Zaawansowane Metody Inteligencji Obliczeniowej

Wykład 2: Uczenie ze wzmocnieniem i jednoręki bandyta

Michał Kempka Marek Wydmuch Bartosz Wieloch

7 marca 2022

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

Plan wykładu

Uczenie ze wzmocnieniem

Uczenie ze wzmocnieniem — będziemy stosowali skrót **RL** (*ang. Reinforcement Learning)*

Problem, którym zajmuje się uczenie ze wzmocnieniem:

Jak uczyć się podejmować **sekwencje decyzji**, które **maksymalizują całkowitą nagrodę**?

Uczenie ze wzmocnieniem

Uczenie ze wzmocnieniem — będziemy stosowali skrót **RL** (*ang. Reinforcement Learning)*

Problem, którym zajmuje się uczenie ze wzmocnieniem:

Jak uczyć się podejmować **sekwencje decyzji**, które **maksymalizują całkowitą nagrodę**?

1. agent **musi odkryć** które akcje są najbardziej korzystne (musi je sprawdzić!)
2. w ogólności każda akcja wpływa na:
   - natychmiastową nagrodę,
   - na kolejne sytuacje w których znajdzie się agent a więc w konsekwencji **na kolejne nagrody**

Powyższe dwie cechy są charakterystyczne dla uczenia ze wzmocnieniem.

RL vs uczenie nadzorowane

W uczeniu nadzorowanym agent ma dostęp do **poetykietowanego** zbioru uczącego. Każdy przykład to:

- opis sytuacji: stan środowiska,
- etykieta: poprawna akcja którą należy wykonać.

Celem uczenia jest generalizacja wiedzy zawartej w zbiorze uczącym na inne przypadki aby w nich również poprawnie działać.

RL vs uczenie nienadzorowane

Uczenie nienadzorowane:

- **brak poetykietowanego** zbioru uczącego
  (tak samo jak w **RL**!)
- zadaniem jest odkrycie struktury w danych (występujących wzorców)

RL vs uczenie (nie)nadzorowane

- Zarówno uczenie nadzorowane jak i nienadzorowane nie adresują bezpośrednio głównego problemu zadania uczenia ze wzmocnieniem: maksymalizacji nagrody.
- Jednak metody uczenia (nie)nadzorowanego są częścią algorytmów uczenia ze wzmocnieniem.

Porównanie — przykład

Agent grający w kółko i krzyżyk

- uczenie nadzorowane — dane to zbiór par:
  (plansza, **jaki ruch najlepiej wykonać**)
  - dla RL drugi element jest niedostępny
- uczenie nienadzorowane — dane to zbiór plansz
  - może np. pogrupować na plansze „przegrywające" i „jeszcze nie wiadomo", albo pogrupować symetryczne/obrócone plansze itd.
- uczenie ze wzmocnieniem — agent gra w grę i zdobywając doświadczenie odkrywa korzystne akcje
  - uczenie się przez interakcję

Elementy RL

Poza **agentem** i **środowiskiem** możemy wyodrębnić kilka istotnych elementów:

- **polityka** *(ang. policy)* – definiuje jak agent zachowuje się w danej sytuacji
- **nagroda** *(ang. reward)* – definiuje cel uczenia
  - w każdym kroku środowisko informuje agenta o nagrodzie (jest to pojedyncza liczba)
  - celem jest maksymalizacja całkowitej nagrody uzyskanej w długim okresie czasu (później doprecyzujemy)
- **funkcja wartości** *(ang. value function)* – definiuje co jest pożądane w długim okresie
  - „długoterminowa nagroda"
  - jest pochodną uzyskiwanych nagród
- **model** środowiska – naśladuje zachowanie środowiska (występuje w części algorytmów)
  - umożliwia wnioskowanie na temat tego jak zmieni się środowisko

Co będzie dalej?

- Metody „tablicowe”:
  - przestrzeń stanów i akcji jest mała – da się przedstawić w postaci tablic
  - główne idee metod uczenia ze wzmocnieniem
  - zaczniemy od najprostszych sytuacji i metod
- Generalizacja:
  - przestrzeń stanów i akcji jest olbrzymia/nieskończona
  - metody przybliżone

Plan wykładu

Wstęp

- w RL informacja ucząca mówi jak dobre były akcje podjęte przez agenta
- nie instruuje jakie akcje powinien był podjąć w danym stanie środowiska!
- w szczególności ocena „jak dobra" była dana akcja nie mówi nic o tym czy była ona najlepsza lub najgorsza
- potrzeba **eksploracji** czyli sprawdzenia jak dobre są pozostałe akcje

Przeanalizujemy teraz wyizolowany aspekt oceny akcji w bardzo uproszczonym (ale nadal praktycznym!) scenariuszu.

Wieloręki bandyta — definicja

Rozważmy problem w którym wielokrotnie:

- wybieramy jedną z k akcji
- po każdym wyborze dostajemy nagrodę ze **stacjonarnego rozkładu prawdopodobieństwa** zależnego od wybranej akcji
- celem agenta jest zmaksymalizowanie spodziewanej całkowitej nagrody w N krokach

Inaczej:

- w kasynie mamy k jednorękich bandytów
- każda maszyna wypłaca nagrody ze swoim własnym (**nieznanym** nam!) rozkładem prawdopodobieństwa
  (czyli jedna może średnio dawać wyższe wypłaty niż inna)
- jaką strategię przyjąć aby zmaksymalizować swój zysk?
  - wybrać jedną i grać tylko na niej?
  - grać cały czas na wszystkich?

Przykładowe problemy

- Lekarz podejmujący eksperymentalne terapie:
  - akcja: wybór konkretnego eksperymentalnego leku
  - nagroda: stan zdrowia po zastosowaniu leku
- Modyfikacja serwisu internetowego:
  - akcja: wybór koloru/rozmieszczenia elementów na stronie
  - nagroda: ilość kliknięć „kup”/czas spędzony w serwisie/itp.

Wartość akcji

- W problemie wielorękiego bandyty każda akcja ma spodziewaną wartość nagrody:

$$q_*(a) = \mathbb{E}[R_t | A_t = a]$$

  - R_t — nagroda otrzymana w kroku t
  - A_t — akcja wybrana w kroku t
- Gdybyśmy znali wartość $q_*(a)$ wystarczyłoby zawsze wybierać akcję maksymalizującą $q_*(a)$
- W praktyce nie znamy prawdziwej wartości akcji $q_*(a)$ ale możemy mieć pewną **estymatę** wartości akcji w kroku t:

$$Q_t(a)$$

Wartość akcji — estymata

- Prawdziwa wartość akcji to średnia wartość nagrody gdy ta akcja jest wybrana
- Możemy ją estymować jako:

$$Q_t(a) = \frac{\text{suma nagród gdy wybrano akcję } a \text{ do chwili } t}{\text{ile razy wybrano akcję } a \text{ do chwili } t}$$

gdy mianownik jest równy zeru przyjmujemy domyślną wartość (np. 0)

Metody

Zachłanna

- wybiera akcję

$$A_t = \arg\max_a Q_t(a)$$

- zawsze eksploatuje aktualną wiedzę
- nie sprawdza czy inna akcja może być lepsza

ϵ-zachłanna

- działa zachłannie, ale...
- z prawdopodobieństwem ϵ wybiera **losową** akcję (niezależnie od aktualnej estymaty ich wartości)
- w nieskończoności mamy gwarancję, że estymata $Q_t(a)$ zbiegnie do prawdziwej wartości $q_*(a)$

Przyrostowe liczenie wartości

Pożądane jest aby wartość estymaty wyliczać inkrementacyjnie bez potrzeby pamiętania całej historii wybieranych akcji i otrzymanych nagród.
Dla dowolnej akcji gdy była ona wybrana/wykonana $n-1$ razy mamy:

$$Q_n = \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

Przyrostowe liczenie wartości

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)Q_n\right) \\
&= \frac{1}{n}\left(R_n + nQ_n - Q_n\right) \\
&= Q_n + \frac{1}{n}\left(R_n - Q_n\right)
\end{aligned}
$$

Przyrostowe liczenie wartości

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

Często spotykana formuła:

$$NowaEstymata = StaraEstymata + \alpha\,(Cel - StaraEstymata)$$

gdzie α — rozmiar kroku

Problemy niestacjonarne

Problem niestacjonarny

Rozkład prawdopodobieństwa otrzymywanych nagród (za daną akcję) zmienia się w czasie.

- Dla problemów niestacjonarnych warto większą wagę przyjmować dla nagród otrzymywanych ostatnio a mniejszą do otrzymanych dawno temu
- Można to uzyskać np. przez stały rozmiar kroku α:

$$Q_{n+1} = Q_n + \alpha \left(R_n - Q_n \right)$$

Optymistyczna wartość początkowa

Przyjęta początkowa wartość $Q_1(a)$ (czyli wartość estymaty zanim akcja a została w ogóle wykonana) ma znaczenie:

- w przypadku stałego α: dla ostatecznej estymaty
  (dla $\alpha = \frac{1}{n}$ wpływ zanika po wybraniu akcji chociaż raz)
- w promowaniu eksploracji (w problemach stacjonarnych).

Optymistyczna wartość początkowa

Eksploracja: jeśli początkowa wartość jest **optymistyczna** (czyli $Q_1(a) > q_*(a)$) to:

1. agent wybiera pewną akcję a
2. otrzymuje nagrodę mniejszą od optymistycznego oszacowania — jest „zawiedziony" tą akcją (obniża oszacowanie jej wartości), więc
3. następnym razem spróbuje innej akcji — efekt: zanim estymaty się zbiegną wszystkie akcje będą kilkukrotnie przetestowane (nawet dla agenta w pełni zachłannego)

Optymistyczna wartość początkowa

Eksploracja: jeśli początkowa wartość jest **optymistyczna** (czyli $Q_1(a) > q_*(a)$) to:

1. agent wybiera pewną akcję a
2. otrzymuje nagrodę mniejszą od optymistycznego oszacowania — jest „zawiedziony" tą akcją (obniża oszacowanie jej wartości), więc
3. następnym razem spróbuje innej akcji — efekt: zanim estymaty się zbiegną wszystkie akcje będą kilkukrotnie przetestowane (nawet dla agenta w pełni zachłannego)

Pytanie

Czy taka metoda eksploracji zadziała dla problemów niestacjonarnych?

Metoda UCB

UCB — ang. Upper Confidence Bound

Idea: wybranie akcji zależy od szansy, że jest ona najlepszą, biorąc pod uwagę:

- jak bardzo jej estymata jest bliska maksymalnej wartości
- niepewności samej estymaty

Działanie:

$$A_t = \arg\max_a \left[Q_t(a) + c \cdot \sqrt{\frac{\ln t}{N_t(a)}} \right],$$

gdzie:
$N_t(a)$ — ile razy akcja a była wybrana do chwili t
$c > 0$ — stopień eksploracji

Bibliografia

[1] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, third edition.

[2] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

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