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

Tim Kalan

Mentor: izr. prof. dr. Marjetka Knez

Fakulteta za matematiko in fiziko

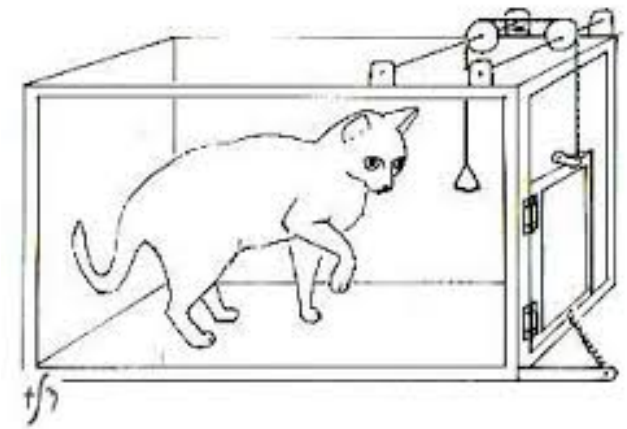
30. marec 2021

Napovednik

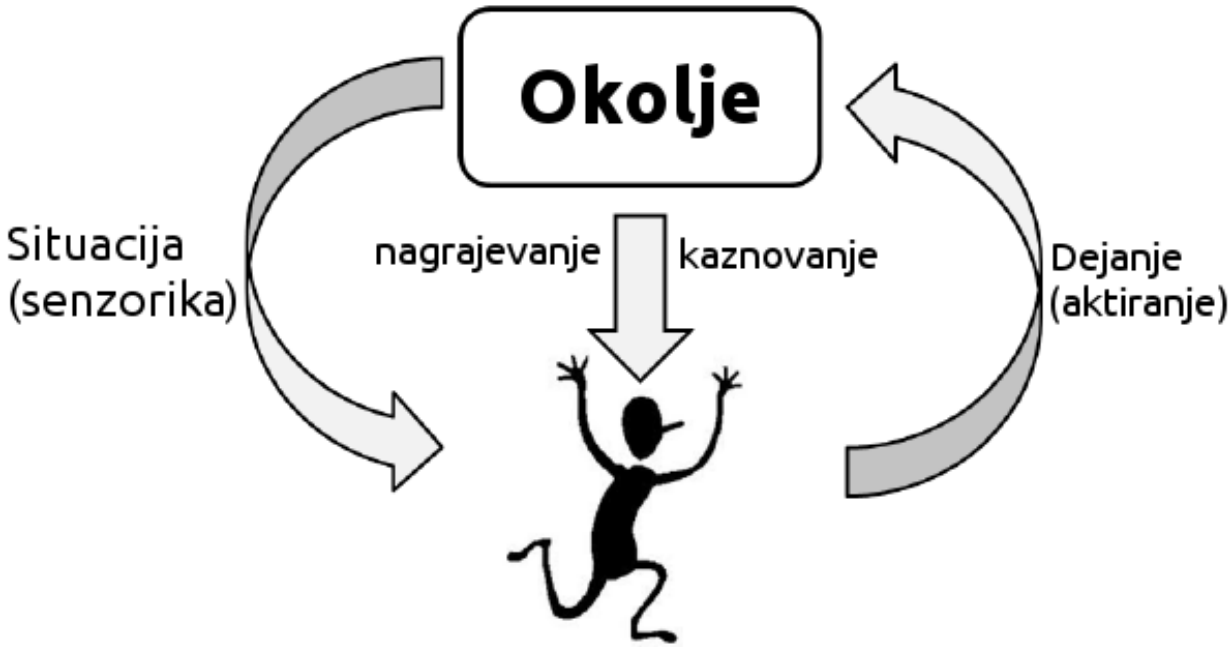
- Motivacija,
- problem spodbujevalnega učenja,
- algoritmi,
- namizne igre.

Motivacija: Instrumentalno pogojevanje

- Psihološko motivirana podlaga.
- **Nagrade in kazni**.



Okvir

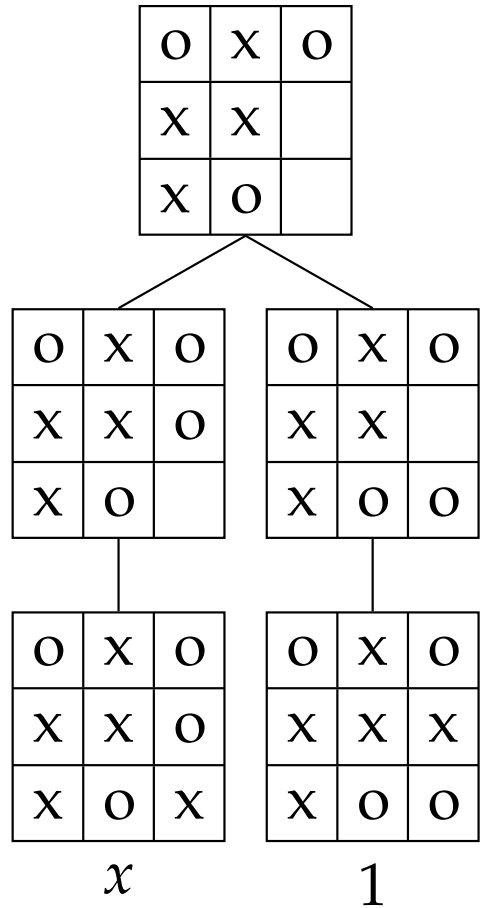

Okolje
Situacija
(senzorika)
nagrajevanje
kaznovanje
Dejanje
(aktiranje)

Primer 1: robot se uči hoje

- **Situacija/Stanje**: položaj v sobi in stanje nog,
- **Nagrada**: 1 za doseg vrat, 2 za ključ, -0.5 za časovni korak,
- **Okolje**: soba in senzorji, ki govorijo o položaju,
- **Akcija**: Premik noge.

Primer 2: križci in krožci

- **Situacija/Stanje**: stanje na plošči,
- **Nagrada**: 1 za zmago, -1 za poraz, x za izenačenje/korak,
- **Okolje**: nasprotnik, plošča, sodnik, nagrajevalec,
- **Akcija**: postavitev X oz. O na ploščo.



Ideja

- Agent »pade« v okolje.
- S poskušanjem se nauči pravilnih akcij.
- Svoje znanje izkoristi za maksimizacijo nagrade.

hipoteza o nagradi??

Formalizacija: Markovski proces odločanja 1

Definicija 1 (Markovska veriga).
Slučajni proces $(S_t)_{t=0}^{T}$ *na končnem verjetnostnem prostoru* $(\Omega, \mathcal{F}, P)$ *je* ***Markovska veriga****, če velja Markovska lastnost*

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

Formalizacija: Markovski proces odločanja 1

Definicija 1 (Markovska veriga).
Slučajni proces $(S_t)_{t=0}^{T}$ *na končnem verjetnostnem prostoru* $(\Omega, \mathcal{F}, P)$ *je* ***Markovska veriga****, če velja Markovska lastnost*

$$P(S_{t+1} = s_{t+1} \mid S_t = s_t, \ldots, S_0 = s_0) = P(S_{t+1} = s_{t+1} \mid S_t = s_t)$$

- Prihodnost je neodvisna od preteklosti, če poznamo sedanjost

Formalizacija: Markovski proces odločanja 1

Definicija 1 (Markovska veriga).
Slučajni proces $(S_t)_{t=0}^{T}$ *na končnem verjetnostnem prostoru* $(\Omega, \mathcal{F}, P)$ *je* ***Markovska veriga****, če velja Markovska lastnost*

$$P(S_{t+1} = s_{t+1} \mid S_t = s_t, \ldots, S_0 = s_0) = P(S_{t+1} = s_{t+1} \mid S_t = s_t)$$

- Prihodnost je neodvisna od preteklosti, če poznamo sedanjost
- $p_{ss'} := P(S_{t+1} = s' \mid S_t = s) \rightarrow \mathcal{P} := [p_{ss'}]_{s,s' \in \mathcal{S}}$, $\mathcal{S}$ je množica stanj
- Markovska veriga je torej dvojica $(\mathcal{S}, \mathcal{P})$

Formalizacija: Markovski proces odločanja 2

Definicija 2 (Markovski proces nagrajevanja).

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

- $\mathcal{S}$ *je (končna) množica stanj*
- $\mathcal{P}$ *je prehodna matrika, kjer* $p_{ss'} = P(S_{t+1} = s' \mid S_t = s)$
- $\mathcal{R}$ *je nagradna funkcija* $\mathcal{R}_s = E[R_{t+1} \mid S_t = s]$
- $\gamma \in [0, 1]$ *je diskontni faktor*

Formalizacija: Markovski proces odločanja 3

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

Primer: MDP

Agent 1

- Strategija (angl. *Policy*)
- Vrednostna funkcija (angl. *Value function*)
- (Model)

Agent 2: strategija

Definicija 4.

- ***Deterministična strategija*** *stanju s priredi akcijo a,*

$$\pi(s) = a.$$

- ***Stohastična strategija*** *za vsako stanje s pove verjetnosti vseh možnih akcij a,*

$$\pi(a|s) = P(A_t = a \mid S_t = s).$$

Agent 3: povračilo

Definicija 5 (Povračilo).

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Agent 4: vrednostna funkcija

Definicija 6 (Vrednostna funkcija).

- ***Vrednostna funkcija stanja*** *je pričakovana vrednost povračila, če se vedemo skladno s strategijo* π

$$v_\pi(s) = \mathrm{E}[G_t \mid S_t = s].$$

- ***Vrednostna funkcija akcije*** *je podobna prejšnji, le da sprosti prvo akcijo*

$$q_\pi(s, a) = \mathrm{E}[G_t \mid S_t = s, A_t = a].$$

Primer - križci in krožci

Algoritmi - dinamično programiranje

Algoritmi - Monte Carlo

Algoritmi - TD(0)

Algoritmi - TD(λ)

Primer - kje se zatakne?

- velike plošče
- pri zgornjih algoritmih hranimo vse v tabeli

Nevronske mreže

Motivacija - namizne igre

- Abstraktno mišljenje.
- Model življenja.
- Testiranje algoritmov.

Namizne igre - postanja

Namizne igre - trening

Namizne igre - tradicionalne metode

Namizne igre - združevanje

Nekaj rezultatov

Literatura


Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An introduction*. The MIT Press, 2015.


Imran Ghory. *Reinforcement learning in board games*. 2004.