Zaawansowane Metody Inteligencji Obliczeniowej

Lab 2: Wieloręki bandyta - eksploracja i eksploatacja

Michał Kempka Marek Wydmuch

11 marca 2021

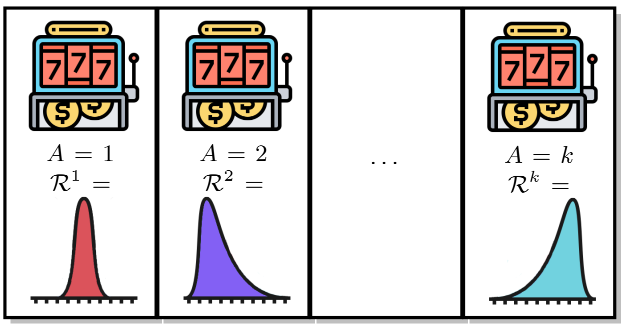
"Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)",
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

1 Wprowadzenie

Cechą charakterystyczną uczenia ze wzmocnieniem (ang. reinforcement learning (RL)) jest to, że bazuje ono na ocenie podjętych akcji (evaluative feedback), a nie instrukcji wskazującej poprawne akcje (instructive feedback), jak w przypadku np. uczenia nadzorowanego (ang. supervised learning). Te dwa rodzaje informacji zwrotnej różnią się tym, że pierwszy jest całkowicie zależy, a drugi całkowicie nie zależy od podjętych akcji. Ta różnica, w wypadku uczenia ze wzmocnieniem, stwarza potrzebę aktywnej **eksploracji** w celu poszukiwania dobrych akcji.

1.1 Wieloręki bandyta

Wieloręki bandyta (ang. multi-armed bandit albo k-armed bandit), nazwany tak z powodu analogi do automatów hazardowych zwanych "jednorękimi bandytami", to problem gdzie stajemy przed wyborem k-różnych opcji działań. Po każdym wyborze otrzymujemy nagrodę liczbową wybraną z rozkładu prawdopodobieństwa, który zależy od wybranej akcji. Czyli każda akcja jest jak zagranie na jednym z "jednorękich bandytów", a nagroda to wygrana.



Rysunek 1: Schemat problemu wielorekiego bandyty

Bardziej formalnie, zdefiniujmy problem jako:

- Wieloręki bandyta to krotka: $\langle \mathcal{A}, \mathcal{R} \rangle$,
- $\mathcal{A}$ jest zbiorem akcji ("ramion"),
- $\mathcal{R}^a = \mathbb{P}[R|A = a]$ jest nieznanym stacjonarnym rozkładem prawdopodobieństwa nagród.
- w każdym kroku t agent wybiera akcję $A_t \in \mathcal{A}$,
- środowisko generuje nagrodę $R_t \sim \mathcal{R}_{A_t}$,

- celem jest zmaksymalizowanie całkowitej nagrody $\sum_{t=1}^{T} R_t$.

Uwaga: obecnie termin problemu "wielorękiego bandyty" jest często używany do uogólnienia problemu kontekstowego wielorękiego bandyty (ang. contextual multi-armed bandit):

- Kontekstowy wieloręki bandyta to krotka: $\langle \mathcal{A}, \mathcal{S}, \mathcal{R} \rangle$,
- $\mathcal{S} = \mathbb{P}[S]$ jest nieznanym rozkładem stanów,
- $\mathcal{R}_s^a = \mathbb{P}[R|S = s, A = a]$ jest nieznanym rozkładem prawdopodobieństwa nagród,
- przed wyborem ramienia prezentowany jest stan z wyżej wymienionego rozkładu (kontekst).

Na tych laboratoriach jednak pozostaniemy przy prostej wersji problemu.

1.2 Wartość akcji (ang. action-value)

W naszym problemie wielorękiego bandyty, każda z akcji a ma oczekiwaną (średnią) nagrodę, biorąc pod uwagę, że ta akcja jest wybrana - nazwijmy to wartością akcji:

$$q^*(a) = \mathbb{E}[R_t | A_t = a]\,. \tag{1}$$

Gdybyśmy znali wartość każdej akcji, rozwiązanie problemu wielorękiego bandyty byłoby trywialne. Niestety nie znamy $q^*(a)$. Możemy ją jednak oszacować. Oszacowanie wartości akcji a w kroku t oznaczamy jako $Q_t(a)$. Zależy nam by $Q_t(a)$ było jak najbliższe $q^*(a)$.

W dowolnym kroku czasowym istnieje co najmniej jedna akcja, której wartość $Q_t(a)$ jest największa. Wybór takiej akcji nazywamy działaniem **zachłannym (ang. greedy)** i mówimy wtedy, że **eksploatujemy** naszą wiedzę o wartości akcji. Jeśli wybieramy akcję nie zachłannie, lecz by poszerzyć naszą wiedzę, dokonujemy wtedy **eksploracji**. Eksploatacja jest właściwym działaniem jeśli chcemy zmaksymalizować oczekiwaną nagrodę w danym kroku podczas gdy eksploracja **może** przynieść nam wiedzę pozwalającą nam osiągnąć wyższą nagrodę w dłuższym horyzoncie czasowym.

1.3 Estymacja wartość akcji

Przyjrzyjmy się bliżej prostej i jednoczesnej naturalnej metodzie estymacji wartości akcji - uśredniania otrzymanych nagród dla danej akcji a:

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i = a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}} \tag{2}$$

gdzie $\mathbb{1}$ oznacza indykator zbioru ($\mathbb{1}_{\text{predykat}} = 1$, jeśli predykat jest prawdziwy, w innym wypadku 0). Metoda ta często nazywana jest "prostą średnią" (ang. sample-average).

1.4 ϵ-greedy

Jest to bardzo prosta, lecz powszechnie stosowana metoda eksploracji bazująca na estymacji wartości stanów. Miesza ona wykorzystywanie aktualnej wiedzy z kompletnie przypadkowymi ruchami, zależnie od zadanego parametru $\epsilon \in [0, 1]$. W każdym kroku:

- z prawdopodobieństwem ϵ wykonaj losową akcję (np. wybierz losowe ramię)
- w przeciwnym razie wykonaj akcję **zachłanną** (***ang. greedy***), czyli maksymalizującą przewidywany zysk biorąc pod uwagę aktualny stan wiedzy: $A_t = \text{argmax}_{a \in \mathcal{A}} Q_t(a)$.

1.5 Żal (*ang. regret*)

By lepiej badać "jakość" algorytmów, zamiast rozważać całkowitą nagrodę, możemy również wykorzystać koncepcję żalu, która porównuje jakość danego algorytmu do najlepszego możliwego.

Optymalna wartość akcji:

$$v^* = q^*(a^*) = \max_{a \in \mathcal{A}} q^*(a)\,. \tag{3}$$

Całkowity oczekiwany żal:

$$L_T = \mathbb{E}\left[\sum_{t=1}^{T} \big(v^* - q^*(A_t)\big) \right]. \tag{4}$$

Maksymalizacja całkowitej nagrody $\equiv$ minimalizacja całkowitego żalu.

1.6 Inkrementacyjna estymacja wartości

W celu uproszczenia notacji oznaczmy Q_n jako estymację wartości akcji, po tym jak została ona wybrana $n-1$ razy i R_i jako nagrodę otrzymają po i-tym wyborze tej akcji:

$$Q_n = \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} R_i \tag{5}$$

Żeby nie musieć przechowywać całej historii nagród dla każdej akcji, powyższą średnią można przekształcić we wzór rekurencyjny:

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i = Q_n + \frac{1}{n}(R_n - Q_n) \tag{6}$$

Podczas przedmiotu często będziemy używać podobnych wzorów, o następującej ogólnej formie:

$$NowaEstymata \leftarrow StaraEstymata + RozmiarKroku[Cel - StaraEstymata]$$

Zauważ, że parametr "$RozmiarKroku$" (ang. $StepSize$), oznaczany jako α jest używany we wzorze 6 ($\alpha = 1/n$) i zmienia się z korku na krok.

Omówiona powyżej metoda uśredniania jest odpowiednia dla problemu stacjonarnego wielorękiego bandyty, czyli takiego gdzie ze stacjonarnymi rozkładami prawdopodobieństw - prawdopodobieństwa nagród nie zmienia się w czasie. Częściej jednak napotykamy problemy, które są niestacjonarne. W takich przypadkach sensowne jest przypisanie większej wagi nagrodom otrzymanym niedawno niż nagrodom z dalekiej przeszłości. Jednym z najpopularniejszych sposobów jest użycie stałego parametru wielkości kroku $\alpha \in [0, 1]$.

$$Q_{n+1} = Q_n + \alpha[R_n - Q_n] = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \tag{7}$$

2 Zadania

2.1 Wieloręki bandyta a prawdziwy świat

Wymień kilka problemów z prawdziwego świata analogicznych do problemu wielorękiego bandyty. Powiedz na czym polegałaby w danym problemie eksploatacja, a na czym eksploracja.

2.2 ϵ-greedy i żal

1. W strategii wyboru akcji ϵ-greedy, w wypadku gdy $|\mathcal{A}| = 4$ i $\epsilon = 0.5$, jakie jest prawdopodobieństwo, że zostanie wybrana zachłanna ("greedy") akcja?
2. Rozważ 4-rękiego bandytę, z akcjami $\mathcal{A} = \{1, 2, 3, 4\}$, do rozwiązania tego problemu zastosowano algorytm ϵ-greedy. Do estymacji wartości akcji użyto średniej dotychczasowych wartości i początkowymi estymatami $Q_1(a) = 0$, dla wszystkich a. Zaobserwowano następującą sekwencje akcji i nagród: $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = 2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. W których krokach nastąpiła eksploracja, a w których eksploatacja?
3. Jaka jest oczekiwana całkowita nagroda algorytmu, który z równym prawdopodobieństwem wybiera jedną z k dostępnych akcji ($|\mathcal{A}| = k$), o oczekiwanych nagrodach $\boldsymbol{q}^* = [q_1^*, \ldots, q_k^*]$?
4. Jaki jest oczekiwany żal algorytmu przedstawionego w punkcie 2.2.3?
5. Wyznacz dolne ograniczenie na oczekiwany żal dla algorytmu ϵ-greedy.
6. Na podstawie wyniku z 2.2.5 powiedz dla jakich wartości ϵ algorytm ϵ-greedy będzie najlepszy z punktu widzenia długiego horyzontu czasowego (dla bardzo dużego T).
7. Zauważ, że ϵ, k, wektor $\boldsymbol{q}^*$ to stałe, dla ϵ-greedy żal rośnie więc liniowo z czasem. Jak możemy zmodyfikować algorytm ϵ-greedy by potencjalnie osiągnąć mniejszy żal?

2.3 Wartość akcji

1. Pokaż, że wzór rekurencyjny na utrzymywanie średniej jest poprawny:

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i = Q_n + \frac{1}{n}(R_n - Q_n)\,.$$

2. Rozważ wzór 7. Jeśli parametr α nie będzie stałą, wtedy estymata Q_n jest inną średnią ważoną poprzednio otrzymanych nagród niż ta we wzorze 7. Analogicznie do wzoru 7 wyznacz ważenie dla każdej poprzedniej nagrody (wzór zamknięty nie rekurencyjny).

Literatura

[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