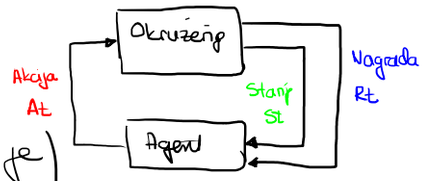
Učenje potkrepljivanjem (eng. Reinforcement learning)
Beleške za kurs Veštačke inteligencije na Matematičkom fakultetu

NEMANJA MIĆOVIĆ



Markovljev proces odlučivanja : (S, A, R, J, P, γ)

$\rightarrow S_0, A_0, R_0, S_1, A_1, R_1, S_2, A_2, R_2, \ldots$

S – skup svih stanja

A – skup svih akcija

R – skup svih nagrada ($R \subseteq \mathbb{R}$)

J – trajektorija

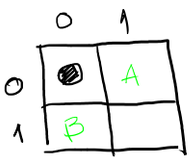
P – funkcija prelaska u okruženju

P je "teška" funkcija, a u praksi je zgodno da je svedemo na verovatnoću

$$P: S \times R \times S \times A \longrightarrow \mathbb{R}$$

$$P(s', r \mid s, a) = P(S_{t+1} = s', R_t = r \mid S_t = s, A_t = a)$$

γ – faktor umanjenja $\in [0,1]$



Igrač je na (0,0) Nagrada ume ista, +1

dostupne akcije: $a_0 = \downarrow$ $a_1 = \rightarrow$

$$P((0,1), 1 \mid (0,0), \rightarrow) = 1 = P(S_{t+1} = (0,0), R_t = 1 \mid S_t = (0,0), A_t = \rightarrow)$$

$$P((0,1), 1 \mid (0,0), \downarrow) = 0 = P(S_{t+1} = (0,0), R_t = 1 \mid S_t = (0,0), A_t = \downarrow)$$

A

Markovljevo svojstvo

prethodna istorija — Nije važno kako smo ovde došli

$$P(S_t, R_{t-1} \mid \cancel{S_0, A_0, R_0, \ldots}, S_{t-1}, A_{t-1}) = P(S_t, R_{t-1} \mid S_{t-1}, A_{t-1})$$

Ako napustiš partiju Šaha u stanju S, onda drugi igrač može da se takmiči

Ako je poznato trenutno stanje procesa, za zaključivanje o budućnosti poznavanje prošlosti nije neophodno

Kada agent "interaguje sa okolinom" generišu se epizode

Politika π → funkcija pomoću koje agent donosi odluke.

↳ raspodela verovatnoće

$\pi(a \mid s)$ → verovatnoća da će agent preduzeti akciju a ako se nađe u stanju s

↓

Mogu biti deterministička i nedeterministička

Dobitak od koraka t

Tipično nas zanima G_0.

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \qquad (\gamma^0 = 1)$$

γ^k → Ceni više skorije nagrade nego daleke nagrade

Zavisi od politike

Deo definicije MDP-a (faktor umanjenja)

$$G_t = \gamma^0 R_t + \gamma^1 R_{t+1} + \gamma^2 R_{t+2} + \ldots$$

$$G_t = R_t + \gamma G_{t+1}$$

γ - faktor umanjenja vaga znači kratkoročnih i dugoročnih nagrada

$\gamma = 0$ → maksimalan fokus na kratkoročnu nagradu → pohlepna politika

U praksi na primer $\gamma = 0.9$; sličnu

T $\gamma^0 = 1$

Šta znači rešiti Markovljev proces odlučivanja ?

Naći politiku koja maksimizuje očekivanu dobitak agenta

→ Teško biti egzaktan usled stohastičnosti

$$J = S_0, A_0, R_0, S_1, A_1, R_1, S_2, A_2, R_2$$

$$P(J) = P(S_0) \prod_{t=0}^{\infty} \pi(A_t \mid S_t) P(S_{t+1}, R_t \mid S_t, A_t)$$

Raspodela po svim mogućim trajektorijama

Raspodela po mogućim početnim stanjima

Politika koju prati agent

Verovatnoća da će agent da preduzme A_t kada dođe u S_t

Funkcija prelaska okoline

Verovatnoća da će da se desi ono što je agent uradio

Očekivana nagrada koju agent dobije prateći politiku π iz stanja s

$$v^{\pi}(s) = \mathbb{E}_{\pi}\left[G_t \mid S_t = s\right]$$

($\mathbb{E}_{\pi}$: prateći politiku π)

Funkcija vrednosti stanja

$$q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[G_t \mid S_t = s, A_t = a\right]$$

Funkcija vrednosti akcije u stanju

→ v_π i q^π se mogu povezati

$$v_\pi(s) = \mathbb{E}\left[G_t \mid S_t = s\right]$$

$v^\pi \to q^\pi$:

$$= \sum_a \pi(a|s)\, \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right]$$

($\pi(a|s)$: verovatnoća akcije)

$$= \sum_a \pi(a|s)\, q^\pi(s,a)$$

Otežano verovatnoćom da ćemo izabrati akciju a u stanju s

kao i obrnuto

($G_t = R_t + \gamma G_{t+1}$)

$q^\pi \to v^\pi$:

$$q^\pi(s,a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right]$$

$$= \mathbb{E}\left[R_t + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, \mathbb{E}\left[G_t \mid S_t = s'\right]\right]$$

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, v_\pi(s')\right]$$

Više nam odgovara $v^\pi \to q^\pi$ ⟸ Ovo je u praksi teže dobiti jer zahteva da nam je poznato kako funkcioniše okruženje

Kako da poredimo 2 politike po kvalitetu?

kontrast: Postoje pristupi koji modeluju i okruženje

Ako važi za sva stanja $s \in S$

$v_{\pi_1}(s) \geq v_{\pi_2}(s)$, jasno je da π_1 bolja od π_2

(ili barem jednaka ...)

Za MDP postoji optimalna politika π_* koja je bolja (ili jednaka) svim ostalim politikama

→ Želimo (je) naći! → One dele v_* i q_* → Imaće ne bi bile optimalne

$$v_*(s) = \max_\pi v_\pi(s) \qquad q_*(s,a) = \max_\pi q_\pi(s,a)$$

Važi: $v_*(s) = \max_a q_*(s,a)$

Najbolja akcija a u stanju s

Ako poznajemo funkciju q_*, možemo izračunati bar jednu optimalnu politiku.

$$\pi_*(a|s) = \begin{cases} 1, & \text{ako } a = \arg\max_a q_*(a,s) \\ 0, & \text{inače} \end{cases}$$

→ Deterministička politika

⇒ Želimo neke načine da izračunavamo q_*

$$v_*(s) = \max_a q_*(s,a)$$

$$= \max_a \mathbb{E}_{\pi_*}\left[G_t \mid S_t = s, A_t = a\right]$$

$$= \max_a \mathbb{E}_{\pi_*}\left[R_t + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$= \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, \mathbb{E}\left[G_{t+1} \mid S_t = s', A_t = a\right]\right]$$

(dobili → usrednjujemo)

$$= \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, v_*(s')\right]$$

($v_*(s')$)

Bellmanova (rekurentna) jednačina za v_*

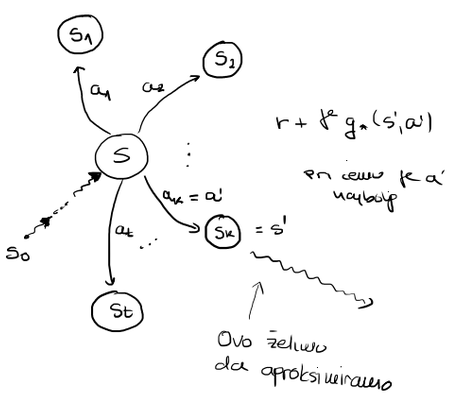
Bellmanova jednakost za q je

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right]$$

($q_*(s',a')$)

Otežavamo nagradu i budućo dobitke verovatnoćama mogućih prelaza

Najbolje što možemo iz stanja gde ćemo dospeti.



Pomoću pokazanog može doći do iterativnog postupka

$$v_{i+1}(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, v_i(s')\right]$$

$$q_{i+1}(s,a) \leftarrow \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_i(s',a')\right]$$

13.2 Učenje u nepoznatom okruženju

Prethodni iterativni postupak nam ne odgovara jer zahteva poznavanje funkcije okruženja P.

$\Rightarrow$ Agent treba da **istražuje** okruženje

→ novi element u našem razmišljanju

Podsećanje, Bellmanova jednačina za q glasi

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right]$$

Ekstrahujemo sumiranje kada aproksimiramo samo povratnom jednom opažanja

$$r + \gamma \max_{a'} q_*(s',a')$$

q_* nam nije poznata → pokušaćemo da je ocenimo

Q learning

Inverzna aproksimacija povodom koje uvodimo važnost

$$q(s,a) \leftarrow (1-\alpha_t)\, q(s,a) + \alpha_t\left(r + \gamma \max_{a'} q(s',a')\right)$$

Popravljamo aproksimaciju kroz 'učenje' (iteracije)

α_t – korak učenja (eng. learning rate)

↳ na primer: $\sum_{t=0}^{\infty} \alpha_t = \infty$ $\quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty$ (Robins-Monro uslovi)

Koliko verujemo staroj, našoj aproksimaciji? Hiperparametar α_t.

Na primer $\frac{1}{t}$ (t broj iteracija)

ε pohlepna politika

Politika koja je pohlepna „do na ε". Postoji neka verovatnoća da politika ne bude pohlepna. Precizirje, politika

$$\pi_{q,\varepsilon}(a|s) = \begin{cases} 1-\varepsilon, & \text{ako } a = \arg\max_{a'} q(s,a') \\ \frac{\varepsilon}{|\mathcal{A}|-1}, & \text{inače} \end{cases}$$

Primer Neka je S veliko i neuređeno stanje $\mathcal{A} = \{\downarrow, \rightarrow, \uparrow, \leftarrow\}$

Ako je samo pohlepna na primer	Ako je $\varepsilon = 0.3$ pohlepna
$\pi_q(\uparrow|s) = 1$	$\pi_q(\uparrow|s) = 1-\varepsilon = 1-0.3 = 0.7$
$\pi_q(\downarrow|s) = 0$	$\pi_q(\downarrow|s) = \frac{\varepsilon}{|\mathcal{A}|-1} = \frac{0.3}{4-1} = \frac{0.3}{3} = 0.1$
$\pi_q(\rightarrow|s) = 0$	$\pi_q(\rightarrow|s) = 0.1$
$\pi_q(\leftarrow|s) = 0$	$\pi_q(\leftarrow|s) = 0.1$

Verovatnoća koja se ravnomerno rasporedi na ostale akcije u odnosu na (dosadašnju) najbolju.

Često ε smanjujemo tokom algoritma kako bi manje istraživali. Na primer $\varepsilon_t = 1/t$, gde je t broj trenutne iteracije.

Algoritam q-učenja
Ulaz: Broj iteracija N
Izlaz: Aproksimacija funkcije q_*

1. Nasumično inicijalizuj za sve parove s i a
2. Inicijalizuj S na početno stanje
3. $t \leftarrow 1$
4. ponavljaj ↳ ε pohlepna politika sa $\varepsilon = 1/t$
5. preduzmi akciju $a \sim \pi_{q,1/t}(a|s)$; opazi nagradu r, novo stanje S'
6. $\alpha_t \leftarrow 1/t$ // ažurira se korak učenja
7. $q(s,a) \leftarrow (1-\alpha_t)\, q(s,a) + \alpha_t\left(r + \gamma \max_{a'} q(s',a')\right)$ // ovde se dešava „učenje"
8. ako je S' završno stanje onda
9. inicijalizuj S na početno stanje
10. inače
11. $s \leftarrow s'$
12. $t \leftarrow t+1$
13. dok nije ispunjen uslov $t = N$
14. Vrati q kao rešenje

\+ Vrlo jednostavno

\- Memorijski (potencijalno) zahtevna tabela q

\- Nema moć generalizacije sa stanja na stanje

Možemo li ovo primeniti na „svet"?

Teorijski: da!

Praktično: NE! Možemo li mi čuvati q tabelu u memoriji?

To je matrica dimenzija $|S| \times |A|$

Nastavićemo se! :)