Raport

Alesia Filinkova

336180

1 Treść ćwiczenia

Zaimplementować algorytm Q-learning, a nastepnie użyć go do wytrenowania agenta rozwiazujacego problem Cliff Walking
(https://gymnasium.farama.org/environments/toy$_text/cliff_walking/$).
Stworzyć wizualizacje wyuczonej polityki

2 Doprecyzowanie

- Środowisko CliffWalking-v0 zawiera plansze o wymiarach 4x12, gdzie celem jest dotarcie z pola startowego (w lewym dolnym rogu) do celu (w prawym dolnym rogu) bez upadku z klifu.
- Użyto tablicy Q table, aby przechowywać wartość oczekiwanych nagród dla każdej kombinacji stanu i akcji.
- Metryka oceny jest suma nagród uzyskanych podczas treningu oraz wizualizacja wyuczonej polityki.

3 Cel i opis eksperymentów

3.1 Cel

Celem ćwiczenia było:

1. Zaimplementowanie algorytmu Q-learning.
2. Wytrenowanie agenta, który skutecznie nauczy sie omijać klif i osiagnie stan końcowy z minimalnym kosztem.
3. Wizualizacja wyników i analiza zachowania agenta.

3.2 Opis eksperymentów:

- Zbiór danych: Plansza CliffWalking (4x12), dostepna w gymnasium.
- Algorytm: Q-learning z tablica q table
- Metryka: Wizualizacja polityki, dystrybucje odwiedzanych stanów i wykonanych akcji
- Parametry:
  - Współczynnik uczenia (learning rate)
  - Współczynnik dyskontowania (discount factor)
  - Prawdopodobieństwo eksploracji (epsilon)
  - Liczba epizodów treningowych (num episodes)

4 Przygotowanie środowiska i danych

Skrypt można uruchomić przez terminal za pomoca polecenia:

1. git clone https://gitlab-stud.elka.pw.edu.pl/afilinko/wsi.git
2. python3 -m venv venv
3. source venv/bin/activate
4. cd /lab6
5. pip install -r requirements.txt
6. `python3 main.py`

5 Wyniki

Eksperymentalnie ustawiono, że najlepsze wyniki daja nastepujace wartości hiperparametrów: learning$_rate = 0,15, gamma = 0,95, epsilon = 0,1, episodes =$ 500.

Można ustawić wieksza liczbe iteracji, jednak to nie bardzo polepsza wynik, skoro cały algorytm jest oparty o dość losowe zdarzenia.

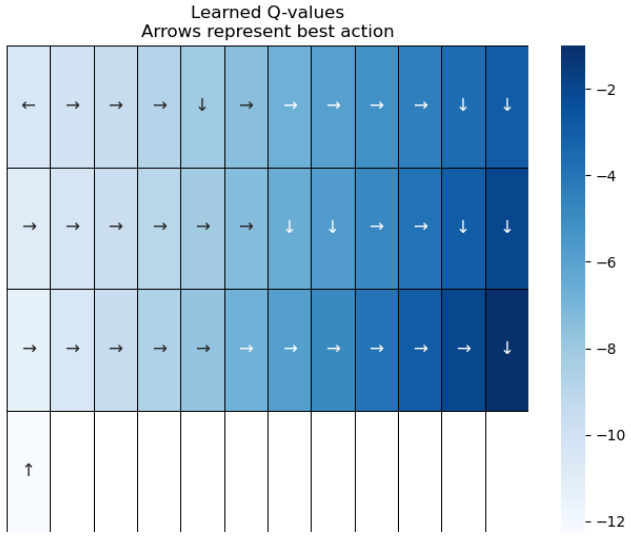


Figure 1: Wizualizacja wyuczonej polityki

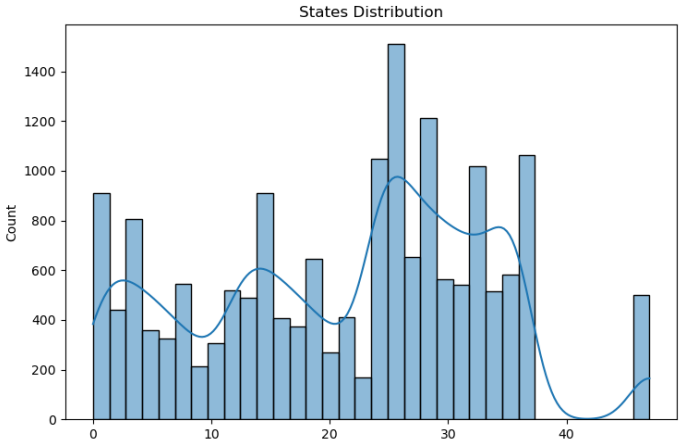
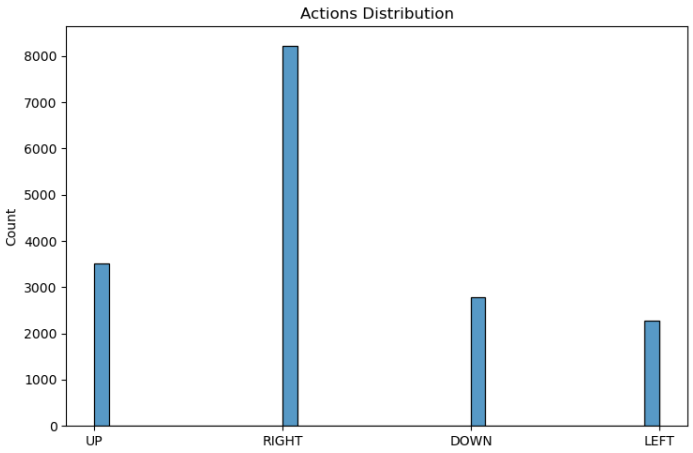


Figure 2: Ilość kroków



6 Wnioski

- Agent skutecznie nauczył sie omijać klif, poruszajac sie wzdłuż bezpiecznych pól, co widać po spójnej polityce ($\rightarrow$ w pierwszym rzedzie, a nastepnie $\downarrow$ w celu dotarcia do końca).
- Liczba odwiedzanych stanów i akcji wskazuje, że agent w wiekszości epizodów trzymał sie bezpiecznych pól
- Wizualizacja polityki pokazuje, że agent preferuje ruchy w kierunku celu, a rozkłady stanów i akcji wskazuja na efektywne eksplorowanie środowiska.
- W przyszłości można by zastosować inne techniki uczenia, takie jak Deep Q-Learning, aby zbadać ich wpływ na wyniki.