znamy mechanikę świata i wiemy, że akcja w stanie da nam konretny rezultat (inny stan).

MDP

Znamy mechanikę świata i wiemy, że akcja w stanie da nam pewien rozkład prawdopodobieństwa na następnych stanach.

Nie wiemy co dokładnie się stanie, ale wiemy co **może** się stać i z jakim prawdopodobieństwem.

Własność Markowa

- Przyszłość zależy od ostatniego stanu.
- Nie zależy od historii...
- Chyba, że jej fragment (o długości N) uznamy za część stanu.

Uwaga na wulkany (1)

- Dobrze omawia się MDP na prostych światach na prostokątnej kratce.
- I od takich modeli zaczniemy.

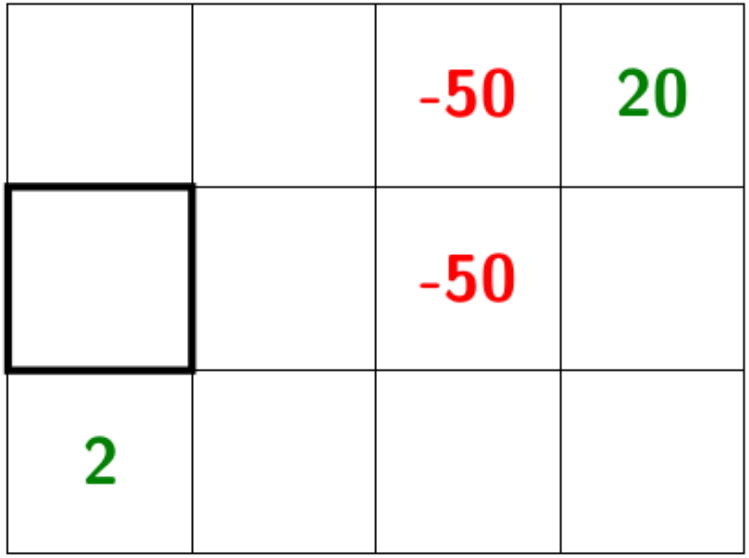
Generalnie myślimy na początku o przestrzeni stanów na tyle małej, że nie będzie kłopotów z pamiętaniem różnych wartości dla **każdego stanu**.

Uwaga na wulkany (2)

Volcano crossing

		-50	20
		-50	
2			

Mechanika świata wulkanów



- Możliwe 4 akcje (UDLR)
- W normalnym przypadku efekt oczywisty (próba wyjścia poza planszę oznacza pozostanie na polu)
- Z prawdopodobieństwem p możemy się poślizgnąć, wówczas poruszamy się w losowym kierunku.
- Dojście do pola z liczbą kończy grę (i odpowiednią dostajemy wypłatę).

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.
- rezygnacja oznacza wypłatę **10$**
- pozostanie to wypłata **4$** po której rzucamy kostką.
- Wynik:
  - 1,2 – koniec gry
  - 3,4,5,6 – gramy dalej

Pytanie

Ile mamy stanów? Odpowiedź: 2

Dla gry w kości

- 1 stan z decyzją, dwie polityki (schemat na tablicy).
- Możemy policzyć oczekiwaną wartość dla każdej:
  - rezygnacja – 10
  - pozostanie – (na tablicy)

MDP – formalna definicja

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów
2. Stan startowy, $s_{\text{start}} \in S$
3. Actions(s) – zbiór możliwych akcji w stanie s
4. T(s,a,s') – prawdopodobieństwo przejścia z s do s' w wyniku akcji a
5. Reward(s,a,s') – nagroda (wypłata) związana z tym przejściem
6. IsEnd(s) – czy stan jest końcowy?
7. Discount factor, $0 < \gamma \leq 1$ – sprawia, że nagrody w przyszłości cieszą mniej.

MDP – komentarz do definicji

- Można też myśleć, że dla pary (s, a) mamy rozkład prawdopodobieństw po parach (nowy-stan, nagroda).
- Nagroda może być pozytywna bądź negatywna

Uwaga

Oczywiście MDP jest ogólniejsze niż zadanie przeszukiwania (bo wystarczy przypisać niektórym rezultatom p-stwo 1, reszcie 0 i mamy zwykłe zadanie przeszukiwania)

Czym jest rozwiązanie MDP?

- Przypominamy: rozwiązaniem zadania przeszukiwania jest ciąg akcji (ale to nie tu nie działa, bo?)
  - (wyniki akcji są niedeterministyczne, więc nie wystarczy podać jednego ciągu akcji)
- Rozwiązanie: agent musi wiedzieć, co zrobić w każdym stanie.

Polityka

Definicja 1

Polityką deterministyczną nazwiemy funkcję, która każdemu stanowi przypisuje akcję (możliwą w tym stanie).

Definicja 2

Polityką nazwiemy funkcję, która każdemu stanowi przypisuje rozkład prawdopodobieństwa na akcjach (możliwych w tym stanie).

Wartościowanie polityki

- Gdy używamy polityki, otrzymujemy ciąg stanów, akcji i nagród
- Dla takiej ścieżki możemy zsumować nagrody, otrzymując użyteczność polityki dla tej ścieżki
- Wartością polityki jest oczekiwana użyteczność polityki (tzn. wartość oczekiwana zmiennej losowej wyrażającej użyteczność takiej ścieżki)

Discounting

- Realizując politykę, otrzymaliśmy ciąg stanów, nagród i akcji
  - $s_0, a_1, r_1, s_1, a_2, r_2, s_2 \ldots, s_n, a_{n+1}, r_{n+1}, s_{n+1} \ldots$
- Nagroda po uwzględnieniu zniżek:

$$r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \ldots$$

- Widzimy że:
  - Dla $\gamma = 1$ po prostu sumujemy nagrody cząstkowe
  - Dla $0 < \gamma < 1$ mamy możliwość mówienia o wartości nieskończonych ciągów akcji.

Uwaga o γ

- Zwróćmy uwagę, że discounting ma sens również w przypadku, gdy nagroda wypłacana jest jedynie w stanie końcowym.
- Jeżeli wypłata jest tylko w ostatnim stanie, to:
  a) Agent, który wygrywa ($R > 0$) woli dostać ją wcześniej,
  b) agent, który przegrywa ($R < 0$) woli dostać ją później.

Przyśpieszanie zwycięstwa i opóźnianie porażki jest „sensownym” zachowaniem.

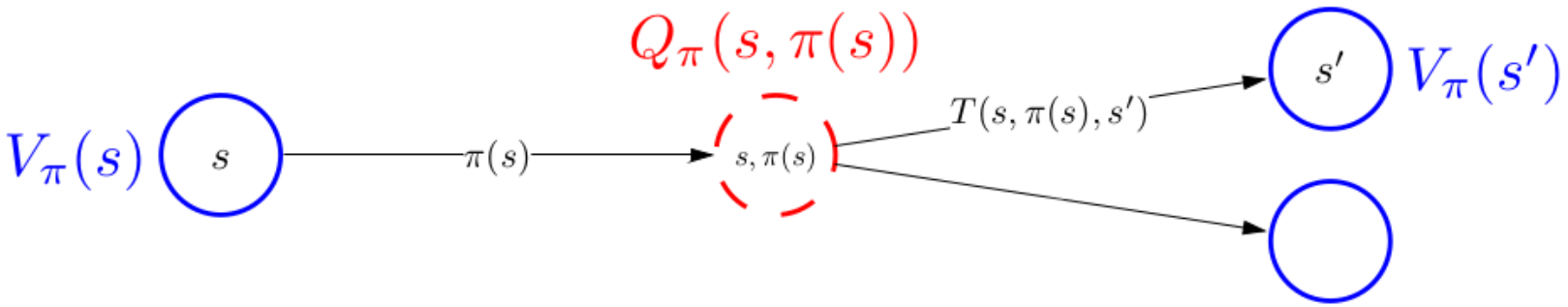
Wartość polityki

Definicja

Wartość $V_\pi(s)$ jest oczekiwaną użytecznością dla agenta startującego w stanie s i działającego zgodnie z polityką π

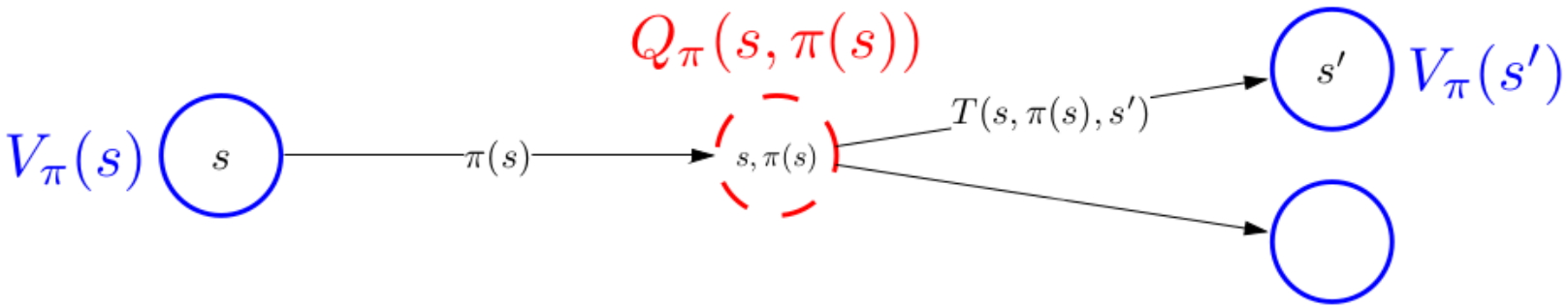
Definicja

Wartość $Q_\pi(s, a)$ jest oczekiwaną użytecznością dla agenta startującego w stanie s, wykonującego w tym stanie akcję a i **dalej** działającego zgodnie z polityką π



Źródło: CS221 / Autumn 2017 / Liang & Ermon

Zależności pomiędzy V i Q



$$Q_\pi(s, a) = \sum_{s'} T(s, a, s')(\text{Reward}(s, a, s') + \gamma V_\pi(s'))$$

Algorytm: policy evaluation

- Napiszmy rekurencyjny wzór dla wartości V (przy zadanej polityce)
- $$V_\pi(s) = \sum_{s'} T(s, \pi(s), s')(\text{Reward}(s, \pi(s), s') + \gamma V_\pi(s'))$$
- Mamy układ równań (liniowych), który można rozwiązywać standardowymi metodami.
- Równań jest tyle co stanów (czyli potencjalnie sporo)

Algorytm: policy evaluation (2)

Możemy ten wzór zmodyfikować, mówiąc: zamiast nieznanego V po prawej stronie weźmiemy poprzednie przybliżenie V:

$$V_\pi^{(t+1)}(s) = \sum_{s'} T(s, \pi(s), s')(\text{Reward}(s, \pi(s), s') + \gamma V_\pi^{(t)}(s'))$$

Algorytm: policy evaluation (3)

1. Zainicjuj $V_\pi^{(0)}(s) \leftarrow 0$, dla wszystkich s
2. Powtarzaj dla $t = 1, \ldots, t_{PE}$
   - Powtarzaj dla każdego stanu s

$$V_\pi^{(t+1)}(s) \leftarrow \sum_{s'} T(s, \pi(s), s')(\text{Reward}(s, \pi(s), s') + \gamma V_\pi^{(t)}(s')$$

Uwagi implementacyjne

- Kończymy, gdy dla każdego stanu zmiana mniejsza niż ε
- Oczywiście nie musimy pamiętać całej historii, tylko dwa ostatnie jej elementy (stany zmieniane i poprzednie)

Złożoność

$O(t_{PE}SS')$, gdzie S to liczba stanów, a S' (maksymalna) liczba stanów z niezerową $T(s, a, s')$.

Polityka optymalna

- Interesuje nas wyznaczanie polityki (a nie tylko ocenianie jej wartości.

Definicja

Optymalną wartością stanu $V_{opt}(s)$ jest maksymalna wartość stanu (ze względu na wszystkie polityki).

Rekurencja dla polityki optymalnej

Jaka polityka jest optymalna? Taka, która wybiera stany o optymalnej wartości

- Przypominamy, dla **każdej** polityki mamy:

$$Q_\pi(s,a) = \sum_{s'} T(s,a,s')(\text{Reward}(s,a,s') + \gamma V_\pi(s'))$$

- Dla polityki optymalnej: $V_{\text{opt}}(s) = \max_{a \in \text{Actions(s)}} Q_{\text{opt}}(s,a)$

Możemy podstawić do drugiego wzoru wzór na Q_π dla $\pi = \text{opt}$.

$$V_{\text{opt}}(s) = \max_{a \in \text{Actions(s)}} \sum_{s'} T(s,a,s')(\text{Reward}(s,a,s') + \gamma V_{\text{opt}}(s'))$$

Algorytm Iteracji wartości – Bellman, 1957

Polityka optymalna (do poprzedniego slajdu)

$\pi_{\text{opt}}(s) = \arg\max_{a \in \text{Actions(s)}} Q_{\text{opt}}(s, a)$

Nasz wzorek zmieniony na wersję do iterowania

$$V_{\text{opt}}^{(t+1)}(s) = \max_{a \in \text{Actions(s)}} \sum_{s'} T(s, a, s')(\text{Reward}(s, a, s') + \gamma V_{\text{opt}}^{(t)}(s'))$$

Algorytm Bellmana (value iteration)

- Mamy dodatkową pętlę wybierającą optymalną akcję (zamiast akcji danej przez politykę)
- Reszta bez zmian, tak jak w policy evaluation.

Warunki zbieżności

Algorytm jest zbieżny, jeżeli zachodzi któryś z warunków

- $\gamma < 1$
- Graf MDP jest acykliczny

Uwaga

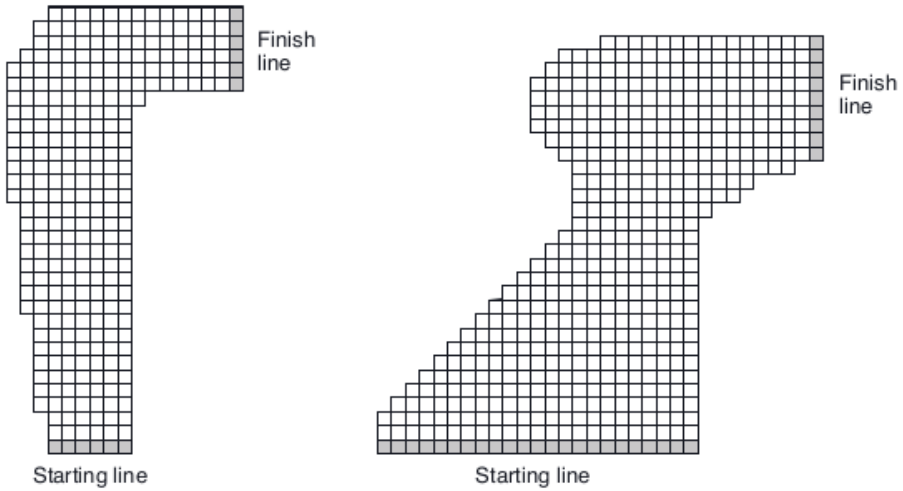
W tym ostatnim przypadku wymagana jest jedna iteracja, w której stany przeglądane są w odwrotnym porządku topologicznym (wyjaśnienie na tablicy)

Podsumowanie algorytmów

- Wartościowanie polityki (policy evaluation): $(\text{MDP}, \pi) \to V_\pi$
- Iteracja wartości (value iteration): $(\text{MDP}) \to V_{\text{opt}}, \pi_{\text{opt}}$

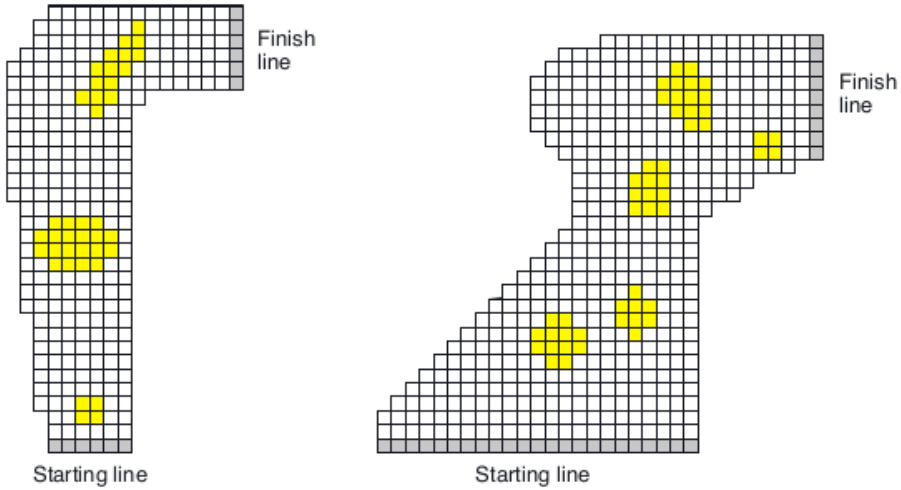
Będą jeszcze inne, w których założymy nieznajomość świata...

Przykład. Wyścigi samochodzików.



- Prędkość autka jest wektorem $(dx, dy) \in \{-3, -2, \ldots, 2, 3\} \times \{-3, -2, \ldots, 2, 3\}$
- Akcja: zmiana prędkości (każda składowa o co najwyżej 1)
- Celem jest (przejechać) przez metę (możemy to uprościć poszerzając metę i mówiąc, że celem jest dotarcie do piksela mety)
- W pełni deterministyczny świat (BFS, A^*?)

Przykład. Wyścigi samochodzików. (2)



- Dodajemy plamy po oleju (pomysł MZ)
- Ruch z pola oleju dodaje dodatkową składową losową do prędkości (znamy rozkład).

W tym momencie klasyczne MDP + algorytm Bellmana (czyli iteracji wartości) powinny dać dobry wynik.

Przykład. Wyścigi samochodzików. (3)

- Problem: dużo większa plansza, dużo większa liczba stanów.
- **Pomysł 1**: położenie „rozmyte”, na przykład w kwadracie 10×10 pikseli.
- **Pomysł 2**: dodatkowo informacja, czy jestem 1, 2, czy 3 raz w takim kwadracie ($3 < 100$)

Fundamentalny problem: nie znamy mechaniki takiego świata (i wielu innych)

Uczenie ze wzmocnieniem

- Zakładamy, że nie dysponujemy modelem (czyli przejściami, prawdpodobieństwami i nagrodami)
- Możemy wszakże wykonywać pewne eksperymenty w naszym systemie, w wyniku których zdobywamy wiedzę jak nam poszło

Uwaga

Zauważmy, że to pasuje do naszych samochodzików z rozmytymi stanami (eksperyment przeprowadzamy na prawdziwym modelu, ale obserwujemy model „rozmyty")

Ogólny schemat uczenia ze wzmocnieniem

Dla $t \in 1, 2, 3, \ldots$

- Wybieramy akcje $a_t = \pi_{act}(s_{t-1})$ (**jak?**)
- Wykonujemy akcję i obserwujemy nowy stan s_t
- Uaktualniamy parametry (**jak?**)

Metody Monte Carlo

- Estymujemy model podczas eksperymentów
- Rozwiązujemy wyszacowane MDP.

Szacowany MDP

- $\hat{T}(s, a, s') = \frac{\text{cnt}(s,a,s')}{\text{cnt}(s,a)}$
- Nagroda: średnie **r** dla zaobserwowanych **s a r s'**

Eksploracja i eksploatacja

- Jaką polityką mamy badać świat?
  a) Wybierającą akcje losowo (strata czasu?)
  b) Wybierającą akcje prowadzącą do stanu o najlepszej wartości V (ale możemy się zafiksować na nieoptymalnej ścieżce)
- Wybór między a) i b) to wybór między eksploracją i eksploatacją
  - Pamiętamy: strategia ε-zachłanna.

Możemy przeplatać etapy wyznaczania modelu i rozwiązywania MDP (bo w kolejnych iteracjach mamy możliwość wykorzystania lepszej strategii eksploatacyjnej).