**Dokumentacja WSI do zadania 6: Analiza Algorytmu Q-Learning w Środowisku Taxi**

**Autor: Artem Kukushkin, 317140**

**Wstęp**

Celem projektu była implementacja algorytmu Q-learning oraz analiza wpływu hiperparametrów (współczynnika uczenia α) i strategii eksploracji na jego działanie w środowisku Taxi-v3 z biblioteki Gymnasium. Środowisko Taxi-v3 symuluje problem taksówki, której zadaniem jest przewiezienie pasażera do celu, minimalizując kary czasowe i unikając nielegalnych akcji.

**Opis Algorytmu i Implementacji**

Algorytm Q-learning to metoda uczenia ze wzmocnieniem, która aktualizuje wartości Q według wzoru:

Q(s, a) ← Q(s, a) + α [ r + γ max(Q(s', a')) - Q(s, a) ],

gdzie:

s: obecny stan,

a: wybrana akcja,

r: nagroda,

s': kolejny stan,

α: współczynnik uczenia,

γ: współczynnik dyskontowania.

Implementacja w Pythonie wykorzystuje bibliotekę Gymnasium i obejmuje:

Tablicę Q zainicjalizowaną zerami dla 500 stanów i 6 akcji.

Dwie strategie eksploracji: ε-greedy oraz softmax z różnymi temperaturami.

Trening przez 1000 epizodów z malejącym ε.

Kompatybilność z Pyodide dzięki asynchronicznym pętlom zdarzeń.

Główne funkcje:

choose\_action: Wybiera akcję na podstawie strategii eksploracji (ε-greedy lub softmax).

train\_q\_learning: Trenuje agenta, aktualizując tablicę Q.

test\_agent: Ocenia wytrenowaną politykę na 100 epizodach testowych.

run\_experiments: Przeprowadza eksperymenty dla różnych wartości α i strategii eksploracji.

**Eksperymenty**

Hiperparametry

Liczba epizodów: 1000.

Współczynnik dyskontowania (γ): 0.99.

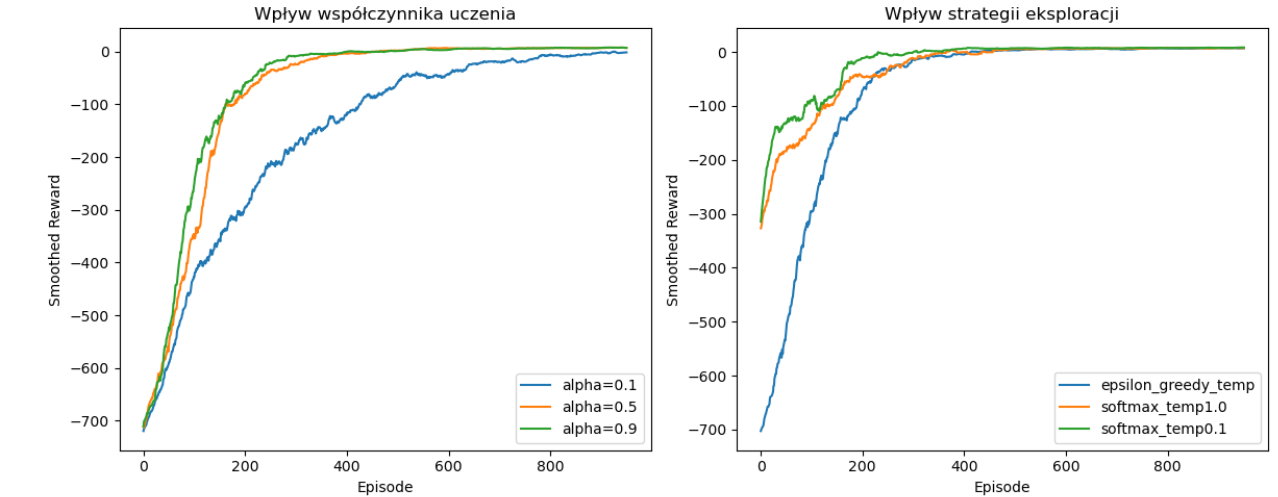
Początkowe ε: 1.0, końcowe ε: 0.01, dekrementacja ε: 0.995.

Wartości α: 0.1, 0.5, 0.9.

Strategie eksploracji: ε-greedy, softmax (temperatura 1.0 i 0.1).

**Wyniki**

Wyniki eksperymentów przedstawiono w analizie statystycznej opartej na testach wytrenowanej polityki.



**Wpływ współczynnika uczenia (α):**

α = 0.1: Średnia nagroda = -103.68, Odchylenie = 104.37.

α = 0.5: Średnia nagroda = 1.27, Odchylenie = 35.48.

α = 0.9: Średnia nagroda = 3.82, Odchylenie = 29.21.

**Wpływ strategii eksploracji:**

ε-greedy: Średnia nagroda = 1.51, Odchylenie = 35.52.

Softmax, temperatura = 1.0: Średnia nagroda = -7.21, Odchylenie = 52.98.

Softmax, temperatura = 0.1: Średnia nagroda = 1.40, Odchylenie = 35.51.

**Analiza Wyników**

**Wpływ współczynnika uczenia (α)**

α = 0.1 prowadzi do bardzo niskich nagród (-103.68) z dużym odchyleniem (104.37), co wskazuje na wolne uczenie i dużą zmienność wyników, prawdopodobnie z powodu zbyt powolnych aktualizacji tablicy Q.

α = 0.5 osiąga dodatnią średnią nagrodę (1.27) z umiarkowanym odchyleniem (35.48), co sugeruje skuteczne uczenie i stabilną politykę.

α = 0.9 daje najwyższą średnią nagrodę (3.82) z najmniejszym odchyleniem (29.21), co wskazuje na szybkie i stabilne uczenie, choć może występować ryzyko nadmiernych oscylacji w tablicy Q.

**Wpływ strategii eksploracji**

ε-greedy osiąga najwyższą średnią nagrodę (1.51) z umiarkowanym odchyleniem (35.52), co potwierdza jej skuteczność w prostym środowisku Taxi-v3 dzięki prostemu mechanizmowi eksploracji.

Softmax z temperaturą 1.0 daje ujemną nagrodę (-7.21) z największym odchyleniem (52.98), co wynika z nadmiernej losowości w wyborze akcji.

Softmax z temperaturą 0.1 osiąga dodatnią nagrodę (1.40) z odchyleniem (35.51) zbliżonym do ε-greedy, wskazując na dobrą równowagę między eksploracją a eksploatacją przy niskiej temperaturze.

**Wnioski**

Wartość α = 0.9 jest optymalna w środowisku Taxi-v3, zapewniając najwyższe nagrody i najlepszą stabilność.

Strategia ε-greedy pozostaje najskuteczniejsza, oferując najwyższe nagrody i dobrą stabilność.

Softmax z temperaturą 0.1 jest dobrą alternatywą, zbliżając się do wyników ε-greedy, podczas gdy wysoka temperatura (1.0) negatywnie wpływa na wyniki.