Zaawansowane Metody Inteligencji Obliczeniowej

Lab 3: Procesy Decyzyjne Markowa (MDP) – wprowadzenie

Michał Kempka Marek Wydmuch

18 marca 2021

Fundusze Europejskie Polska Cyfrowa | Rzeczpospolita Polska | Unia Europejska Europejski Fundusz Rozwoju Regionalnego

"Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)",
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

1 Wprowadzenie

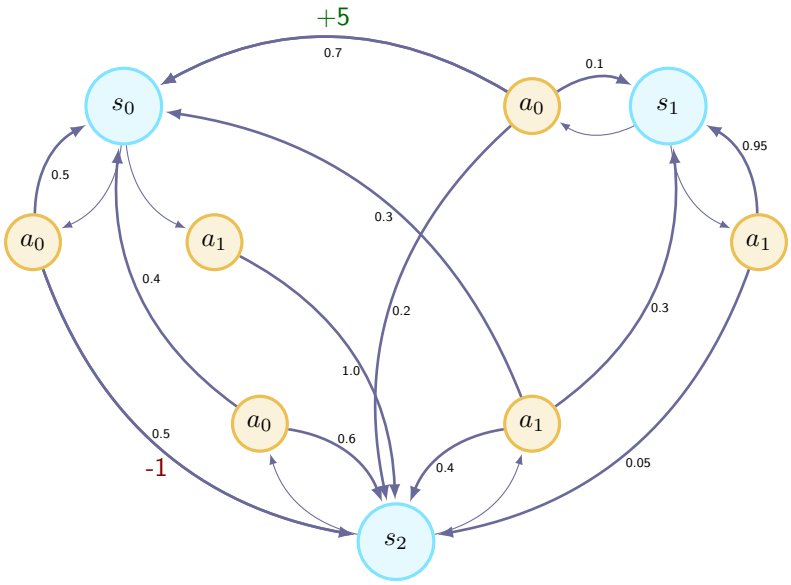
Uczenie ze wzmocnieniem (ang. reinforcement learning), jak wspomnieliśmy wcześniej, zajmuje się **agentami** działającymi w **środowiskach** wykonującymi pewne **akcje**. Akcje te wpływają na stan świata, a w ich konsekwencji agent otrzymuje wzmocnienie (ang. reinforcement). W przypadku kary mówimy o negatywnym wzmocnieniu, a w przypadku nagrody pozytywnego wzmocnienia. Podczas zajęć będziemy odnosić się do skwantyfikowanych wartości (zamiast np. rażenia prądów lub rozdawania cukierków) i będziemy nazywać wzmocnienie nagrodą (ang. **reward**) niezależnie od tego czy jest to wartość pozytywna czy negatywne. Bardzo prostym przykładem takiego zachowanie jest tresowanie psa. Gdy podaje łapę dostaje smakołyk i wzmacniane jest zachowanie podawania łapy. Nas jednak to nie satysfakcjonuje, chcielibyśmy żeby Puszek umiał przynieść piwo z lodówki, wyciągnąć tonącego człowieka z jeziora, albo wywęszył narkotyki, a następnie zaatakował ich właściciela. Są to jednak relatywnie skomplikowane **sekwencje** zależnych stanów i czynności, których trzeba się nauczyć. Właśnie takimi sekwencyjne zależności odróżniają uczenie ze wzmocnieniem od wielu innych problemów decyzyjnych (wieloręki bandyta, klasyfikacja), w których podjęte decyzje są od siebie niezależne.

2 Dyskretny Proces Decyzyjny Markowa

By zamodelować takie środowisko (problem, zadanie optymalizacyjne) możemy użyć Dyskretnego Procesu Decyzyjnego Markowa (ang. Discrete Markov Decision Process - **MDP**):

- Zdefiniowany jest przez krotkę: $MDP = \langle \mathcal{S}, \mathcal{A}, \mathcal{P} \rangle$,
- $\mathcal{S} = \{s_0, s_1 \dots s_{|\mathcal{S}|-1}\}$ - dyskretny zbiór stanów, najczęściej zakłada się, że jest to zbiór skończony (wtedy mówimy o ang. finite MDP). W każdym kwancie czasu **t** (czas też uznajemy za dyskretny) agent obserwuje jeden z dostępnych stanów.
- $\mathcal{A} = \{a_0, a_1 \dots a_{|\mathcal{A}|-1}\}$ - dyskretny, skończony zbiór akcji. W każdym kwancie czasu t agent musi wybrać jedną z akcji, lecz w ogólności nie wszystkie akcje muszę być dostępne w każdym stanie.
- $\mathcal{P}(s', r|s, a) = Pr\{s_{t+1} = s', r_t = r|s_t = s, a_t = a\}$ - model przejść (ang. transition model), czyli łączny rozkład prawdopodobieństwa otrzymywanych nagród i wystąpienia następnych stanów (następników)

Z uwagi na fakt, że istnieje wiele różnych definicji MDP, które są równoważne (tzn. mogłyby być przetransformowane do siebie wzajemnie) będziemy nadużywać tej notacji w zależności od potrzeb. Będzie to jednak zaznaczone i (oby) oczywiste z kontekstu. W szczególności wszelkie sumy lub średnie (np. formy $\sum_a$) będą domniemywać, że sumujemy po wszystkich dostępnych wartościach (w tym wypadku np. po wszystkich możliwych $a \in \mathcal{A}$).



Rysunek 1: Przykładowy diagram procesu decyzyjnego Markowa z niedeterministycznymi przejściami (prawdopodobieństwa przejść małym czarnym drukiem) i deterministycznymi nagrodami (zielone i czerwowne liczby).

2.1 Własność Markowa

Własność Markowa, jest to podejście do modelowania problemu, które mówi nam, że przyszłość i przeszłość są (warunkowo) nie zależne pod warunkiem teraźniejszości. Oznacza to, że wszystko co stanie się w przyszłości zależy jedynie od obecnego stanu i naszych akcji. Jak widać z definicji modelu przejść, MDP posiada własność Markowa.
Pytanie: Czy własność Markowa jest zwykle realistycznym założeniem? Gdy nie jest, z czego to wynika i jakie są tego konsekwencje i jak sobie z nimi poradzić?

2.2 Koniec świata

Powyższy model zakłada, że świat nie ma końca i agent (racjonalny) działa bez strachu i zawsze patrzy w przyszłość, potencjalnie zaniedbując najbliższe zachłanne akcje. Jest to często uzasadnione gdy taki 'koniec świata' nie ma praktycznego ani semantycznego sensu. Często jednak mamy do czynienia z przeciwną sytuacją, w której pojawia się potrzeba stworzenia stanów końcowych/terminalnych (ang. **terminal states**). Są to stany, w których agent kończy swoje działanie (po dotarciu do nich) - nie wykonuje zatem więcej akcji i nie otrzymuje żadnych nagród.

2.3 Nasz cel

Mamy już zdefiniowany świat. Do tej pory jednak, nie poruszyliśmy precyzyjnie co chcemy osiągnąć (co optymalizować). Oczywiście, chcemy żeby nagrody były wysokie i kary niskie. Jest to jednak zbyt mało precyzyjne sformułowanie by uzyskać jakieś praktyczne wyniki, które będą poparte jakąś dozą rygoru matematycznego.
Pytanie: Jaki jest zatem nasz cel? Tzn. co chcemy optymalizować biorąc pod uwagę nasz model MDP?

2.4 Koniec świata nie nadchodzi

Po chwili zastanowienia można dojść do wniosku, że akcje podejmowane teraz mają większy wpływ na najbliższą przyszłość (stany i nagrody). Dodatkowo możemy uznać, że nagrody zbierane w najbliższej przyszłości są ważniejsze (cenniejsze) niż te w dalekiej przyszłości. Prostym analogiem są pieniądze, które wartość przez inflację i mogą być zainwestowane. Możemy uznać, że nagroda z przyszłości(następnego kroku) jest warta tylko pewien ułamek naszej obecnej nagrody. Zwyczajowo ułamek ten oznaczamy jako **gamma** γ i nazywamy

go współczynnikiem dyskontowym (ang. **discount factor**).

Pytanie: Jak zatem uwzględnić w naszej maksymalizacji współczynnik dyskontowy?

Pytanie: Co jeśli współczynnik dyskontowy wynosi 0?

Pytanie: Co jeśli współczynnik dyskontowy wynosi 1?

Pytanie: Co jeśli współczynnik dyskontowy jest większy od 1?

Pytanie: Co jeśli współczynnik dyskontowy jest mniejszy od

Pytanie: Czy istnieje sytuacja gdzie środowisko nie ma końca, a współczynnik dyskontowy nie będzie konieczny?

2.5 Polityka (ang.policy)

Przypomnijmy, że polityką (**ang. policy**, nie ang. politics) nazywamy funkcję agenta, która mapuje stan (lub jego reprezentację) na akcję w przypadku polityki **deterministycznej**, lub rozkład prawdopodobieństwa na akcjach w przypadku polityki **niedeterministycznej**. Zwykle politykę oznaczamy π, a π_* używamy do oznaczenia polityki optymalnej (maksymalizującej nasz cel).

Pytanie: W jakiej sytuacji chcielibyśmy by polityka była niedeterministyczna?

2.6 Funkcja wartości stanu v

Mając daną politykę π oraz dany cel (w postaci oczekiwanej, sumarycznej, zdyskontowanej nagrody) chcielibyśmy ustalić jak dobry jest wybrany stan według naszej polityki - tzn. ile sumarycznie nagród (zdyskontowanych) zbierzemy zaczynając w tym stanie i wykonując dalej akcje według zadanej polityki. Właśnie taką funkcję nazywamy funkcją wartości stanu (ang. **state-value function**) i definiujemy ją jako:

$$v_\pi(s) = \mathbb{E}_\pi\Big[\sum_{i=0}^{\infty} \gamma^i r_{t+i} | s_t = s\Big] \tag{1}$$

2.7 Funkcja wartości akcji q

Analogicznie do funkcji wartości stanu możemy zdefiniować funkcję wartości akcji (ang. **action-value function**), $q_\pi(s, a)$, która właściwie odpowiada parze (stan,akcja). Mówi nam ona ile średnio zdobędziemy sumarycznie zdyskontowanych nagród wykonując akcję wdg polityki π zaczynając od stanu s i wykonania akcji a. Warto zauważyć, że pierwsza akcja a, jest niezależna od polityki.

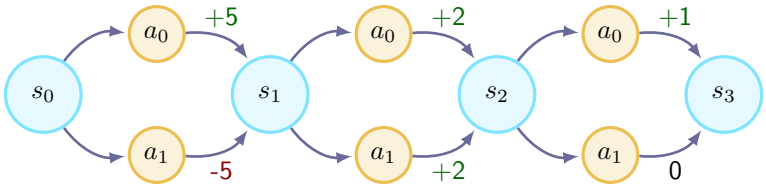
$$q_\pi(s, a) = \mathbb{E}_\pi\Big[\sum_{i=0}^{\infty} \gamma^i r_{t+i} | s_t = s, a_t = a\Big] \tag{2}$$

2.8 Liczenie v i q

Rozważ MDP z rysunku 2, policz v_π i q_π dla wszystkich stanów i akcji, przyjmij $\gamma = 1$ i politykę, która wybiera akcje z równym prawdopodobieństwem.

2.9 MDP z cyklem

Rozważ MDP przedstawione na rysunku 1. Jak policzyć v_π i q_π dla wszystkich stanów i akcji? Co jeśli chcielibyśmy wartości dla optymalnej polityki (więc właściwie maksymalne)?



Rysunek 2: Diagram przedstawiający proste MDP z czterema stanami (jeden początkowy, jeden terminalny). Nagrody i następniki są deterministyczne i zależą tylko od dwóch akcji.

2.10 Równanie Bellmana

Równanie uzyskane w sekcji 1 jest jednak mało użyteczne i możemy je trochę przetransformować.

$$v_\pi(s) = \mathbb{E}_\pi\Big[\sum_{i=0}^{\infty}\gamma^i r_{t+i}|s_t = s\Big] \tag{3}$$

$$= \mathbb{E}_\pi\Big[r_t + \gamma\sum_{i=0}^{\infty}\gamma^i r_{t+i+1}|s_t = s\Big] \tag{4}$$

$$= \sum_a \pi(a|s)\sum_{s',r}\mathcal{P}(s',r|s,a)\Big[r + \gamma\overbrace{\mathbb{E}_\pi\Big[\sum_{i=0}^{\infty}\gamma^i r_{t+i+1}|s_{t+1}=s'\Big]}^{v_\pi(s')}\Big] \tag{5}$$

$$= \sum_a \pi(a|s)\sum_{s',r}\mathcal{P}(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big] \tag{6}$$

Jest to **równanie Bellmana** dla funkcji wartości stanu i jest równaniem rekurencyjnym, którego rozwiązanie pozwoli nam wyznaczyć jakość polityki π. Nie jest to jeszcze rozwiązanie naszego problemu, lecz jesteśmy jeden krok bliżej.

Pytanie: Jak wyglądałoby to równanie gdyby polityka była deterministyczna, a rozkłady nagród i następnych stanów były niezależne od siebie?

2.11 Wyraź v_π jako funkcję q_π

2.12 Wyprowadź równanie Bellmana dla funkcji wartości akcji q_π

2.13 Optymalna polityka

Optymalną polityką π_* nazwiemy politykę, która maksymalizuje funkcję wartości stanu i akcji v_π i q_π (dla każdego stanu i akcji-stanu) dając nam optymalne funkcje v_* i q_*.

2.14 Jak wyglądałyby funkcje wartości stanu i akcji dla takiej polityki?

2.15 Różne zadania

1. Przetransformuj MDP w postaci, którą przedstawiliśmy do takiego gdzie rozkłady nagród i następników są niezależne. Weź pod uwagę współczynnik dyskontowy.
2. Wyobraźmy sobie giełdę walut(można myśleć o zwykłej giełdzie, ale chcemy uniknąć używania słowa akcja w dwóch kontekstach) gdzie możemy sprzedawać lup kupować jedną walutę. Załóżmy, że zawsze możemy kupić lub sprzedać dowolnie małą/dużą ilość waluty po aktualnej cenie (x_t jeśli tylko nas stać, a zaczynamy z danym budżetem (np. M_0)), którego nie będziemy zewnętrznie modyfikować. Jak zamodelujesz ten problem? W szczególności skup się na tym co i jak chcemy maksymalizować i jak ma wyglądać nagroda.
3. Wyobraźmy sobie środowisko gdzie każda akcja, karana jest niewielką sumą (relatywnie do pozostałych 'bardziej sensowny' nagród w tym środowisku) - tzw. ang. **living reward**. Jaki efekt ma to na naszą optymalizację?

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

"Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)",
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20