Wprowadzenie do RL 3

Zadanie 1

A. Rzucamy dwukrotnie monetą. Interesuje nas to ile razy wypadł orzeł. Nagrody:

   R(0)=10, R(1)=100, R(2)=300

   Jaka jest wartość oczekiwana nagrody?

B. Rzucamy kostką do gry. Jeżeli wypadnie liczba nieparzysta wówczas nagroda jest równa tej liczbie. Jeżeli wypadnie liczba parzysta wówczas R = -4. Jaka jest wartość oczekiwana nagrody?

C. Wybieramy 3 różne liczby z przedziału 1,…,12. Jeżeli wskażesz poprawnie liczby wygrywasz 100$. Jakie są Twoje przewidywania co do nagrody jeżeli jedna gra kosztuje 1$.

Zadanie 2

Rzucamy kostką do gry. Jeżeli wypadnie liczba nieparzysta wówczas nagroda jest równa tej liczbie. Jeżeli wypadnie liczba parzysta wówczas R = -4.

A. Jaka jest wartość oczekiwana nagrody pod warunkiem, że wypadnie liczba nieparzysta.
B. Jaka jest wartość oczekiwana nagrody pod warunkiem, że wypadnie liczba parzysta.

Zadanie 3

Opracuj trzy przykładowe problemy, które mogą być zrealizowane jako **MDP**. Dla każdego z nich podaj zbiory stanów S, akcji A i nagród R. Postaraj się, aby przykłady były maksymalnie różne.

Zadanie 4

Rozważ problem jazdy samochodem. Akcje można zdefiniować przynajmniej na 3 sposoby:

- Wykorzystując pedał przyśpieszenia, kierownicę i hamulec, czyli miejsce, w którym ciało kierowcy styka się z maszyną.
- Możemy rozważyć miejsce, gdzie guma spotyka się z drogą. Akcje będą wówczas momentami obrotowymi.
- Możemy też rozważyć miejsce w którym mózg spotyka się z ciałem kierowcy. Działania będą wówczas skurczami mięśni, które kontrolują kończyny.

Jaki jest właściwe miejsce do wyznaczenia granicy między agentem i środowiskiem? Na jakiej podstawie jedna lokalizacja tej granicy jest lepsza od innej? Czy jest jakiś podstawowy powód preferowania jednej jej lokalizacji nad inną, czy też jest to swobodny wybór?

Zadanie 5

Załóżmy, że problem pole-balancing traktujemy jako zadanie epizodyczne ze zdyskontowaniem. Przyjmujemy, że wszystkie nagrody równe są 0, z wyjątkiem -1 po niepowodzeniu.

Jaki będzie zwrot za każdym razem? Jak ten zwrot różni się od przypadku ze zdyskontowaniem i ruchem ciągłym?

Zadanie 6

Wyobraź sobie, że projektujesz robota, który ma wyjść z labiryntu. Przyznajesz nagrodę +1 za ucieczkę z labiryntu i nagrodę 0 w pozostałych sytuacjach. Zadanie w naturalny sposób dzieli się na epizody - kolejne przejścia przez labirynt. Postanawiasz potraktować to zadanie jako zadanie epizodyczne, którego celem jest maksymalizacja oczekiwanej sumy nagroda.

Po uruchomieniu agenta i uczeniu przez pewien czas okaże się, że widoczny jest brak poprawy (agent nie uczy się). Co należy poprawić? Jak skutecznie przekazać agentowi to co chcesz osiągnąć?

Zadanie 7

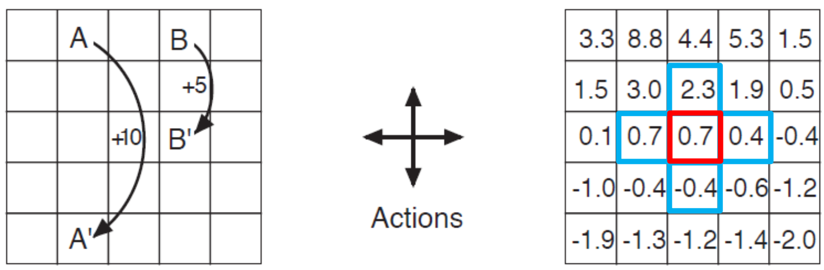
1. Załóżmy, że γ = 0.5 oraz, że agent otrzymuje następujące nagrody $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$ i $T = 5$. Wylicz $G_0, G_1, \ldots, G_5$?
2. Załóżmy, że γ = 0.9 i ciąg nagród wygląda następująco: R_1 =2 po której następuje nieskończona liczba 7. Wylicz G_0, G_1.

Zadanie 8

Udowodnij, że dla $\gamma < 1$ i wszystkich nagród $R_t = 1$ otrzymujemy: $G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$.

Zadanie 9

Poniższy rysunek (po lewej) pokazuje prostokątną reprezentację prostego świata. Komórki siatki odpowiadają stanom środowiska. W każdej komórce możliwe są cztery akcje: góra, dół, lewo i prawo, które deterministycznie powodują, że agent przesuwa jedną komórkę w odpowiednim kierunku na siatce.



Ruch agenta z pola brzegowego poza planszę skutkuje nagrodą R=-1. Pozostałe akcje przynoszą nagrodę R=0 z wyjątkiem stanów A i B. W stanie A wszystkie cztery akcje dają nagrodę +10 i przenoszą agenta do A'. W stanie B wszystkie akcje dają nagrodę +5 i przenoszą agenta do B'.

Pokaż, że równanie Bellmana

$$v_\pi(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

zachodzi dla stanu środkowego o wartości +0.7 ze względu na sąsiadów o wartościach +2.3, +0.4, −0.4, and +0.7. Przyjmij, że:

$\pi(a|s)=¼$

Dla dowolnego stanu *s* i akcji *a*, które mogą być wykonane w tym stanie.