Zadanie 6. Uczenie ze wzmocnieniem

Julia Jodczyk

20 stycznia 2022

1 Treść zadania

Zaimplementuj algorytm Q-learning. Następnie, wykorzystując proste środowisko (np. Taxi-v3), zbadaj wpływ hiperparametrów na działanie algorytmu (np. wpływ strategii eksploracji, współczynnik uczenia).

2 Wstęp

Uczenie się ze wzmocnieniem jest formą maszynowego uczenia się, gdzie agent dynamicznie oddziałuje ze środowiskiem dążąc do maksymalizowania przyszłych nagród wynikających z jego decyzji w danym stanie.

Ogólnym algorytm uczenia się agenta wygląda następująco:

1. wybierz akcję,
2. pobierz odpowiedź od środowiska,
3. pobierz nowy stan,
4. wybierz nową akcję.

Agent może wybierać akcję na dwa sposoby - losowy (eksploracja) lub korzystając ze zdobytej wiedzy (eksploatacja).
Q-learning jest jednym z algorytmów uczenia się ze wzmocnieniem. Wiedza o środowisku zdobyta przez agenta przechowywana jest w strukturze *Q-table* o wymiarach ilość akcji × ilość stanów, która początkowo wypełniona jest zerami. Po otrzymaniu przez agenta nagrody za podjętą decyzję tablica jest aktualizowana. Do wybrania akcji maksymalizującej przyszłą nagrodę korzysta się ze wzoru:

$$Q^{\pi}(s_t, a_t) = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...][s_t, a_t] \tag{1}$$

gdzie γ to dyskonto, czyli parametr, który decyduje jak ważne w danym stanie są długotrwałe nagrody.

Wartości w tabeli Q-table aktualizuje się według wzoru:

$$NewQ(s, a) = Q(s, a) + \alpha[R(s, a) + \gamma maxQ'(s', a') - Q(s, a)] \tag{2}$$

gdzie α to współczynnik uczenia się.

To, czy w danym stanie agent wybierze akcję na podstawie informacji w *Q-table* czy ją wylosuje zależy od wybranej polityki. Może nią być polityka ϵ-zachłanna albo Boltzmanna.

2.1 Polityka ϵ-zachłanna

Polityka zakłada wylosowanie numeru z przedziału (0,1) przed wybraniem akcji. Jeśli number jest mniejszy niż ϵ - parametr polityki - akcja jest losowana. W przeciwny wypadku wybierana jest akcja, dla której wartość w *Q-table* jest największa.

2.2 Polityka Boltzmanna

Polityka zakłada, że za każdym razem akcja jest losowana, ale akcje prowadzące do wyższych nagród mają większe prawdopodobieństwo wybrania. Prawdopodobieństwo akcji oblicza się ze wzoru:

$$P(a|s) = \frac{exp(Q(s,a)/T)}{\sum_{a' \in A} exp(Q(s,a')/T)} \tag{3}$$

gdzie T to tzw. temperatura - parametr polityki.

3 Opis implementacji

Agent zaimplementowany jest jako klasa, środowiska dostarcza biblioteka gym. Obiekt agenta posiada atrybuty Q-table i hiperparametry: współczynnik uczący, dyskonto i polityka oraz metody wykonujące wybór akcji oraz aktualizację Q-table.

4 Opis i wyniki eksperymentów

Przy badaniu wpływu hiperparametrów na działanie agenta będę mierzyła skumulowaną przy uczeniu nagrodę oraz średnią ilość kroków w epizodzie. Spośród tych dwóch badanych wielkości to nagrody mają większe znaczenie, jednak czasami istotne jest też zaobserwowanie czasu dochodzenia do stanu terminalnego. Hiperparamerami w algorytmie są:

- współczynnik uczenia się
- dyskonto
- polityka
- parametr charakterystyczny dla polityki - współczynnik eksploracji (epsilon) dla ϵ-zachłannej oraz temperatura dla Boltzmanna.

Eksperymenty dla parametrów powtarzałam 5 razy, wyniki w tabelach przedstawiają uśrednione wartości.

We wszystkich eksperymentach ilość epizodów wynosiła 10000 dla powtórzeń treningowych, 100 dla testowych, a maksymalna ilość kroków to 100.
Wyjściowe wartości badanych parametrów to:

- współczynnik uczenia się : 1
- dyskonto : 0.9
- współczynnik eksploracji : 0.1
- temperatura : 0.1

4.1 polityka ϵ-zachłanna

4.1.1 współczynnik uczenia się

współczynnik uczenia się	0.25	0.5	0.75	1
skumulowana nagroda	4256.2	10921.4	12465.2	14417.8
średnia ilość kroków	15.2	16	14.2	14.8

Tabela 1: Skumulowana nagroda i średnia ilość kroków przy zmiennym współczynniku uczenia się.

Współczynnik uczenia się decyduje o tym, w jakim stopniu nowa zdobyta przez agenta wiedza nadpisze dotychczasową informację. Im jest mniejszy, tym mniejsze znaczenie mają nowe informacje dostarczane przez środowisko. W deterministycznym środowisku, czyli w takim, jakie testujemy, agent powinien uczyć się najlepiej dla dużych wartości tego współczynnika, najlepiej dla maksymalnej, czyli 1. To właśnie obserwujemy w wynikach eksperymentów. Im większym współczynnik uczenia się, tym większa była skumulowana nagroda. Jeśli chodzi średnią ilość kroków, to nie ma znaczących różnic między ich wartościami.

4.1.2 dyskonto

dyskonto	0.1	0.3	0.6	0.9
skumulowana nagroda	-11825.6	8322.6	11736.8	14417.8
średnia ilość kroków	20.59	16.8	13	14.8

Tabela 2: Skumulowana nagroda i średnia ilość kroków przy zmiennym dyskoncie.

Od dyskonta zależy jak ważne dla agenta są przyszłe nagrody. Przy wartościach bliskich 0 agent jest krótkowzroczny i dąży do maksymalizowania tylko tymczasowych nagród. Natomiast przy większym dyskoncie stara się zwiększać nagrody

długoterminowe. W wynikach przeprowadzonych testów widać, że wraz ze wzrostem dyskonta wzrasta również wartość skumulowanych nagród. Wartość 0.1 jest zbyt mała, aby agent mógł się nauczyć, przez co skumulowana nagroda jest znacząco mniejsza od zera. Żeby przy takim dyskoncie osiągnąć lepsze wyniki należałoby zwiększyć ilość epizodów i maksymalnych kroków w epizodzie.

4.1.3 współczynnik eksploracji

współczynnik eksploracji	0.01	0.1	0.5	0.9
skumulowana nagroda	63737.4	14417.8	-466326.8	-411336
średnia ilość kroków	13.2	14.8	26.2	100

Tabela 3: Skumulowana nagroda i średnia ilość kroków przy zmiennym współczynniku eksploracji.

Współczynnik eksploracji decyduje o tym, jak bardzo agent będzie skłonny do eksploracji swojego środowiska. Im wyższy ten współczynnik, tym częściej wybierany będzie losowy stan zamiast tego o największej potencjalnej nagrodzie, im mniejszy, tym bardziej losowe będzie zachowanie agenta. W testowanym problemie najkorzystniejszy był mały współczynnik eksploracji. Dla wartości większych od 0.1 agent nie korzysta wystarczająco często ze zdobytej wiedzy, przez co nie uczy się wystarczająco dobrze.

4.2 polityka Boltzmanna

4.2.1 współczynnik uczenia się

współczynnik uczenia się	0.25	0.5	0.75	1
skumulowana nagroda	59497.2	67479	67815.2	68604.2
średnia ilość kroków	15.75	14.22764	13.858	13.737

Tabela 4: Skumulowana nagroda i średnia ilość kroków przy zmiennym współczynniku uczenia się.

Podobnie jak przy polityce ϵ-zachłannej współczynnik uczenia równy 1 daje najlepsze wyniki, jednak różnice między osiągniętymi wynikami są stosunkowo małe, a suma nagród jest mniejsza niż przy polityce ϵ-zachłannej.

4.2.2 dyskonto

dyskonto	0.1	0.3	0.6	0.9
skumulowana nagroda	-	63478.2	67940.2	68604.2
średnia ilość kroków	-	14.943	13.885	13.737

Tabela 5: Skumulowana nagroda i średnia ilość kroków przy zmiennym dyskoncie.

Dla dyskonta = 0.1 wartości prawdopodobieństwa były za małe, żeby algorytm zadziałał poprawnie. Najlepszy wynik ponownie uzyskano dla wysokiej wartości.

4.2.3 temperatura

temperatura	0.1	1	10	30
skumulowana nagroda	68604.2	62124.8	-164124.2	-48663
średnia ilość kroków	13.737	15.53	43.83	98.2

Tabela 6: Skumulowana nagroda i średnia ilość kroków przy zmiennej temperaturze.

Od temperatury zależne jest prawdopodobieństwo wyboru każdego ze stanów. Przy większych wartościach wszystkie stany stają się równo prawdopodobne, dlatego przy badanych wartościach 10 i 30 otrzymujemy wyniki podobne, co przy polityce ϵ-zachłannej i dużym współczynniku eksploracji.

5 Wnioski

- Kiedy mamy do czynienia z deterministycznym środowiskiem, to najlepszą wartością współczynnika uczącego się jest 1.
- Agent krótkowzroczny nie bierze pod uwagę długotrwałych nagród, przez co nie uczy się wystarczająco dobrze. Przy większych wartościach dyskonta agent skuteczniej maksymalizuje przyszłe nagrody.
- Polityka Boltzmanna jest bardziej odporna na dobór złych parametrów, ale skumulowane nagrody były mniejsze niż przy polityce ϵ-zachłannej.
- Duże wartości współczynnika eksploracji sprawiają, że agent rzadko korzysta ze zdobytej wiedzy i zamiast tego działa w losowy sposób. Małe wartości sprawiają, że agent bazuje na wiedzy, co jakiś czas losując swoją akcję, co pozwala na eksplorację środowiska.
- Duże wartości temperatury sprawiają, że wszystkie akcje stają się mniej więcej równo prawdopodobne, natomiast mniejsze bardziej faworyzują akcje maksymalizujące przyszłe nagrody.