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

Tim Kalan

Mentor: izred. prof. dr. Marjetka Knez

Fakulteta za matematiko in fiziko

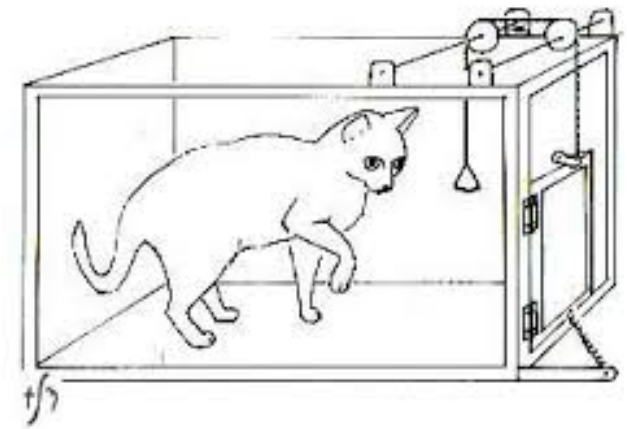
20. november 2020

Strojno učenje

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

Motivacija: Instrumentalno pogojevanje

- Lepa psihološko motivirana podlaga
- **Nagrade in kazni**

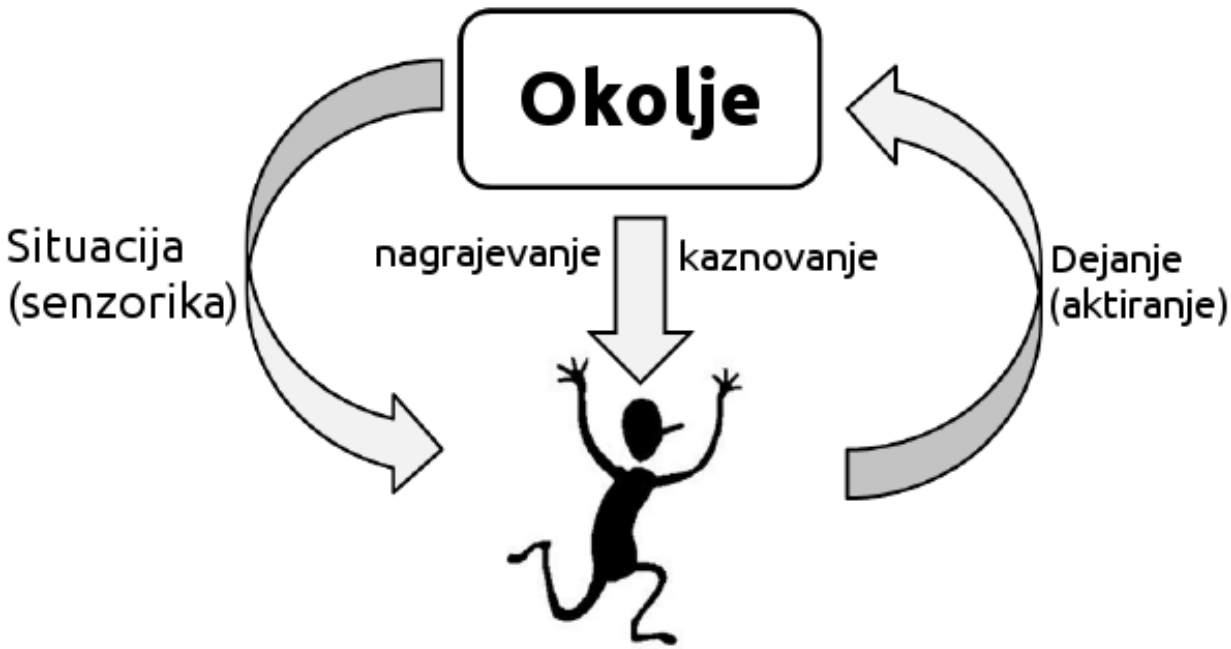


Motivacija: Zakaj namizne igre?

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

Spodbujevano učenje - osnovni koncepti 1

- Nagrada
- Agent
- Okolje
- Stanje
- Akcija



Spodbujevano učenje - osnovni koncepti 2

- Pomemben je čas.
- Ne poznamo »pravilnih« akcij.
- Raziskovanje in izkoriščanje

RL agent

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

Kje je to uporabno?

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

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

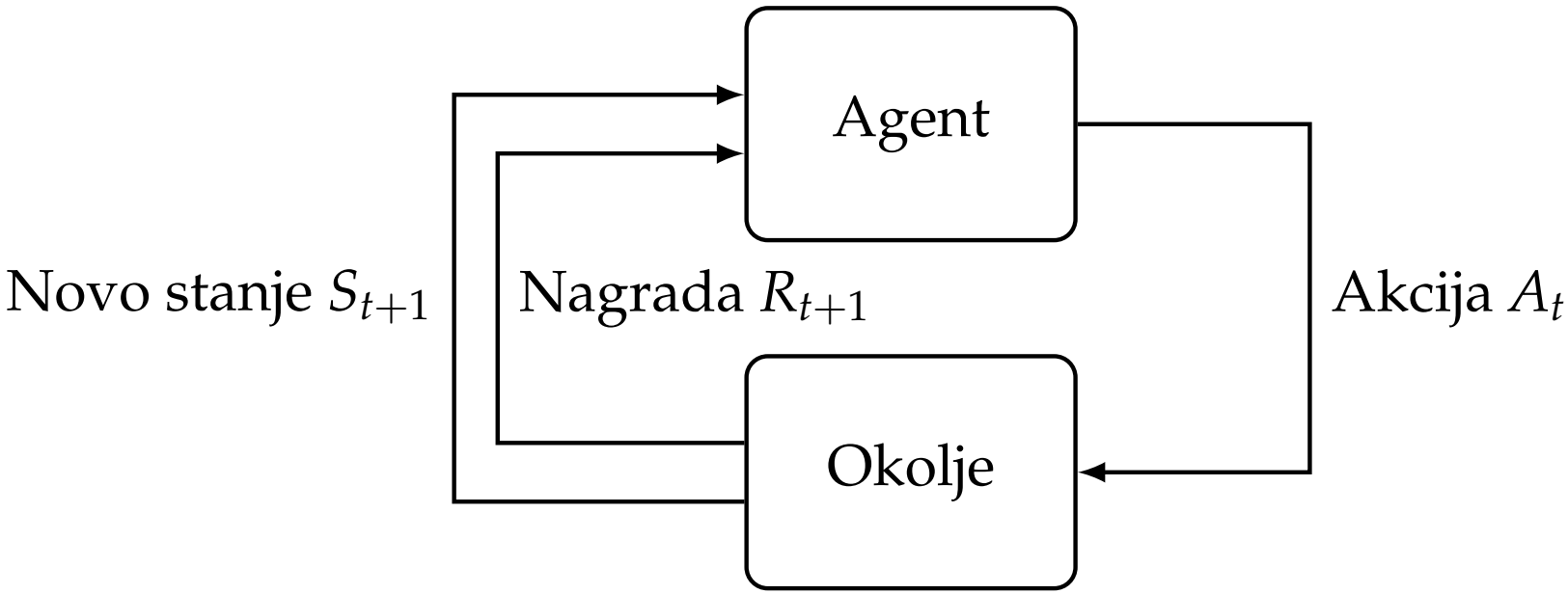
Problem

Definicija 1 (Hipoteza o nagradi).

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

- Je to vedno res?

Primer: Križci in krožci 1



- **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.
- **Akcija:** Postavitev »X« oz. »O« na ploščo.

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:

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

Primer: Križci in krožci 3

- Enostavna ideja:

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

- s je trenutno stanje

Primer: Križci in krožci 3

- Enostavna ideja:

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

- s je trenutno stanje
- V je vrednostna funkcija

Primer: Križci in krožci 3

- Enostavna ideja:

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

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

Primer: Križci in krožci 3

- Enostavna ideja:

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

- s je trenutno stanje
- V je vrednostna funkcija
- α je velikost koraka (hitrost učenja)
- s' je stanje, ki sledi s

Zgornje je primer **učenja s časovno razliko** (angl. *Temporal difference learning*)

Bistvo

$$\textit{nova ocena} \leftarrow \textit{stara ocena} + \textit{korak}[\textit{tarča} - \textit{stara ocena}]$$

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

ϵ-požrešna izboljšava strategije

Izberemo »najboljšo« akcijo.

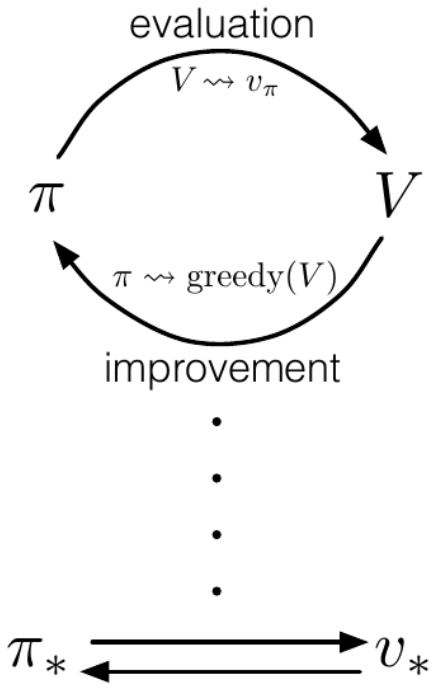
ϵ-požrešna izboljšava strategije

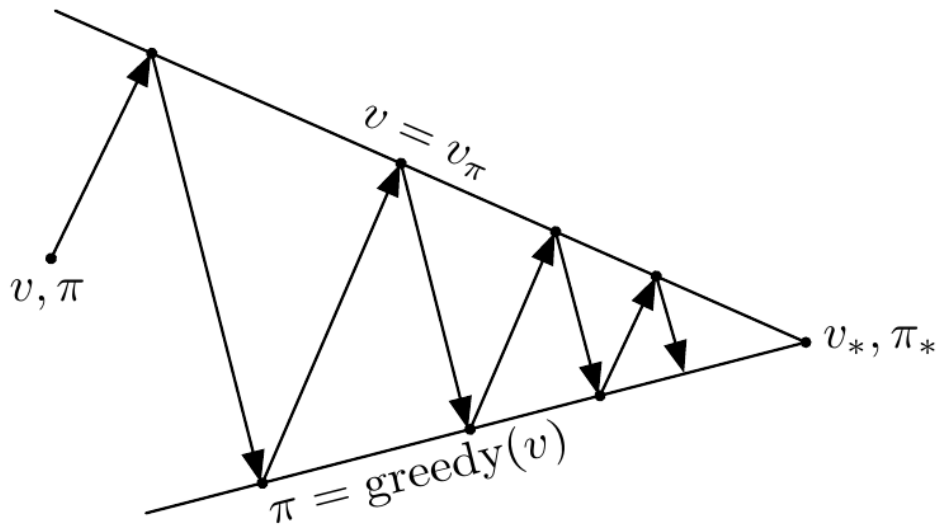
~~Izberemo »najboljšo« akcijo.~~

ϵ-požrešna izboljšava strategije

~~Izberemo »najboljšo« akcijo.~~

- »Ponavadi« izberemo »najboljšo« akcijo.
- Z verjetnostjo ϵ izberemo naključno akcijo.





Alternativa: Monte Carlo spodbujevano učenje 1

Definicija 2.

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

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

- *Naj bodo $R_{t+1}, ..., R_T$ nagrade, ji jih bomo prejeli od trenutka t do konca epizode. **Povračilo** G_t definiramo kot*

$$G_t = R_{t+1} + \gamma R_{t+2} + ... + \gamma^{T-1} R_T.$$

- *Naj bo π dana strategija agenta. **Vrednostna funkcija stanja** glede na dano strategijo $v_\pi(s)$ je*

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

Alternativa: Monte Carlo spodbujevano učenje 2

- Ob obisku stanja s:

$$N(s) \leftarrow N(s) + 1$$
$$S(s) \leftarrow S(s) + G_t$$

- Po koncu učenja:

$$V(s) \leftarrow S(s)/N(s)$$

- Pomni: Računanje povprečja zaporedja $(X_i)_{i \in \mathbb{N}}$

$$\mu_k = \frac{1}{k} \sum_{j=1}^{k} X_j = \mu_{k-1} + \frac{1}{k}(X_k - \mu_{k-1})$$

- Inkrementalni Monte Carlo:

$$V(s) \leftarrow V(s) + \alpha(G_t - V(S_t))$$

Kam gremo od tu?

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

- Monte Carlo metode ponavadi nastopijo, ko nekaj aproksimiramo, kako je s tem v našem primeru?

Ideje za naprej

- Drugi algoritmi
- Problem časovne dodelitve zaslug
- Večje igre - nevronske mreže
- Različni tipi učenja

Literatura


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


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