Zaawansowane Metody Inteligencji Obliczeniowej

Wykład 1: Agent i środowisko

Michał Kempka Marek Wydmuch Bartosz Wieloch

27 lutego 2023

Plan wykładu

Ankieta „demograficzna”

https://tinyurl.com/zmio2023start

Plan wykładu

Prowadzący

mgr inż. Michał Kempka
<mkempka@cs.put.poznan.pl>

mgr inż. Marek Wydmuch
<mwydmuch@cs.put.poznan.pl>

O czym jest ten przedmiot?

Ogólnie, jednym zdaniem:

O tym jak podejmować **dobre sekwencje decyzji**.

A przez większość czasu:

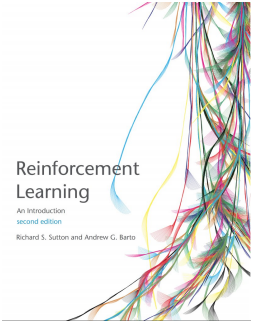
Jak uczyć się podejmować dobre sekwencje decyzji,
gdy nie wiemy jak działa świat.

O czym jest ten przedmiot?

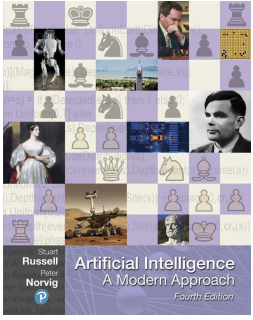
Fundamentalny problem dla sztucznej inteligencji:

Jak uczyć się podejmować dobre sekwencję decyzji **w obliczu niepewności**.

Literatura



„Reinforcement Learning: An Introduction"
Richard S. Sutton and Andrew G. Barto, 2018
(http://incompleteideas.net/book/the-book.html)



„Artificial Intelligence: A Modern Approach"
Stuart J. Russell and Peter Norvig, 2010
(http://aima.cs.berkeley.edu/)

+ wybrane publikacje do ostatnich wykładów.

Przybliżony plan tematów

1. Agent i środowisko
2. Wprowadzenie do uczenie ze wzmocnieniem i wieloręki bandyta
3. Proces Decyzyjny Markowa (MDP) i Programowanie Dynamiczne (DP)
4. Metoda Monte Carlo (MC) i Temporal-Difference Learning (TDL)
5. Optymalizacja metodą spadku wzdłuż gradientu (przypomnienie) i metody aproksymacyjne (przypomnienie?)
6. Ciągła przestrzeń stanów, (Deep) Q-Learning (DQN) i rozszerzenia
7. Problemy z ciągła przestrzenią akcji, Policy Gradient
8. Actor Critic (AC) i Deep Deterministic Policy Gradient (DDPG)
9. Planowanie i Monte Carlo Tree-Search (MCTS)
10. TD-Gammon, Deep Blue, AlphaGo, AlphaZero
11. Wnioskowanie probabilistyczne i sieci Baysowskie
12. State of the art: Proximal Policy Optimization (PPO), World Models i Dreamer oraz Reinforcement Learning from Human Feedback (RLHF) i ChatGPT

Wymagania

- Znajomość języka programowania Python, umiejętność obsługi Jupyter Notebooków/JupyterLab oraz znajomość biblioteki PyTorch będzie pomocna.
- Podstawy z rachunku prawdopodobieństwa, statystyki, rachunku różniczkowego.
- Znajomość optymalizacji metodą spadku wzdłuż gradientu będzie pomocna.
- Wstępna wiedza o uczeniu maszynowym i technikach przeszukiwania będzie pomocna.

Zajęcia i zasady zaliczenia

- Wykłady z elementami interaktywnymi. Ocena końcowa: z kolokwium + bonusy za aktywność.
- Laboratoria: ćwiczenia i zadania programistyczne zakładające znajomość wykładu. Ocena końcowa: mini-projekty - od 3 do 5, różna punktacja + bonusy za aktywność, standardowe przedziały procentowe dla ocen (szczegóły na zajęciach laboratoryjnych).
- Jakiekolwiek wykryte próby oszustwa skutkują niepoprawialną 2. dla wszystkich zamieszanych osób.

Plan wykładu

Agent i środowisko

W ogólnym schemacie zadań które będziemy chcieli rozwiązywać możemy wydzielić dwa główne elementy:

- **Środowisko** – definiuje problem który rozwiązujemy (na zasady jego działania nie mamy wpływu). Przykłady:
  - fizyczne, np. hala produkcyjna czy powierzchnia Marsa po której przemieszcza się robot (działa zgodnie z mechaniką świata)
  - plansza do gry wraz z zasadami gry lub gra komputerowa.
- **Agent** – (od łac. agere, działać/czynić) ktoś lub coś co działa – **wykonuje akcje**. Przede wszystkim mamy tutaj na myśli działanie autonomiczne w celu realizowania jakiegoś celu. Przykłady:
  - robot który fizycznie przemieszcza się na hali i przewozi paczki,
  - program grający w szachy.

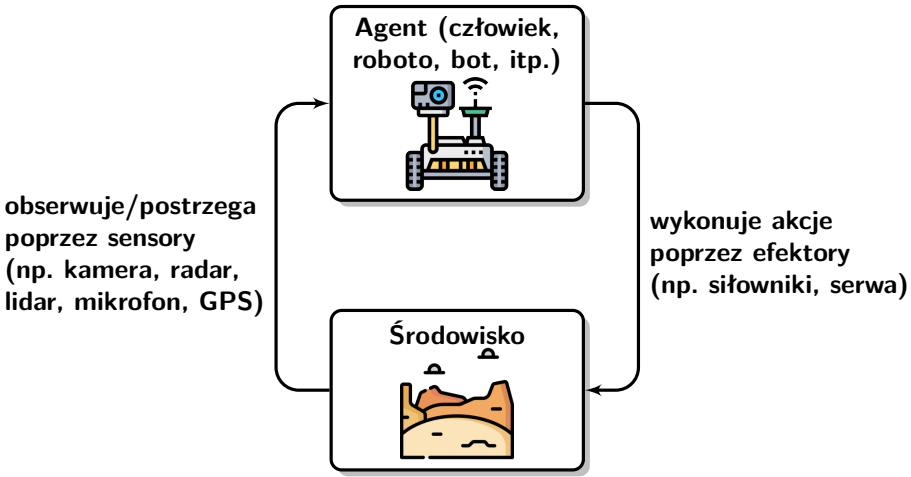
Agent i środowisko

Schemat działania:

- Środowisko jest aktualnie w jakimś stanie (**stan** środowiska)
- Agent pomocą dostępnych sensorów postrzega ten stan (**obserwacja**)
- Agent na podstawie obserwacji, swojej wiedzy, sposobu wnioskowania itd. podejmuje działanie (**akcję**)
- Środowisko potencjalnie zmienia swój stan. Akcja podjęta przez agenta może (ale nie musi) wpłynąć na zmianę stanu środowiska.

Kroki powtarzane się aż do spełnienia warunku stopu (np. śmierci agenta).

Agent i środowisko



Rysunek: Schemat modelu agenta działającego w środowisku.

Przykłady

Robot rozwożący paczki w magazynie

- środowisko: magazyn
  - ściany, regały, alejki, ...
  - paczki do przewiezienia
  - reguły mówiące co jak się zachowuje (prawa fizyki, jak zaprogramowane są maszyny itd.)
- Agent: robot
  - sensory: kamery, lidary, czujniki dotyku, waga, itd.
  - efektory: kółka, chwytak, itd.

Przykłady

Program grający w szachy

- środowisko: gra
  - wirtualna plansza, pionki/figury
  - reguły gry w szachy
- Agent: program komputerowy
  - sensory: zna bezpośrednio cały stan planszy (lub obraz itp.)
  - akcje: wykonuje ruch figurą

Przykłady

Program inwestujący na giełdzie

- środowisko: giełda
  - reguły rynku, sprzedaży i kupna akcji
- Agent: program komputerowy
  - sensory: aktualne ceny kupna i sprzedaży, informacje
  - akcje: decyzje o zakupie lub sprzedaży konkretnej ilości wybranych akcji

Funkcja agenta

- Agent w trakcie swojego życia (całego czasu działania) widzi kolejne stany środowiska przez pryzmat swoich sensor (czyli obserwacje), pojedynczy **zaobserwowany** stan będziemy oznaczać jako S.
- Jego działanie (podejmowane akcje) opisuje **funkcja agenta**:

$$f : \mathcal{H}^* \to \mathcal{A},$$

  gdzie $\mathcal{H}^*$ to przestrzeń historii obserwacji, a $\mathcal{A}$ to przestrzeń akcji.
- Funkcja agenta to abstrakcyjny matematyczny opis agenta
- Funkcja agenta musi być technicznie zaimplementowana (**program agenta**), np.:
  - poprzez stabularyzowana wersja funkcji agenta (na zasadzie zdefiniowania zasady „jeśli jakieś $\{S_1, S_2 \dots S_t\} = h$ to wykonaj akcję a" dla wszystkich $h \in \mathcal{H}$)
  - zbiór reguł, itp.

Plan wykładu

Racjonalność

- Nasz cel: budowanie „inteligentnych agentów".
- Przyjmujemy pogląd, że inteligencja jest równoznaczna z racjonalnością/racjonalnym działaniem.
- **Racjonalność – działanie w celu osiągnięcie jak najlepszego rezultatu lub w wypadku niepewności najlepszego oczekiwanego rezultatu.**

Racjonalność

Agent **racjonalny** to **NIE** to samo co agent **perfekcyjny**:

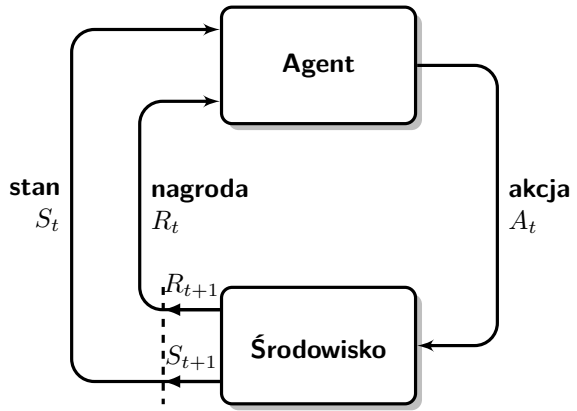
- perfekcja wymaga wiedzy co się stanie **po** wykonaniu akcji
- agent racjonalny nie musi być wszechwiedzący
- racjonalność często wymaga od agenta aktywnego zdobywania informacji o środowisku i uczenia się

Przykład: chodzenie – czy warto:

- patrzeć pod nogi?
- patrzeć w górę?

Uczenie ze wzmocnieniem

- **Uczenia ze wzmocnieniem (ang. reinforcement learning (RL))** – agent wykonuje **akcję (ang. action)**, obserwuje nowy **stan (ang. state)** środowiska i otrzymuje **nagrodę (ang. reward)**, która formalizuje cel agenta.



Rysunek: Schemat modelu uczenia ze wzmocnieniem.

Uczenie ze wzmocnieniem

- W kontekście uczenia ze wzmocnieniem racjonalny agent **zawsze** wybiera akcję, która **maksymalizuje** oczekiwaną wartość sumy wszystkich otrzymanych nagród, biorąc pod uwagę **aktualną wiedzę** agenta.
  - **Zawsze** – dla każdej możliwej historii obserwacji.
  - **Wiedza** – a priori (np. model środowiska) + historia wszystkich dotychczas zaobserwowanych stanów.
- Suma wszystkich otrzymanych nagród jest nazywana **całkowitą nagrodą (ang. total reward)** lub w bardziej ogólnie ujęciu **miarą jakości**.

Racjonalność w uczenie ze wzmocnieniem

Podsumowując racjonalność w uczeniu ze wzmocnieniem zależy od:

- **miary jakości**, która definiuje kryterium sukcesu,
- **wiedzy a priori** agenta o środowisku
- **akcji** które agent wykonuje,
- **historii obserwacji** agenta (zdobytego doświadczenia).

Racjonalność – przykłady

Pytanie

Czy cotygodniowe kupowanie losów Lotto jest racjonalne?

Pytanie

Czy ubezpieczanie samochodu od kradzieży jest racjonalne?

Plan wykładu

Obserwowalność środowiska

- **Całkowicie obserwowalne** – stan całego środowiska znany w każdym momencie (np. szachy), w konsekwencji agent nie musi pamiętać poprzednio zaobserwowanych stanów.
- **Częściowo obserwowalne** – stan zawierający tylko część informacji lub zaszumione/niedokładne informacje (np. poker).
- **Nieobserwowalne** – całkowity brak informacji, agent otrzymuje wyłącznie nagrodę (np. jednoręcy bandyci).

Przykłady – obserwowalność

Pytanie

Pasjans – obserwowalne?

Pytanie

Chińczyk (gra planszowa) – obserwowalne?

Liczność agentów w środowisku

- **Jednoagentowe** – np. gra w *Breakout*.
- **Wieloagentowe** – mogą być:
  - kompetytywne (racjonalność może wymagać losowości, np. poker)
  - kooperacyjne lub mieszane (racjonalność może wymagać komunikacji, np. ruch drogowy)

Przykłady – wieloagentowe?

Pytanie

Pasjans – wieloagentowe?

Pytanie

Chińczyk (gra planszowa) – wieloagentowe?

Determinizm środowiska

- **Deterministyczne** – dana akcja w danym stanie będzie skutkować obserwacją zawsze tego samego, następnego stanu (np. balansowanie).
- **Stochastyczne** – dana akcja w danym stanie może skutkować różnymi stanami i/lub nagrodami, zgodnie ze znanym lub nieznanym rozkładem prawdopodobieństwa (np. gra w ruletkę).

Uwagi:

- Agent nie jest w stanie przewidzieć następnego stanu gdy środowisko nie jest całkowicie obserwowalne lub gdy nie jest deterministyczne.
- Częściowa obserwowalność może wyglądać jakby środowisko nie było deterministyczne.

Przykłady – determinizm

Pytanie

Pasjans – deterministyczne?

Pytanie

Chińczyk (gra planszowa) – deterministyczne?

Statyczne i dynamiczne środowiska

- **Statyczne** – środowisko ‘czeka’ na akcję (np. szachy, gry turowe).
- **Dynamiczne** – środowisko zmienia się podczas gdy agent decyduje o kolejnej akcji (np. ruch drogowy, regulowanie temperatury).
- **Semidynamiczne** – środowisko się nie zmienia, ale czas podejmowania decyzji wpływa na otrzymaną nagrodę (np. teleturniej).

Przykłady – statyczne?

Pytanie

Autonomiczny samochód – statyczne?

Pytanie

Kontrola jakości na linii produkcyjnej – statyczne?

Ciągłość środowiska

- **Dyskretne** – akcje i stany są skończonymi zbiorami np. szachy. Warto zauważyć, że nagrodę uznaje się zwykle za niedyskretną i pomija w rozważaniach ciągłości środowisk.
- **Ciągłe** – akcje i/lub stany są zmiennymi ciągłymi (liczbami rzeczywiste) np. ruch drogowy.

Uwaga:

- Wiele środowisk łączy w sobie elementy ciągłe i dyskretne. Przykładem takiego środowiska może być jazda samochodem (np. stan diod, zmiana biegu), a część ciągła (np. prędkość, prędkość skręcania kierownicą)

Przykłady – dyskretne?

Pytanie

Autonomiczny samochód – dyskretne?

Pytanie

Kontrola jakości na linii produkcyjnej – dyskretne?

Model środowiska

Nie jest to cecha środowiska, lecz stan wiedzy agenta (lub projektanta algorytmu), lecz z praktycznego punktu widzenia możemy tę wiedzę uznać za cechę środowiska.

- **Znany model środowiska** – konsekwencje akcji lub ich rozkłady prawdopodobieństw (gdy środowisko jest stochastyczne) jest znany.
- **Nieznany model środowiska**
  - konsekwencje akcji nie są znane
  - racjonalność wymaga uczenia się (poznania środowiska)
  - cecha nie ma związku z tym czy środowisko jest całkowicie lub częściowo obserwowalne (układanie pasjansa vs. pisanie na maszynie z poprzestawianymi klawiszami)

Przykłady – znany model?

Pytanie

Chińczyk – znany model?

Pytanie

Autonomiczny samochód – znany model?

Plan wykładu

Agent odruchowy (ang. reflex agent)

Odpowiada bezpośrednio na zaobserwowane stan, ignorując poprzednio zaobserwowane stany.

- akcja zależy tylko od **aktualnej obserwacji**
- ignoruje historię obserwacji (nie potrzebuje pamięci)
- **może być racjonalny**

Agent z modelem/pamięcią (ang. model-based agent)

Utrzymuje wewnętrzny stan aktualizowany po każdej nowej obserwacji i podejmuje decyzje na jego podstawie. Czyli śledzi informacje na temat aspektów środowiska, które nie są zawsze obserwowalne.

- aktualizuje swoją wiedzę o stanie świata po każdej obserwacji.
- wiedza ta może być niepewna (np. wnioskowanie o nieobserwowanej części środowiska)
- akcje podejmuje na podstawie swojej **aktualnej wiedzy** (w przeciwieństwie do aktualnej obserwacji)

Agent celowy (ang. goal-based agent)

Agent, który działa by osiągnąć jakiś cel. Zazwyczaj cel jest dyskretny (konkretny stan).

- agent zna cel który chce osiągnąć
- potrafi ocenić konsekwencje swoich akcji
- akcje zazwyczaj podejmuje na podstawie **planowanie/przeszukiwania**
- planowanie wymaga by model środowiska był znany

Agent z funkcją użyteczności (ang. utility-based agent)

- agent posiada funkcję użyteczności, za pomocą której ocenia użyteczność stanów i akcji
- wykonuje akcje na podstawie użyteczności akcji w aktualnym
- agent jest racjonalny jeśli funkcja użyteczności jest zgodna z miarą jakości

Zalety:

- kompromis między wieloma celami
- działanie na podstawie oczekiwanej użyteczności (gdy cele osiągalne stochastycznie)

Plan wykładu

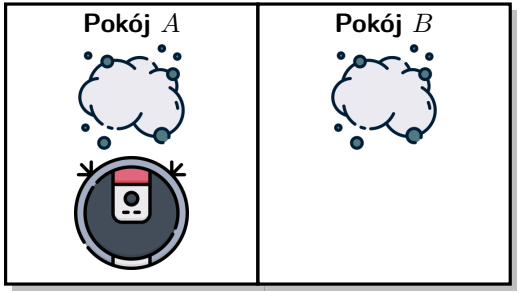
Świat odkurzacza

Środowisko (pokoje do sprzątania):

- dwa pokoje: A (z lewej) i B (z prawej)
- w każdym pokoju jest albo *czysto* albo *brudno*
- robot zaczyna w losowym pokoju

Agent (robot odkurzacz):

- obserwacje to para: [lokalizacja robota, stan pokoju]
- robot może wykonać jedną z akcji: {(idź w) lewo, prawo, odkurz, czekaj}



Rysunek: Świat odkurzacza robota z dwoma pokojami

Pytanie

Ile jest możliwych różnych stanów tego świata?

Świat odkurzacza – funkcja agenta

Przykładowa funkcja agenta (postać stabularyzowana):

Historia obserwacji	Akcja
$[A,$ *czysto*$]$	*prawo*
$[A,$ *brudno*$]$	*odkurz*
$[B,$ *czysto*$]$	*prawo*
$[B,$ *brudno*$]$	*odkurz*
$[A,$ *czysto*$]$, $[A,$ *czysto*$]$	*prawo*
$[A,$ *czysto*$]$, $[A,$ *brudno*$]$	*odkurz*
$\vdots$	$\vdots$

Świat odkurzacza – program agenta

Przykładowy program agenta:

- Jeśli $S_t = [A,$ *brudno*$]$ lub $S_t = [B,$ *brudno*$]$, to *odkurz*.
- Jeśli $S_t = [A,$ *czysto*$]$, to *prawo*.
- Jeśli $S_t = [B,$ *czysto*$]$, to *lewo*.

Tworząc agenta trzeba sobie odpowiedzieć na pytania:

- Jaka funkcja agenta jest odpowiednia?
- Jak ją zwięźle zaimplementować?

Świat odkurzacza – miara jakości

- **Ocena agenta** – na podstawie **stanów środowiska** (potencjalnie będących konsekwencjami podejmowanych akcji)
- **Miara jakości** agenta dokonuje oceny **sekwencji stanów środowiska**, np.
  - $+1$ punkt: za każdy pokój posprzątany do momentu T
  - $+1$ punkt: za każdy czysty pokój w danym kroku t i -1 punkt: za każdy wykonany ruch
- Miarę jakości często wybiera **projektant agenta** i powinna:
  - być oparta na stanach środowiska
  - odzwierciedlać faktyczny cel na którym nam zależy (nasze oczekiwania dotyczące środowiska)

Uwaga

- Miara nie powinna brać pod uwagę tego jak nam się wydaje, że agent powinien działać.

Świat odkurzacza – pytania

Załóżmy następującą miarę jakości:

- Agent jest nagradzany 10 punktami za każdy czysty pokoju po 100 krokach.

Pytanie

Pokaż, że następująca funkcja agenta odruchowego jest racjonalna:

- Jeśli $S_t = [A,$ *brudno*$]$ lub $S_t = [B,$ *brudno*$]$, to *odkurz*.
- Jeśli $S_t = [A,$ *czysto*$]$, to *prawo*.
- Jeśli $S_t = [B,$ *czysto*$]$, to *lewo*.

Świat odkurzacza – pytania

Funkcja agenta:

- Jeśli $S_t = [A,$ *brudno*$]$ lub $S_t = [B,$ *brudno*$]$, to *odkurz*.
- Jeśli $S_t = [A,$ *czysto*$]$, to *prawo*.
- Jeśli $S_t = [B,$ *czysto*$]$, to *lewo*.

Pytanie

Rozważmy zmodyfikowaną wersję środowiska, w której każdy ruch do sąsiedniego pokoju kosztuje agenta 1 punkt.

1. Czy agent opisany w poprzednim pytani (powyżej) jest nadal racjonalny?
2. Co w wypadku agenta z pamięcią? Jaka powinna być funkcja takiego agenta?
3. Co jeśli agent będzie w stanie obserwować obecność kurzu w obu pokojach, czy wtedy agent odruchowy może być racjonalny?

Świat odkurzacza – pytania

Pytanie

Rozważmy niedeterministyczną (stochastyczną) wersję środowiska w której robot odkurzacz jest dodatkowo wadliwy:

- w 50% przypadków akcja *odkurzać* się nie udaje i pozostawia kurz w pokoju nawet jeśli ten był czysty,
- w 25% przypadków czujniki kurzu w pokoju podaje nieprawidłową informację.

Co powinien uwzględniać racjonalny agent w tym wypadku?

Świat odkurzacza – pytania

Pytanie

Rozważmy niedeterministyczną wersję środowiska, w której w oprócz robota odkurzacza w pokojach znajdują się koty z kłaczącym futerkiem:

- W każdym kroku, pokój ma 10% szansy by znowu znowu się ubrudzić.
- Agent jest nagradzany 2 punktami za każdym razem kiedy wyczyści brudny pokój.

Co powinien uwzględniać racjonalny agent w tym wypadku?

Ankieta na koniec wykładu

https://tinyurl.com/zmio2023

Bibliografia

[1] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, third edition.

[2] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.