WSI - Uczenie się ze wzmocnieniem

Marcin Szymczak 314910

1 Opis badanej metody

Pseudokod:

1. Inicjalizacja tablicy $Q[s, a]$ wartościami zerowymi, gdzie s - stan, a - akcja
2. Pętla przechodząca po maksymalnej liczbie epizodów
   (a) Resetowanie środowiska s = stan początkowy.
   (b) Przerwanie jeśli stan jest zakończony (done lub truncated)
      i. Wybór akcji a według strategii eksploracji:
         - Jeżeli strategia to **epsilon-greedy**:
           – Wybranie losowej akcji, która musi być mniejsza od ϵ.
           – Jeżeli nie, wybierana jest akcja maksymalizującą $Q(s, a)$: $a = \arg\max_a Q(s, a)$.
         - Jeżeli strategia to **Boltzmann**:
           – Obliczenie rozkładu prawdopodobieństwa dla akcji:

$$P(a) = \frac{\exp(Q(s, a)/\epsilon)}{\sum_{a'} \exp(Q(s, a')/\epsilon)}.$$

           – Wybranie akcji na podstawie rozkładu $P(a)$.
      ii. Wykonanie akcji a:
         - Następny stan s',
         - Nagrodę r,
         - Informację, czy epizod został zakończony (`done`) lub dla środowiska Taxi v-3 *epizod* >= 200 (*truncated*).
      iii. Aktualizacja wartości $Q(s, a)$:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \cdot \left[r + \gamma \cdot \max_{a'} Q(s', a') - Q(s, a) \right].$$

      iv. Nowy stan: $s \leftarrow s'$.
   (c) Zmniejszenie eksploatacji ϵ: $\epsilon \leftarrow \epsilon \cdot$ `epsilon_decay`.
3. Zwraca tablicę $Q[s, a]$ i listę nagród.

2 Planowane eksperymenty numeryczne

Zbadanie wpływu współczynnika uczenia oraz strategii eksploracyjnej na działanie algorytmu.
Badane wartości *learning_rate*:

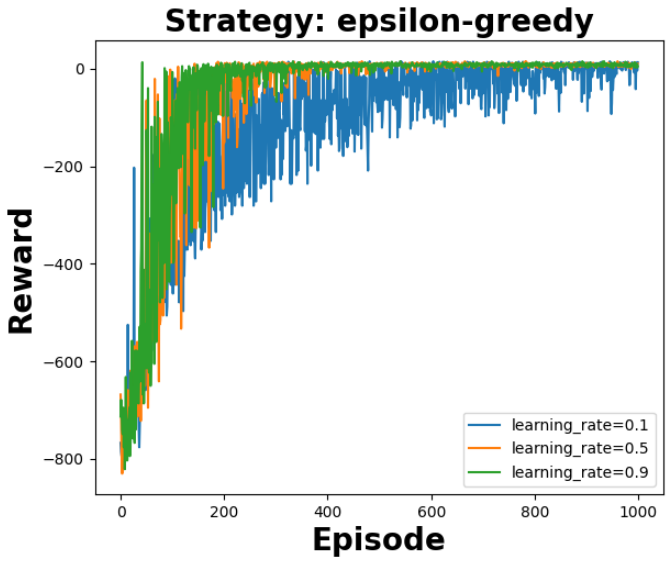
$$0.1, 0.5, 0.9$$

Strategie eksploracyjne: **epsilon-greedy**, **Boltzmann**

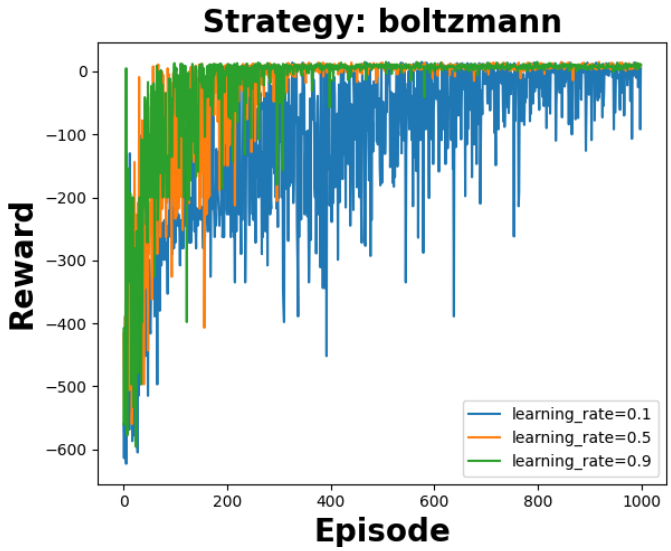
3 Warunki eksperymentów numerycznych

- Środowisko "Taxi-v3" z biblioteki gymnasium
- Maksymalna liczba epizodów = 1000
- Współczynnik dyskontowania $gamma = 0.99$
- Współczynnik eksploracji $epsilon = 1$
- Współczynnik zmniejszenia eksplorcaji $\epsilon_decay = 0.99$

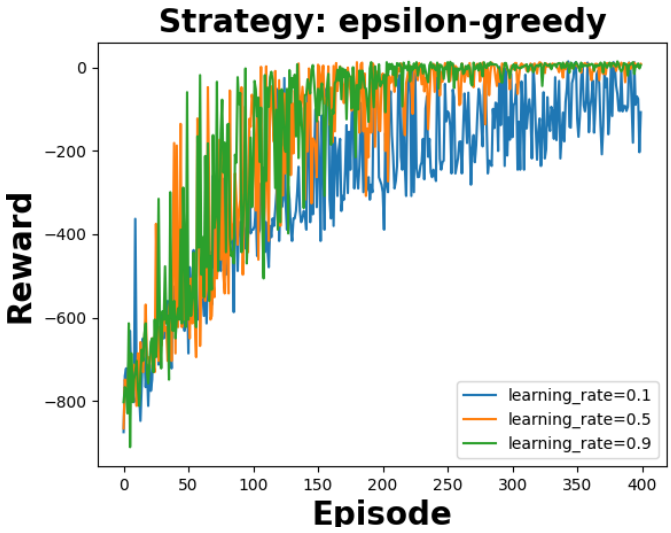
4 Otrzymane Wyniki



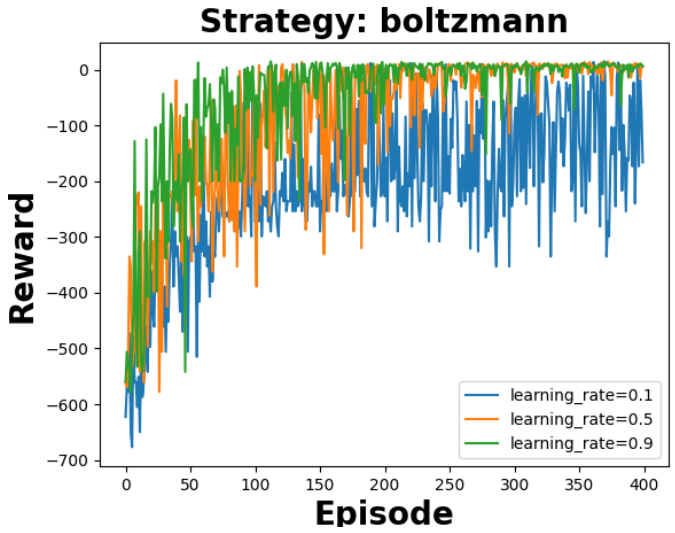
Rysunek 1: Wyniki dla strategii epsilon-greedy: porównanie nagród w kolejnych epizodach dla różnych wartości współczynnika uczenia.



Rysunek 2: Wyniki dla strategii Boltzmanna: porównanie nagród w kolejnych epizodach dla różnych wartości współczynnika uczenia.



Rysunek 3: Wyniki dla strategii Boltzmanna: porównanie nagród w kolejnych epizodach dla różnych wartości współczynnika uczenia dla liczby epizodów = 400.



Rysunek 4: Wyniki dla strategii Boltzmanna: porównanie nagród w kolejnych epizodach dla różnych wartości współczynnika uczenia dla liczby epizodów = 400.

5 Wnioski

- Współczynnik uczenia wpływa na szybkość nauki. Wyższe wartości współczynnika w obu metodach szybciej się zaadoptowały.
- Najniższa wartość współczynnika = 0.1 osięgneła oczekiwany wynik dużo później.
- Natomiast zmiana wartości dla najmniejszego współczynnika wydaje się być w miarę stabilna.
- Co do metod to widać jednoznacznie że metoda Epsilon-greedy jest stabilniejsza, ponieważ stopniowo zwiększa eksplorację.
- Boltzmann dla pierwszych epizodów ma bliższe wartości niż epsilon-greedy.
- Boltzmann natomiast ma zdecydowane większe wahania wyników. Jest to spowodowane tym, że ta metoda bardziej eksploruję przestrzeń.