Spodbujevano učenje pri igranju namiznih iger
(angl. *Reinforcement learning in board games*)

Tim Kalan

Fakulteta za matematiko in fiziko

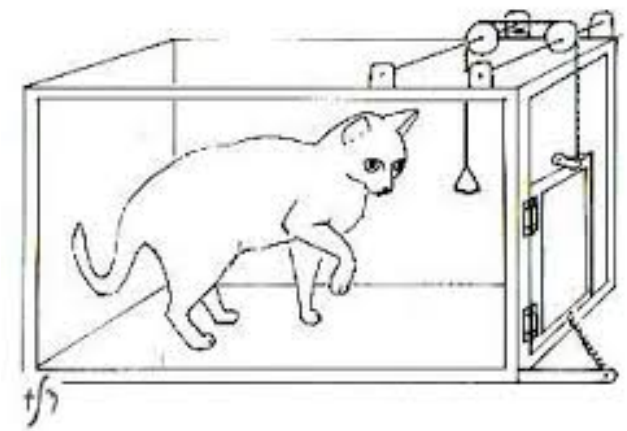
10. november 2020

Strojno učenje

- Nadzorovano učenje *(npr. prepoznavanje števk)*
- Nenadzorovano učenje *(npr. razvrščanje)*
- **Spodbujevano učenje**

Motivacija: Instrumentalno pogojevanje

- Tu bo slika (http://www.edugyan.in/2017/03/edward-lee-thorndike-theory-of-learning.html, https://en.wikipedia.org/wiki/Reinforcement)
- Lepa psihološko motivirana podlaga
- **Nagrade in kazni**



Motivacija: Zakaj namizne igre?

- Aplikacija abstraktnega mišljenja

Motivacija: Zakaj namizne igre?

- Aplikacija abstraktnega mišljenja
- Spremljajo človeštvo že zelo dolgo

Motivacija: Zakaj namizne igre?

- Aplikacija abstraktnega mišljenja
- Spremljajo človeštvo že zelo dolgo
- »Modelirajo« resnično življenje

Motivacija: Zakaj namizne igre?

- Aplikacija abstraktnega mišljenja
- Spremljajo človeštvo že zelo dolgo
- »Modelirajo« resnično življenje
- Uporabno mesto za testiranje algoritmov

Spodbujevano učenje - osnovni koncepti

- **Okolje, agent, nagrada, (model)**
- Pomemben je čas
- Ne poznamo »pravilnih« akcij
- Raziskovanje in izkoriščanje
- Vrednostna funkcija

Kje je to uporabno?

- Naučiti robota hoje
- Upravljati s portfeljem
- Igrati namizne igre
- Igrati katerekoli igre
- ...

Torej res praktično karkoli, kjer lahko cilj modeliramo kot numerične nagrade, ne poznamo pa optimalnih akcij za dostop do teh nagrad.

Problem

Definicija 1 (Hipoteza o nagradi).

Vse cilje je mogoče opisati kot maksimizacijo neke kumulativne numerične nagrade.

- To ni nujno res

Primer: Križci in krožci 1

- tu slika tistega loopa
- **Stanje:** Kje je prazno, kje »X« in kje »O«
- **Agent:** Program, ki se odloča, kako igrati
- **Okolje:** Agentu sporoča nagrade in stanje
- **Nagrada:** Pozitivna za zmago, negativna za poraz

Primer: Križci in krožci 2

- Agent igra igre, posodablja svoje vrednosti stanj glede na odgovor okolja
- Kako naj to stori?

Primer: Križci in krožci 3

- Enostavna ideja:

Primer: Križci in krožci 3

- Enostavna ideja:

$$V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$$

Primer: Križci in krožci 3

- Enostavna ideja:

$$V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$$

- s je trenutno stanje

Primer: Križci in krožci 3

- Enostavna ideja:

$$V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$$

- s je trenutno stanje
- V je vrednostna funkcija

Primer: Križci in krožci 3

- Enostavna ideja:

$$V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$$

- s je trenutno stanje
- V je vrednostna funkcija
- α je velikost koraka (hitrost učenja)

Primer: Križci in krožci 3

- Enostavna ideja:

$$V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$$

- s je trenutno stanje
- V je vrednostna funkcija
- α je velikost koraka (hitrost učenja)
- R je nagrada

Primer: Križci in krožci 3

- Enostavna ideja:

$$V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$$

- s je trenutno stanje
- V je vrednostna funkcija
- α je velikost koraka (hitrost učenja)
- R je nagrada
- γ je diskontni faktor (pomemben je čas)

Primer: Križci in krožci 3

- Enostavna ideja:

$$V(s) \leftarrow V(s) + \alpha[R(s') + \gamma V(s') - V(s)]$$

- s je trenutno stanje
- V je vrednostna funkcija
- α je velikost koraka (hitrost učenja)
- R je nagrada
- γ je diskontni faktor (pomemben je čas)
- s' je stanje, ki sledi s

Bistvo

$$\text{nova ocena} \leftarrow \text{stara ocena} + \text{korak}[\text{cilj/tarča} - \text{stara ocena}]$$

- Tako ocenimo dano strategijo
- Kako pa strategijo dejansko spremenimo?

Formalizacija: Markovski proces odločanja 1

Definicija 2 (Markovska veriga).

Skučajni proces $(S_t)_{t=0}^{T}$ *na končnem verjetnostnem prostoru* (Ω, P) *je* ***Markovska veriga****, če velja Markovska lastnost*

$$P(S_{t+1} = s_{t+1} | S_t = s_t, ..., S_0 = s_0) = P(S_{t+1} = s_{t+1} | S_t = s_t)$$

Formalizacija: Markovski proces odločanja 1

Definicija 2 (Markovska veriga).
Skučajni proces $(S_t)_{t=0}^{T}$ *na končnem verjetnostnem prostoru* (Ω, P) *je* ***Markovska veriga****, če velja Markovska lastnost*

$$P(S_{t+1} = s_{t+1} | S_t = s_t, ..., S_0 = s_0) = P(S_{t+1} = s_{t+1} | S_t = s_t)$$

- Prihodnost je neodvisna od preteklosti, če poznamo sedanjost

Formalizacija: Markovski proces odločanja 1

Definicija 2 (Markovska veriga).
Skučajni proces $(S_t)_{t=0}^T$ *na končnem verjetnostnem prostoru* (Ω, P) *je* ***Markovska veriga****, če velja Markovska lastnost*

$$P(S_{t+1} = s_{t+1} | S_t = s_t, ..., S_0 = s_0) = P(S_{t+1} = s_{t+1} | S_t = s_t)$$

- Prihodnost je neodvisna od preteklosti, če poznamo sedanjost
- Definiramo $p_{ss'} := P(S_{t+1} = s' | S_t = s)$ in to združimo v matriko $\mathcal{P} := [p_{ss'}]_{s,s' \in \mathcal{S}}$, kjer je $\mathcal{S}$ množica stanj

Formalizacija: Markovski proces odločanja 2

Definicija 3 (Markovski proces odločanja).

Markovski proces odločanja *je nabor* $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$*, kjer je*

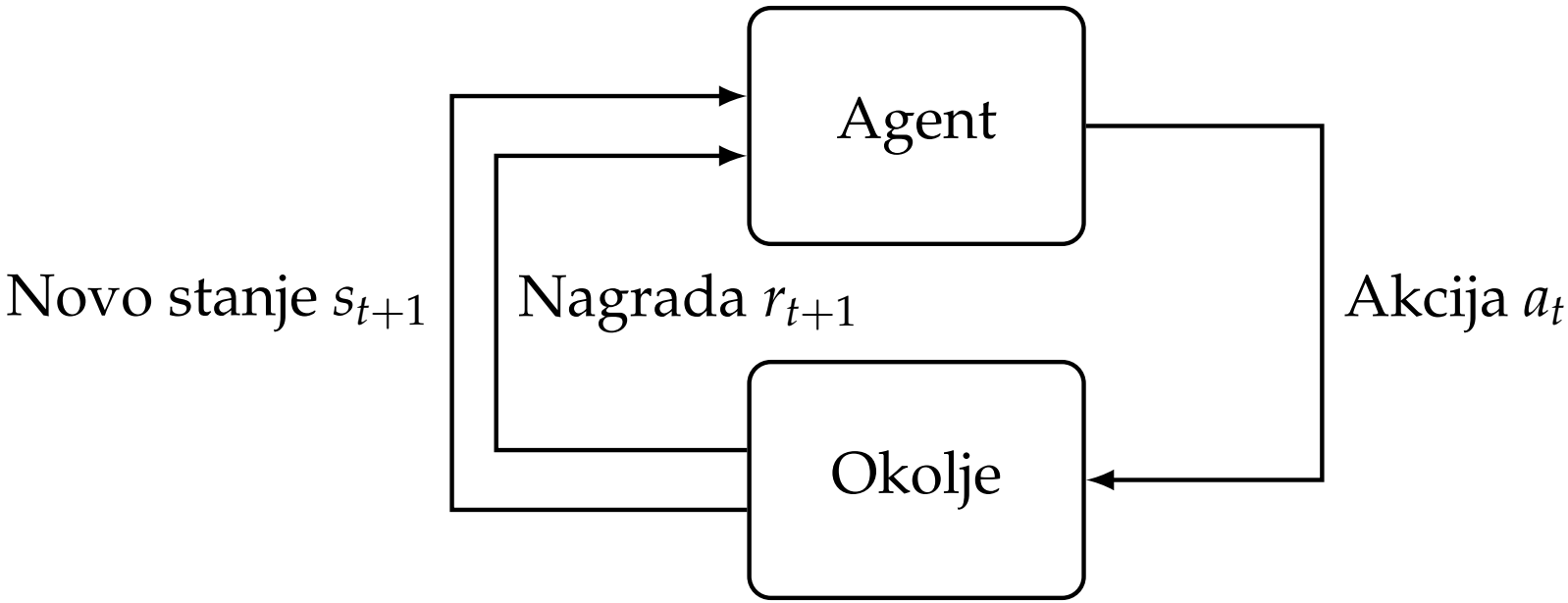
- $\mathcal{S}$ *je (končna) množica stanj*
- $\mathcal{A}$ *je (končna) množiza akcij oz. dejanj*
- $\mathcal{P}$ *je prehodna matrika, kjer* $p^a_{ss'} = P(S_{t+1} = s' | S_t = s, A_t = a)$
- $\mathcal{R}$ *je nagradna funkcija* $\mathcal{R}^a_s = E[R_{t+1} | S_t = s, A_t = a]$
- $\gamma \in [0, 1]$ *je diskontni faktor*

Kako lahko to posplošimo

- Koliko stanj imamo?
- Do kje lahko pridemo?
- Kdaj odpove?
- Kaj je rešitev?

Demonstracija: Križci in krožci

Morda kakšna slika/grafikon


Agent
Okolje
Novo stanje s_{t+1}
Nagrada r_{t+1}
Akcija a_t

Ideje

- Formalizacija, V, Q, pi, ...