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

Tim Kalan

Fakulteta za matematiko in fiziko

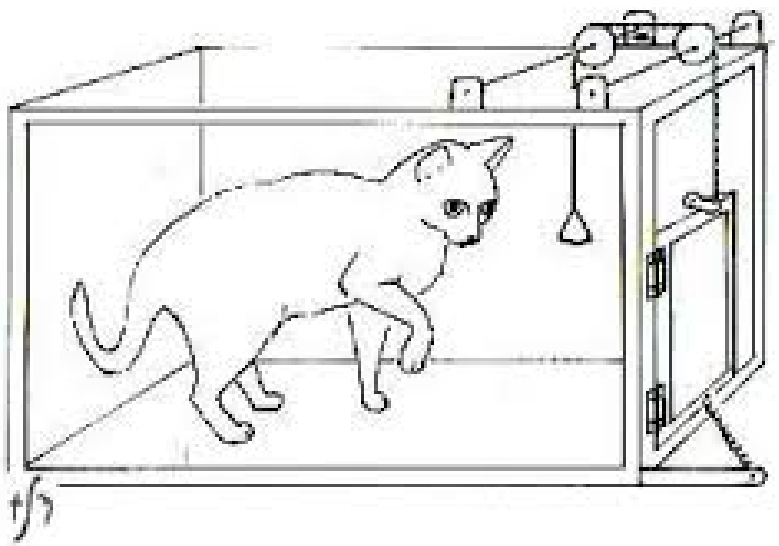
6. november 2020

Strojno učenje

- Nadzorovano učenje
- Nenadzorovano učenje
- **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**



Spodbujevano učenje

- **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

Reward hypothesis

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:

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

Primer: Križci in krožci 3

- Enostavna ideja:

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

- s je trenutno stanje

Primer: Križci in krožci 3

- Enostavna ideja:

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

- s je trenutno stanje
- V je vrednostna funkcija

Primer: Križci in krožci 3

- Enostavna ideja:

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

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

Primer: Križci in krožci 3

- Enostavna ideja:

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

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

Primer: Križci in krožci 3

- Enostavna ideja:

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

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

Bistvo

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

Kako lahko to posplošimo

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

Demonstracija: Križci in krožci

Morda kakšna slika/grafikon