Dokumentacja zadania laboratoryjnego - badanie perceptronu dwuwarstwowego

Krzysztof Wyrzykowski, nr indeksu 331455

22 stycznia 2025

1 Algorytm Q-learning

1.1 Opis działania algorytmu

Algorytm Q-learning to algorytm uczenia ze wzmocnieniem. Działa on w środowisku o zadanym skończonym zbiorze stanów S, w których można wykonać jedną operację ze skończonego zbioru stanów A. Działanie algorytmu polega na iteracyjnym modyfikowaniu wartości dwuwyiarowej tablicy $Q(s, a), s \in S, a \in A$. Aby umożliwić naukę konieczne jest zdefiniowane nagród r za wykonanie konkretnych działań, np. nagrody za dotarcie do celu lub kary za każdy wykonany krok. Najpierw inicjalizujemy tablicę Q (gdy nie posiadamy wiedzy o środowisku to inicjalizujemy zerami). Następnie rozpoczyamy proces uczenia. Agent wybiera akcję zgodnie z przyjętą strategią i aktualizuje stan tablicy Q zgodnie ze wzorem:

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

gdzie:

- $Q(s, a)$ — aktualna wartość Q dla stanu s i akcji a,
- α — współczynnik uczenia (learning rate),
- r — nagroda uzyskana po wykonaniu akcji a w stanie s,
- γ — współczynnik dyskontowania (discount factor), który kontroluje wpływ przyszłych nagród,
- $\max_{a'} Q(s', a')$ — maksymalna wartość Q w nowym stanie s'.

1.2 Strategie wyboru akcji

Zaimplementowano trzy strategie wyboru akcji:

1. Epsilon-Greedy

Działa na zasadzie kompromisu między eksploracją a eksploatacją. W tym podejściu, z prawdopodobieństwem ϵ, agent wykonuje akcję losową, a z prawdopodobieństwem $1 - \epsilon$, wybiera akcję o najwyższej wartości Q. Parametr ϵ kontroluje stopień eksploracji, gdzie duża wartość ϵ sprzyja eksploracji, a mała wartość sprzyja eksploatacji.

2. Strategia Boltzmanna

Strategia `Boltzmann` wybiera akcję na podstawie rozkładu prawdopodobieństwa, które jest związane z wartościami Q. Prawdopodobieństwo wyboru akcji a w danym stanie s jest proporcjonalne do eksponencjalnej funkcji wartości $Q(s, a)$, a temperatura T kontroluje stopień eksploracji: przy wysokiej temperaturze akcje są wybierane bardziej losowo, podczas gdy przy niskiej temperaturze akcje z wyższymi wartościami Q mają wyższe prawdopodobieństwo wyboru.

$$P(a) = \frac{e^{Q(s,a)/T}}{\sum_{a'} e^{Q(s,a')/T}}$$

3. Strategia oparta na liczbie odwiedzin (Count-Based)

Strategia `Count-Based` stawia na eksplorację rzadziej odwiedzanych stanów i akcji. Każdy stan-akcja jest liczony, a eksploracja nowych stanów jest premiowana poprzez dodanie tzw. "błąd eksploracji"do wartości Q. Dodatkowa wartość eksploracji zależy od liczby odwiedzin danego stanu-akcji, co zmusza agenta do eksploracji stanów, które były mniej odwiedzane w przeszłości. Funkcja błędu eksploracji może przyjmować różne postacie, w tej implementacji zdecydowano się na funkcję w następującej postaci.

$$\frac{1}{\sqrt{N(s,a)+1}}$$

gdzie $N(s,a)$ to liczba odwiedzin danego stanu s i akcji a. Akcja jest wybierana na podstawie największej wartości z tablicy Q powiększonej o błąd eksploracji.

$$Q(s,a) = Q(s,a) + \frac{1}{\sqrt{N(s,a)+1}}$$

2 Planowane eksperymenty numeryczne

2.1 Cel eksperymentów

Celem eksperymentów było zbadanie wpływu współczynnika uczenia α oraz zaimplementowanych strategi wyboru akcji na działanie alogorytmu Q-learning.

2.2 Parametry eksperymentów

W badaniach przyjęto następująco wartości parametrów:

- współczynnik dyskontowania $\lambda = 0.99$
- w strategi Epsilon-Greedy $\epsilon = 0.1$
- w strategi Boltzmanna $T = 1$

Dla każdego badanego zestawu parametrów przeprowadzono trening składający się z 1000 epizodów. Każdy trening został powtórzony 5-krotnie w celu skompensowania losowości wynikającej z generowania środowiska oraz sposobu wyboru akcji. Prezentowane na wykresach wartości są średnią arytmetyczną z 5 wykonanych prób.

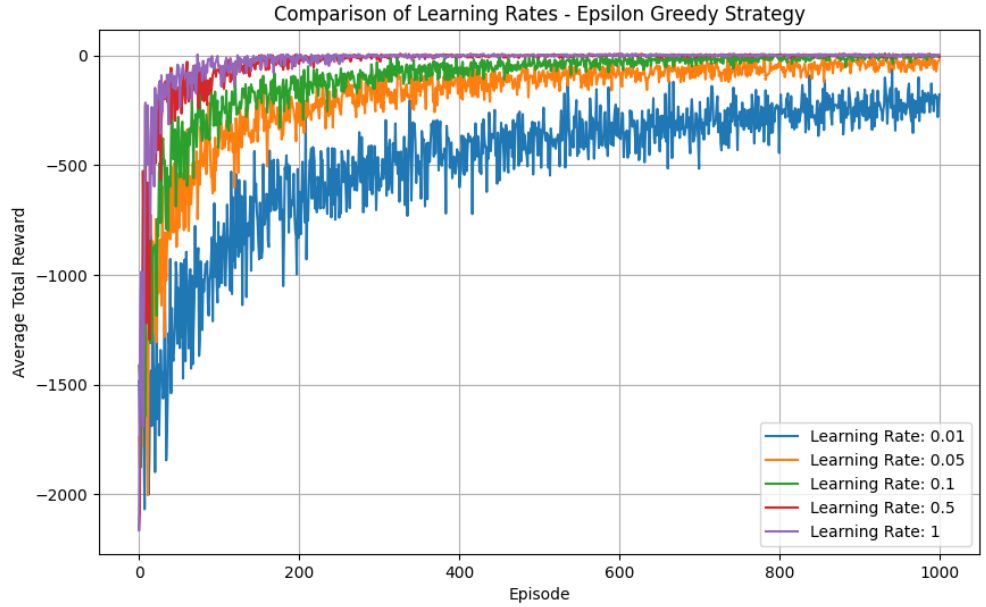
2.3 Środowisko

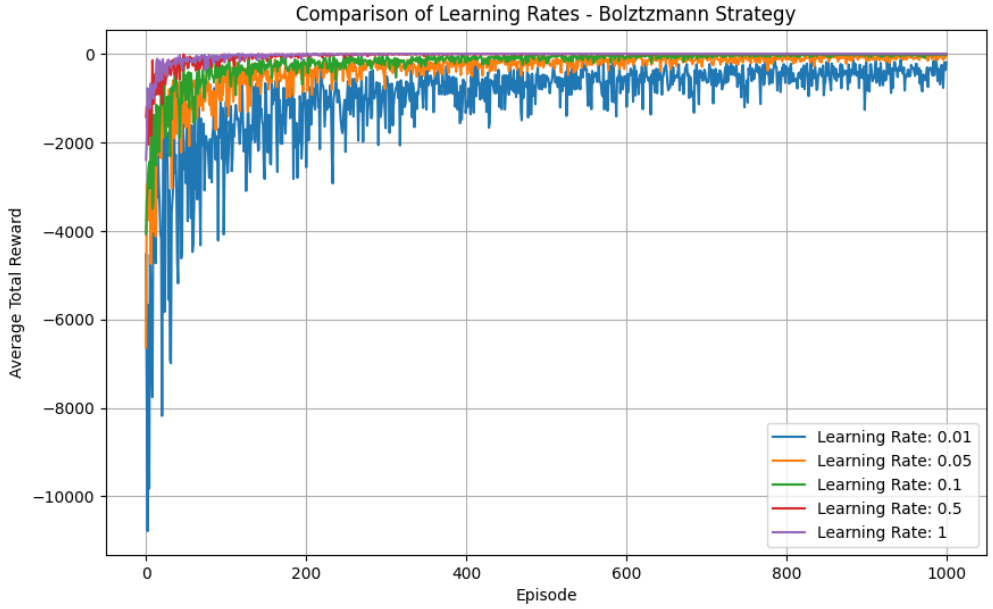
Algorytm został przetestowany w środowisku Taxi-v3 z biblioteki OpenAI Gym, które symuluje problem przewozu pasażera w mieście. Celem agenta jest dostanie się do punktu odbioru pasażera, a następnie dowiezienie go do punktu końcowego. Agent w kazdym kroku wykonuje jedną z sześciu dostępnych akcji: ruch o jedno pole w dowolnym kierunku, odebranie pasażera lub zostawienie pasażera. Za wykonane akcje zdefiniowane są następujące nagrody:

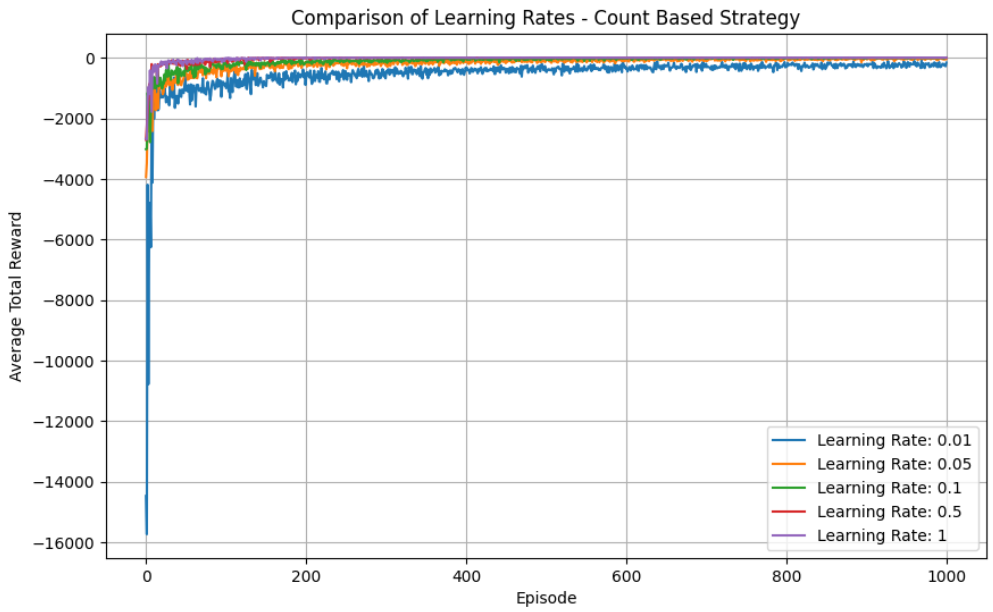
- -1 - każdy wykonany ruch
- -10 - odebranie lub zostawienie pasażera w niewłaściwym miejscu
- +20 - dotarcie do celu

3 Wyniki badań

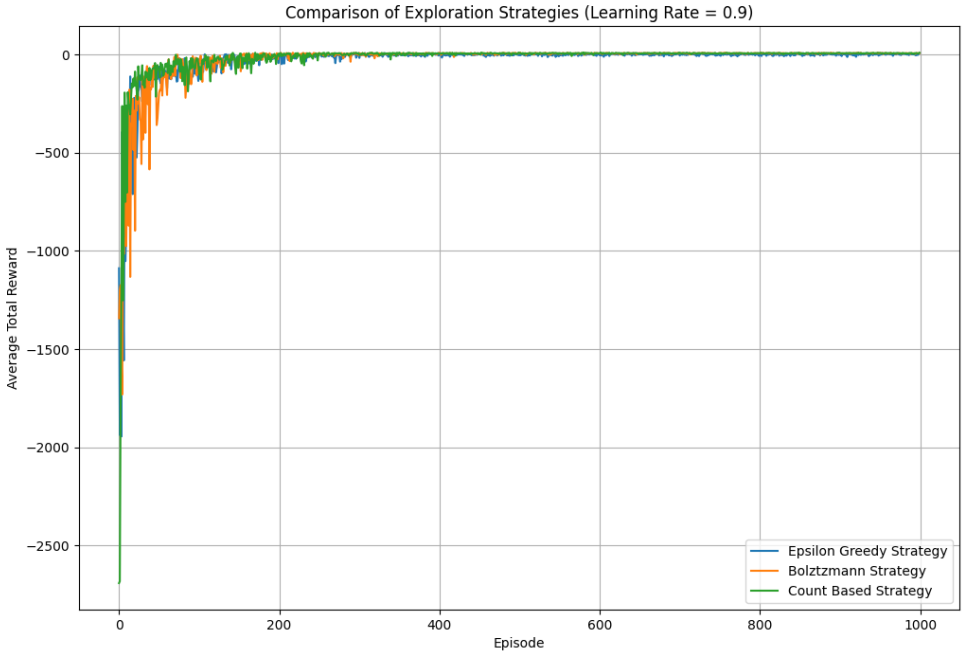
3.1 Wpływ współczynnika uczenia







Dla wszystkich badanych wartości algorytm Q-learning dąży do optimum. Z wykresów wynika, że najlepsze wyniki osiąga dla współczynnika uczenia równego 1. Dla mniejszych wartości, algorytm wolniej zbiega do optimum, a oscylacje są większe. W dalszych badaniach przyjęto $\lambda = 1$.



Z powyższego wykresu wynika, że zachowanie algorytmu dla wszystkich badanych strategi jest niemal identyczne. Jedyną zauważalną różnicą jest wielkość oscylacji, które w początkowej fazie są większe dla strategi Boltzmanna.

4 Wnioski

Z badań wynika, że odpowiedni dobór współczynnika uczenia, może poprawić jakość uzyskanych wyników. Zastosowanie zbyt niskiej wartości nie pozwala osiągnąć zadowalajcych efektów w zadanej liczbie epizodów. Natomiast algorytm dąży do optimum dla szerokiego zakresu wartości tego parametru, co ułatwia strojenie algorytmu. Wpływ współczynnika uczenia jest zbliżony dla

wszystkich zaimplementowanych strategii, co pozwala na łatwe zmienianie strategi w zależności od badanego problemu.

Dla badanego problemu wpływ zastosowanej strategi wyboru nie wykazał znaczącego wpływu na działanie algorytmu. Jedyną widoczną przewagą strategi `Count-based` jest brak konieczności strojenia jej parametrów, co jest jej przewagą w mniej złożonych problemach.