

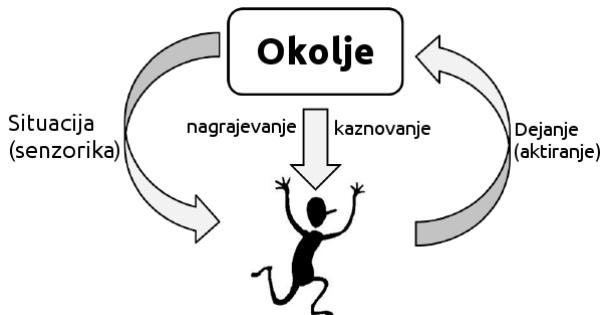
Spodbujevano učenje pri igranju namiznih iger (angl. *Reinforcement learning in board games*)

Tim Kalan

Mentor: izr. prof. dr. Marjetka Knez

Fakulteta za matematiko in fiziko

30. marec 2021

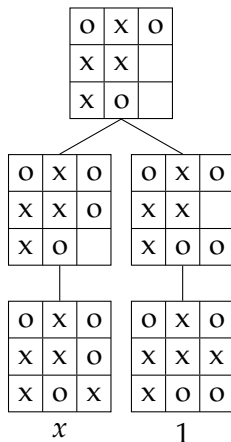


Primer 1 - Robot se uči hoje

- ▶ **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.



Formalizacija: Markovski proces odločanja 1

Definicija 1 (Markovska veriga).

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

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

Formalizacija: Markovski proces odločanja 1

Definicija 1 (Markovska veriga).

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

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

- ▶ 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 (Ω, \mathcal{F}, P) je **Markovska veriga**, če velja Markovska lastnost

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

- ▶ 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 $p_{ss'}^a = P(S_{t+1} = s' \mid S_t = s, \mathbf{A}_t = \mathbf{a})$
- ▶ \mathcal{R} je nagradna funkcija $\mathcal{R}_s^a = E[R_{t+1} \mid S_t = s, \mathbf{A}_t = \mathbf{a}]$
- ▶ $\gamma \in [0, 1]$ je diskontni faktor

Agent

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

Strategija

Agentova **strategija** je takšna preslikava $\pi : S \rightarrow A$ da velja:

$$a = \pi(s) \text{ oz.}$$

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

Pri čemer prva formula definira *deterministično* strategijo, druga pa *stohastično*. a in s sta realizaciji akcije in stanja v času t .

Vrednostna funkcija

Literatura



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



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