Wprowadzenie do sztucznej inteligencji - Laboratorium 6

Ireneusz Okniński 310228

Styczeń 2022

1 Treść zadania

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

2 Implementacja

2.1 Trenowanie modelu

Zaimplementowany został algorytm Q-learning z epizodami. Metoda train() posiada hiperparametry:

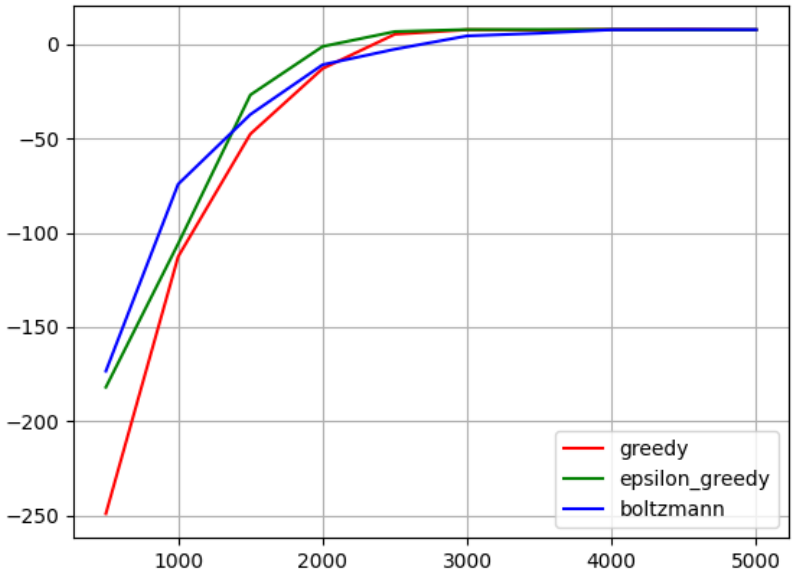
- episodes - liczba powtórzeń algorytmu
- learning_rate - współczynnik uczenia, mówi o tym jak bardzo będą zmieniać się wartości tablicy Q.
- discount_factor - współczynnik dyskontowania, reguluje względną ważność krótko i długoterminowych wzmocnień.
- strategy - definiuje zasady zgodnie z którymi wybierane są akcje. Dostępne strategie:
  - greedy - strategia zachłanna, agent zawsze wybiera najlepszą akcję.
  - epsilon_greedy - strategia zachłanna z parametrem epsilon, mówiącym o tym jak często agent wybiera losową decyzję.
  - boltzmann - strategia oparta na rozkładzie Boltzmanna, posiada parametr temperature, który reguluje stopień losowości wyboru.

2.2 Ocena modelu

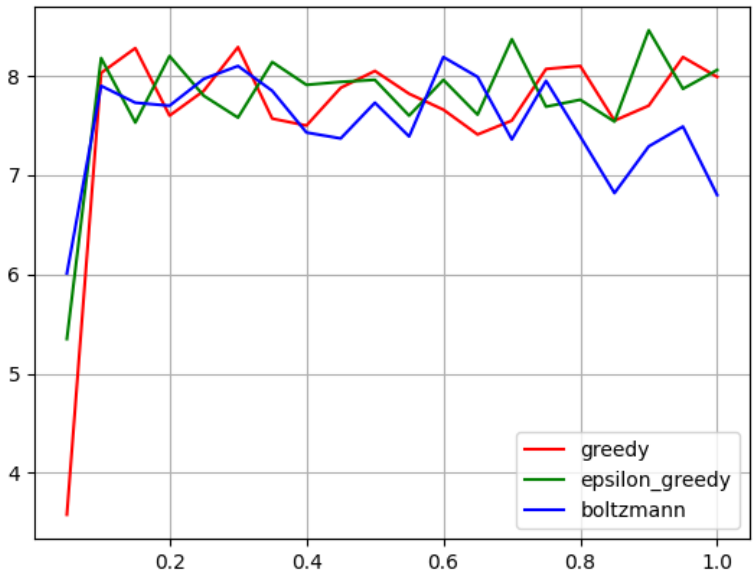
Ocena wytrenowanego modelu odbywa sie w metodzie evaluate. Przyjmuje ona jako paremetr ilość epizodów. Zwraca ona średnią nagrodę ze wszystkich uruchomień algorytmu. Dodatkowo możliwe jest ustawienie w niej parametru render, który umożliwia wyświetlanie środowiska oraz agenta podczas wykonywania zadań.

3 Eksperymenty numeryczne

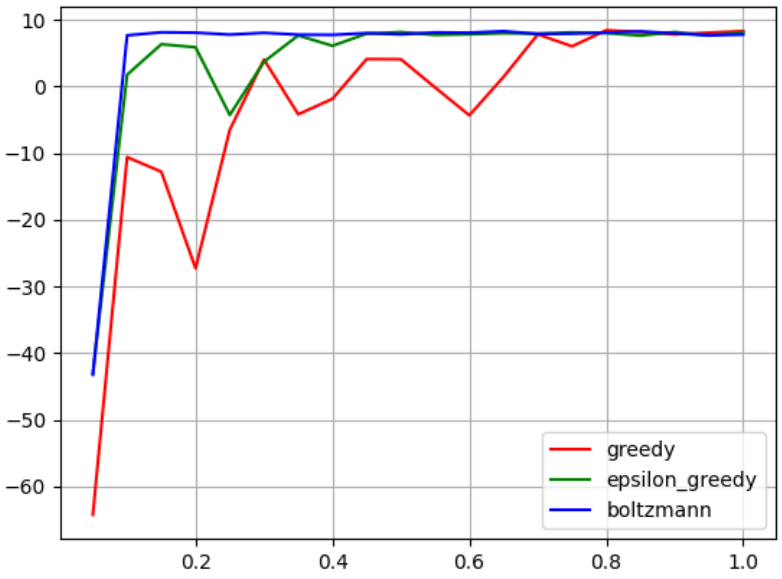
3.1 Wpływ ilości epizodów



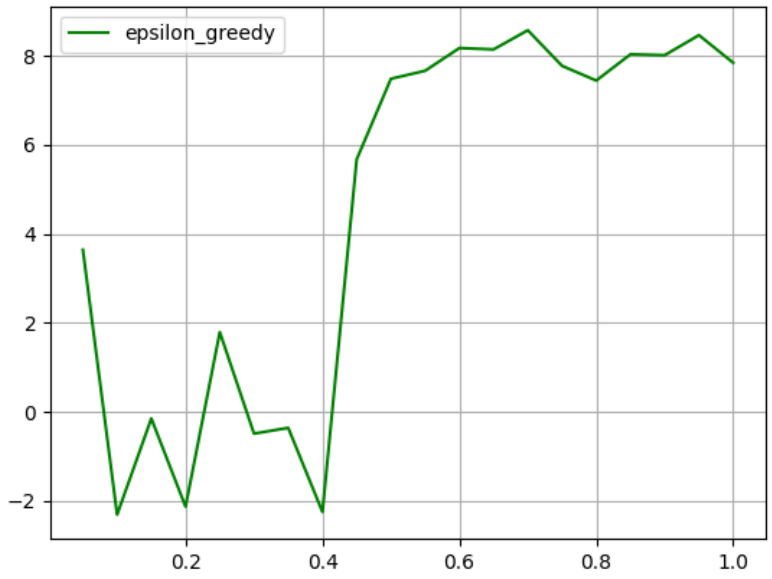
3.2 Wpływ parametru learning rate



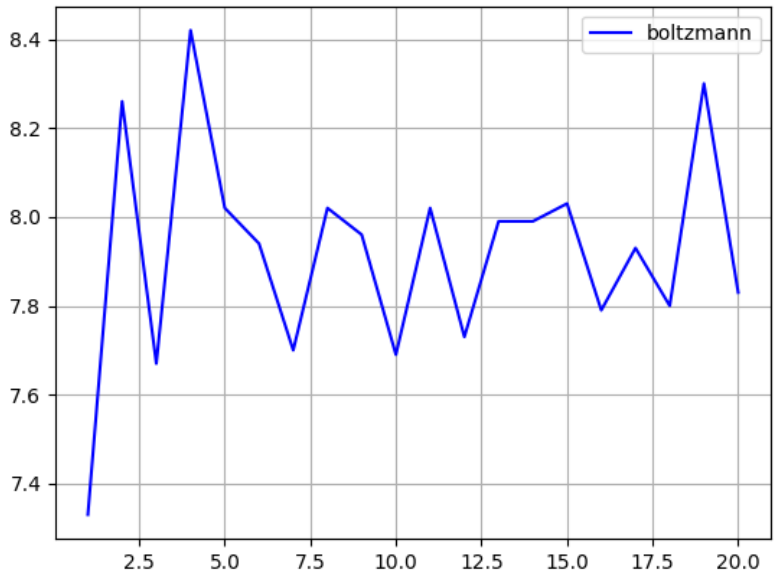
3.3 Wpływ parametru discount factor



3.4 Wpływ parametru epsilon w strategii epsilon-zachłannej



3.5 Wpływ parametru temperature w strategii Boltzmanna



4 Wnioski

4.1 Wpływ ilości epizodów

Wyskokość średniej nagrody rośnie wraz ze wzrostem ilości epizodów. Algorytm ma więcej czasu na odpowiedznie dostosowanie tablicy Q, tak aby nagroda była jak najwyższa. Strategia epsilon-zachłanna najszybciej osiąga dodatnie wartości średniej nagrody. Od ok. 4000 epizodu wszystkie algorytmy przynoszą bardzo podobne rezultaty. Dalsze kontynuowanie treningu nie przyniosłoby poprawy.

4.2 Wpływ parametru learning rate

Dla wartości learning rate poniżej 0,1 każda ze strategii przynosi słabsze rezultaty. Dla pozostałych wartości parametru średnia nagroda jest w przybliżeniu stała. Nieznacznym wyjątkiem jest strategia Boltzmanna, której wyniki mają tendencję spadkową dla learning rate większego niż 0,6. Odpowiednia wartość learning rate zapewnia dopasowanie algorytmu do środowiska.

4.2.1 Wpływ parametru discount factor

Każda ze strategii wymaga do najlepszego działania innej wartości parametru discount factor. Strategia Boltzmanna osiąga najlepsze wyniki dla parametru wiekszego od ok. 0,15. W przypadku strategii epsilon-zachłannej discount factor powinien być większy od ok. 0,35. Natomiast strategia zachłanna wymaga wartości parametru powyżej 0,7. Algorytm preferuje większe wartości tego parametru, co oznacza, że waga długoterminowych nagród jest ważna dla optymalnego rozwiązania.

4.3 Wpływ parametru epsilon w strategii epsilon-zachłannej

Poniżej wartości epsilon równej 0,4 średnia nagroda algorytmu jest losowa. Powyżej 0,4 wyniki stabilizują się na optymalnym poziomie.

4.4 Wpływ parametru temperature w strategii Boltzmanna

Wyniki treningu wykorzystującego strategię Boltzmanna nie są bardzo zróżnicowane. Wynik nieco mniejszy niż mogłoby to wynikać z losowości algorytmu występuje dla temperature = 1. Może to oznaczać, że ta wartość jest mniej korzystna.