Wybrane zagadnienia sztucznej inteligencji
laboratorium

Ćwiczenie 3
Uczenie ze wzmocnieniem i nadzorowane w grach

opracowanie: J. Jakubik

Opis ćwiczenia

W ćwiczeniu zapoznamy się z problematyką uczenia ze wzmocnieniem, w szczególności algorytmem Q-learning oraz możliwym zastosowaniem metod uczenia nadzorowanego w grach

Uczenie ze wzmocnieniem oznacza typ uczenia maszynowego, w którym agent SI, na podstawie doświadczeń uczy się, jak kierować własnymi akcjami w pewnym środowisku tak aby maksymalizować otrzymaną nagrodę.

Celem ćwiczenia jest zapoznanie się z koncepcją uczenia ze wzmocnieniem, oraz jej podstawowymi ograniczeniami.

Uczenie ze wzmocnieniem

Uczenie ze wzmocnieniem typowo zajmuje się sytuacją w której agent SI świadomy swojego *stanu* ma wybrać najlepszą spośród zbioru możliwych *akcji* tak aby zmaksymalizować pewną *nagrodę.* Akcja może zaowocować zmianą stanu, która nie musi jednak być deterministyczna. Jeżeli przejścia między stanami są probabilistyczne, ale zależne wyłącznie od obecnego stanu i podjętej akcji (a nie od historii), mówimy o *procesie decyzyjnym Markowa*.

Uczenie ze wzmocnieniem jest interesujące ze względu na analogie do rzeczywistych procesów kognitywnych (w rzeczywistości uczymy się w oparciu o nagrody i kary wynikające z naszych poczynań), ale ma również zastosowanie w grach.

Zwróćmy uwagę na problem gier dwuosobowych w sytuacji, gdy nieznana jest logika kierująca przeciwnikiem. Jeżeli uznamy przeciwnika za część otoczenia, mamy *stan* gry, problem wyboru *akcji* oraz ewentualną *nagrodę* w postaci przegranej lub wygranej. Decyzję przeciwnika możemy uznać za probabilistyczny element modelu – agent nie wie dokładnie do jakiego stanu przejdzie po podjęciu decyzji.

Q-Learning

Algorytm Q-Learning jest sposobem wyznaczenia funkcji Q(s,a) oceniającej jakość pewnej akcji w obecnym stanie agenta SI. W najprostszym przypadku wartość Q możemy przechowywać jako tabelę stan-akcja i aktualizować konkretne pole w każdej iteracji. Aktualizacje wartości Q odbywają się jednocześnie z wykonywaniem akcji i obserwacji przez agenta, według reguły:

$$Q(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

https://en.wikipedia.org/wiki/Q-learning

Modelowanie oczekiwanej nagrody

Podstawowa definicja Q-learningu napotyka problemy w sytuacji, gdy liczba możliwych stanów zaczyna być problemem wydajnościowym. W takiej sytuacji warto sformułować model zdolny do generalizowania wniosków z pewnych doświadczeń odnośnie stanów obserwowanych na stany podobne.

Model taki można zbudować wykorzystując metody uczenia nadzorowanego. Jeżeli opiszemy stan w postaci wektora, możemy ocenić go na podstawie doświadczeń wykorzystując dowolną metodę regresji, wyuczonej na podstawie par stan gry – zmienna binarna (wygrana/przegrana, oceniona po końcowym wyniku gry). Problemem w takiej sytuacji jest zbudowanie odpowiedniego zbioru uczącego.

Zadania

Student ma możliwość wyboru między dwoma wariantami zadania.

Wariant pierwszy: implementacja Q-learningu oraz rozwiązania opartego o uczenie nadzorowanego w grze kółko i krzyżyk. W tym przypadku konieczna jest również implementacja gry.

Wariant drugi: implementacja rozwiązania opartego o uczenie nadzorowane w grze Hearthstone, wykorzystująca kod opracowany w ramach zadania pierwszego. W takim wypadku implementacja Q-learningu nie jest wymagana.

Przez rozwiązanie oparte o uczenie nadzorowane rozumiemy tutaj model, który na podstawie rozegranych gier i samego wyniku wygrana/przegrana ocenia stany gry. Przedmiotem badań w sprawozdaniu powinna być strategia budowania zbioru uczącego dla takiej metody.

Metody uczenia nadzorowanego powinny wykorzystywać gotowe biblioteki, np. scikit-learn, keras etc. Q-learning powinien być zaimplementowany własnoręcznie.

Harmonogram oczekiwanych postępów

Wariant 1:

Tydzień 1: Wprowadzenie do ćwiczenia
Tydzień 2: Implementacja Q-learningu
Tydzień 3: Propozycja rozwiązania opartego o uczenie nadzorowane
Tydzień 4: Implementacja rozwiązania opartego o uczenie nadzorowane
Tydzień 5: Oddanie sprawozdania z wynikami badań

Wariant 2:

Tydzień 1: Wprowadzenie do ćwiczenia
Tydzień 2: Propozycja rozwiązania opartego o uczenie nadzorowane
Tydzień 3: Implementacja rozwiązania opartego o uczenie nadzorowane
Tydzień 4: Integracja rozwiązania z MCTS
Tydzień 5: Oddanie sprawozdania z wynikami badań

Za brak systematyczności w tygodniach 2–4 można stracić po punkcie za każdy.

Ocena ćwiczenia

Wariant 1:

2pkt	Implementacja gry
2pkt	Implementacja Q-learningu
2pkt	Propozycja i implementacja rozwiązania opartego o uczenie nadzorowane
2pkt	Ocena zaproponowanego rozwiązania
2pkt	Punkty za kreatywność

Wariant 2:

2pkt	Propozycja i implementacja rozwiązania opartego o uczenie nadzorowane
2pkt	Badanie skuteczności uczenie nadzorowanego
2pkt	Integracja rozwiązania z MCTS
2pkt	Ocena zaproponowanego rozwiązania
2pkt	Punkty za kreatywność