Wprowadzenie do sztucznej inteligencji - ćwiczenie 6

Igor Kraszewski
310164

Warszawa, Styczeń 2023

Spis treści

1. Zadanie

Proszę zaimplementować algorytm Q-Learning i użyć go do wyznaczenia polityki decyzyjnej dla problemu FrozenLake8x8 (w wersji domyślnej, czyli z włączonym poślizgiem). W problemie chodzi o to, aby agent przedostał się przez zamarznięte jezioro z pozycji 'S' do pozycji 'G' unikając punktów 'H'. Symulator dla tego problemu można pobrać z podanej strony lub napisać własny o takiej samej funkcjonalności.
Oprócz zbadania domyślnego sposobu nagradzania (1 za dojście do celu, 0 w przeciwnym przypadku) proszę zaproponować własny system nagród i kar, po czym porównać osiągane wyniki z wynikami systemu domyślnego.
Za wynik (podczas testowania) uznajemy procent dojść do celu w 1000 prób (10x więcej prób używamy w treningu). W każdej próbie można wykonać maksymalnie 200 akcji.

2. Wyniki

Wszystkie wyniki zbierane są jako statystyki dla testów przy uruchomieniu 25 razy. Parametr ϵ używany jest przy wyborze akcji - strategia $\epsilon - zachanna$. Wybór akcji w ten sposób zakłada rozstrzygnięcie wymiany pomiędzy eksploracją, a eksploatacją na zasadzie wyboru akcji losowej z prawdopodobieństwem ϵ, a akcji zachłannej z prawdopodobieństwem $1 - \epsilon$.

2.1. Bazowa funkcja ewaluacji

W bazowej funkcji ewaluacji otrzymuje się 1 punkt za dojście do celu, natomiast 0 za każde inne zakończenie gry.

Tab. 2.1. Badanie wpływu wartości ϵ

ϵ	γ	β	Średnia	Std	Najlepszy	Najgorszy
0,1	0,9	0,1	47,43	16,81	72	17,9
0,2	0,9	0,1	45,75	17,53	72,1	17,9
0,3	0,9	0,1	48,82	18,56	79,9	12,4
0,4	0,9	0,1	49,94	18,03	81,3	12,4
0,5	0,9	0,1	44,23	17,36	74,2	12,0
0,6	0,9	0,1	44,17	15,63	74,2	12,0

Tab. 2.2. Badanie wpływu wartości γ

ϵ	γ	β	Średnia	Std	Najlepszy	Najgorszy
0,4	0,8	0,1	34,16	19,87	86,1	5,2
0,4	0,9	0,1	49,94	18,03	81,3	12,4
0,4	0,99	0,1	59,9	24,34	90,4	11,3

Tab. 2.3. Badanie wpływu wartości β

ϵ	γ	β	Średnia	Std	Najlepszy	Najgorszy
0,4	0,99	0,1	78,58	8,97	88,7	58,0
0,4	0,99	0,2	74,87	12,43	90,4	53,3
0,4	0,99	0,3	68,59	14,40	87,7	27,1
0,4	0,99	0,4	62,51	16,97	87,7	5,2

2.2. Własna funkcja nagradzania

Funkcja nagradzania, którą wybrałem za dojście do celu daje 100 punktów, za wpadnięcie do dziury -100, a za zakończenie na lodzie 0 punktów.

Tab. 2.4. Badanie wpływu wartości ϵ

ϵ	γ	β	Średnia	Std	Najlepszy	Najgorszy
0,1	0,9	0,1	77,29	9,68	87,6	59,6
0,2	0,9	0,1	77,68	8,94	88,4	57,7
0,3	0,9	0,1	77,28	9,41	88,4	56,9

Tab. 2.5. Badanie wpływu wartości γ

ϵ	γ	β	Średnia	Std	Najlepszy	Najgorszy
0,2	0,8	0,1	73,24	10,42	86,9	54,5
0,2	0,9	0,1	74,34	10,20	88,6	54,5
0,2	0,99	0,1	78,20	10,27	91,4	54,5

Tab. 2.6. Badanie wpływu wartości β

ϵ	γ	β	Średnia	Std	Najlepszy	Najgorszy
0,2	0,99	0,1	82,85	5,99	88,8	67,1
0,2	0,99	0,2	81,37	7,97	90,6	51,8
0,2	0,99	0,3	82,14	8,23	92,2	51,8

Dobór parametrów jest istotny dla dobrego działania algorytmu. Zarówno dobór parametru ϵ do wyboru akcji aby zachodził dobry podział pomiędzy eksploracją, a eksploatacją, jak również parametrów γ oraz β. Wartość parametru ϵ dla bazowej funkcji nagrody była wyższa, może to wynikać z tego, że jedyną nagrodę różną od zera otrzymuje się, gdy gra zakończy się sukcesem. Parametr β czyli współczynnik uczenia wpływa na to jak duże poprawki chcemy wprowadzać po każdym ruchu. Dla obu sposobów nagradzania najlepsze wyniki otrzymano dla wartości 0,1. Współczynnik γ nazywany też współczynnikiem dyskontowania reguluje ważność krótko i długoterminowych wzmocnień - jako, że w naszej grze wzmocnienie otrzymujemy na jej końcu współczynnik za każdym razem przyjmował dużą wartość (0,99). Widoczna jest spora losowość wyników (duże wartości odchyleń standardowych, różnice pomiędzy najlepszym oraz najgorszym wynikiem), mimo uśredniania wyników z 25 prób wciąż możliwe było otrzymanie znacząco różnych wyników przy tych samych parametrach, co widać w tabelach. Może być to związane z elementem gry jakim jest poślizg. Najlepszy wynik dla własnej funkcji nagrody jest wyższy. Odpowiednie dobranie funkcji nagrody jest ważne dla poprawnego działania uczenia ze wzmocnieniem, dzięki temu algorytm jest w stanie szybciej i lepiej reagować na stan w jakim się znajduje nawet mimo poślizgów wprowadzających losowość. Uważam, że nałożenie kary za wpadnięcie do dziury znacząco różnej od tej za zakończenie gry na lodzie mogło mieć duże znaczenie.