Zadanie 6

Q-Learning

Kacper Kania

28.04.2024

Treść zadania

Q-Learning jest podstawowym algorytmem uczenia ze wzmocnieniem. Jest to metoda uczenia bez nadzoru, która polega na uczeniu się strategii zachowań w środowisku, aby osiągnąć maksymalną nagrodę. W algorytmie Q-Learningu agent uczy się funkcji wartości akcji, która określa wartość każdej akcji w każdym stanie. W każdym kroku agent wybiera akcję, która maksymalizuje wartość funkcji wartości akcji. Wartość funkcji wartości akcji jest aktualizowana na podstawie nagrody otrzymanej po wykonaniu akcji i wartości funkcji wartości akcji w nowym stanie. Państwo zrealizujecie funkcję Q jako tablicę, natomiast Q może być dowolną funkcją, nawet wyuczalną. Na przykład, jedna ze słynniejszych prac DeepMindu wykorzystała głębokie sieci neuronowe do nauczenia funkcji Q z pikseli obrazu gry [1].

Państwa zadaniem jest implementacja uczenia funkcji Q jako tablicy w środowisku **FrozenLake-v1**[1] z biblioteki **gym** z ustawieniami:

```
gym.make('FrozenLake-v1', desc=None, map_name="8x8", is_slippery=True)
```

Sama biblioteka zwraca dla podanej akcji obserwację (zmianę stanu środowiska), nagrodę oraz informację o zakończeniu epizodu.

```
import gym
env = gym.make("LunarLander-v2", render_mode="human")
observation, info = env.reset(seed=42)
for _ in range(1000):
   action = policy(observation)  # User-defined policy function
   observation, reward, terminated, truncated, info = env.step(action)

   if terminated or truncated:
      observation, info = env.reset()
env.close()
```

Dla przypomnienia jeden krok uczenia Q wygląda następująco:

$$Q^{\text{new}}(S_t, A_t) \leftarrow \left(1 - \underbrace{\alpha}_{\text{learning rate}}\right) \cdot Q(S_t, A_t) + \alpha \cdot \underbrace{\left(\underbrace{R_{t+1}}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\pi_a(Q(S_{t+1}, a))}_{\text{Q value in next state after selecting action according to the policy}}\right)}_{\text{new value}}, \tag{1}$$

gdzie:

[1] https://www.gymlibrary.dev/environments/toy_text/frozen_lake/

- α to współczynnik uczenia,
- R_{t+1} to nagroda otrzymana po wykonaniu akcji A_t w stanie S_t,
- γ to współczynnik dyskontowania,
- $Q(\cdot, \cdot)$ nasza funkcja do wyuczenia,
- $\pi_{a(\cdot)}$ zwraca akcję a w zależności od określonej polityki wyboru i wartości Q (np. polityka ϵ-zachłanna).

W sprawozdaniu należy:
- zbadać wpływu parametru uczenia α na zbieżność algorytmu,
- zbadać wpływ parametru dyskontowania γ na zbieżność algorytmu,
- zbadać wpływ parametru eksploracji ϵ na zbieżność algorytmu w podejsciu ϵ-zachłannym, oraz parametr T w strategii opartej na rozkładzie Boltzmanna.

Zbieżność algorytmu proszę przedstawic jako wykres zależności końcowej nagrody od kroku uczenia funkcji Q. W tabelach natomiast proszę przestawić średnią nagrodę w ostatnich 10 krokach uczenia wraz z odchyleniem standardowym oraz liczbę rozgrywek zakończonych sukcesem (10 rozgrywek). Proszę przedstawić wyniki dla 5 różnych random seedów.

Bibliografia

[1] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.