I007 Osnove umjetne inteligencije

Tema: Učenje s podrškom - vježbe

4. 5. 2020.

predavač: Darija Marković

asistent: Darija Marković

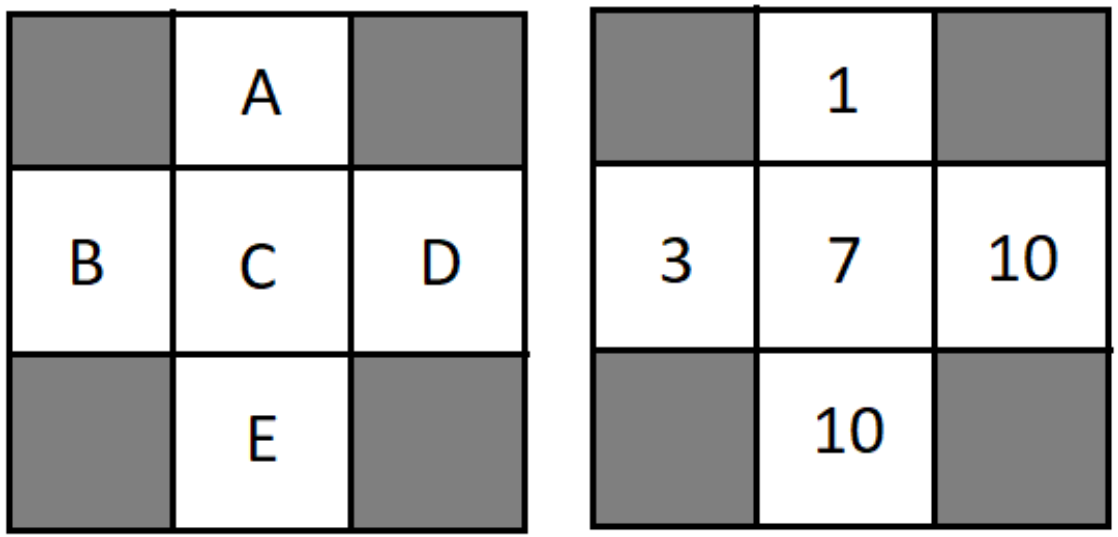
Učenje vremenskih razlika (Temporal Difference Learning - TDL)

- želimo naučiti iz svakog primjera, te ažuriramo vrijednost $V(s)$ nakon svakog novog primjera (s, a, s', r)
- vjerojatniji ishod s' će češće doprinijeti ažuriranju
- uvodimo faktor učenja α
- TDL učenje vrijednosti (uz fiksiranu strategiju π):
  - novi primjer za $V(s)$: $R(s, \pi(s), s') + \gamma V^{\pi}(s')$
  - ažuriramo staru vrijednost $V^{\pi}(s)$:

$$V^{\pi}(s) \leftarrow (1 - \alpha)V^{\pi}(s) + \alpha[R(s, \pi(s), s') + \gamma V^{\pi}(s')]$$

Zadatak 1.

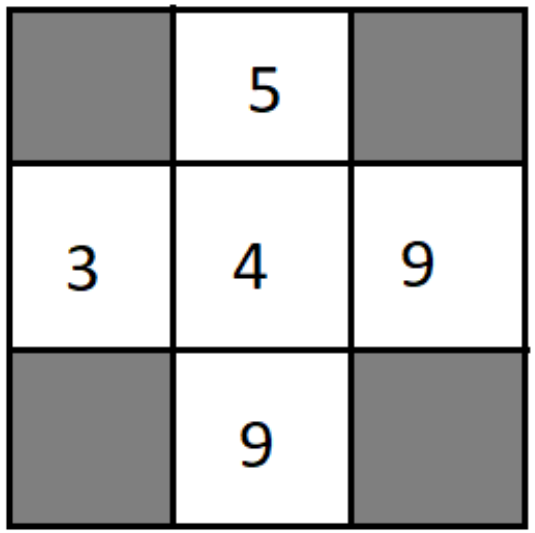
Zadana je sljedeća mreža s procijenjenim vrijednostima



Nakon toga dobijemo sljedeći primjer: $B, \rightarrow, C, -3$. Uz $\gamma = 1$ i faktor učenja $\alpha = 0.5$ odredite nove procjene za vrijednosti.

Zadatak 2.

Za mrežu iz prethodnog zadatka i za dane vrijednosti



odredite nove procjene za vrijednosti u stanjima nakon izvršavanja primjera $B, \rightarrow, C, -1$ pa $C, \uparrow, A, 2$ uz $\gamma = 0.9$ i $\alpha = 0.4$.

Q-učenje (Q-learning)

- iteracija q-vrijednosti bazirana na primjerima

$$Q(s,a) \leftarrow \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma Q(s',a')]$$

- TDL za učenje q-vrijednosti:
  - dobijemo novi primjer (s, a, s', r)
  - uzmemo u obzir staru procjenu $Q(s,a)$
  - uzmemo u obzir procjenu za novi primjer: $R(s,a,s') + \gamma \max_{a'} Q(s',a')$
  - ažuriramo staru vrijednost $Q(s,a)$:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha[R(s,a,s') + \gamma \max_{a'} Q(s',a')]$$

Zadatak 3.

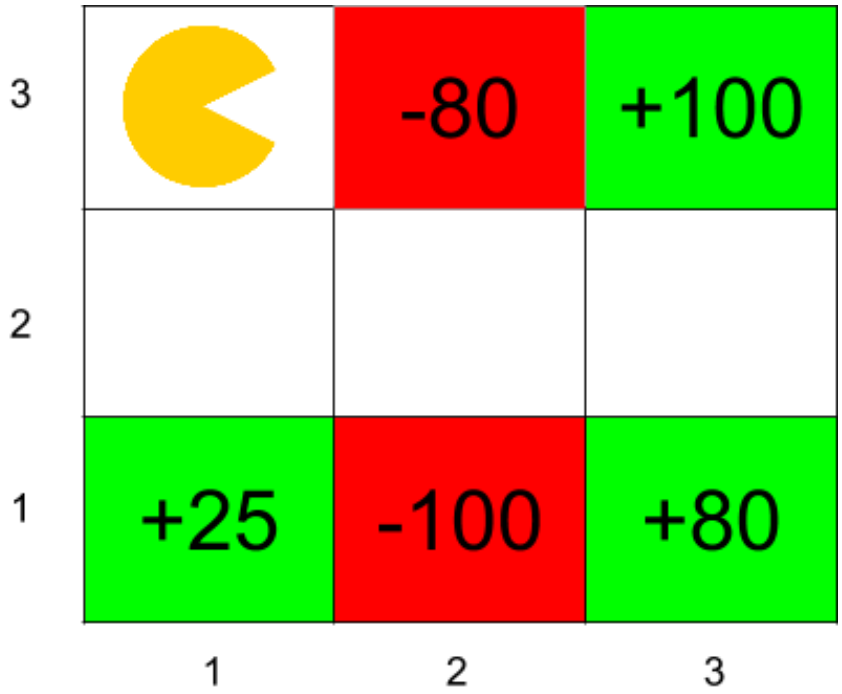
Razmatramo MPO za tri stanja A,B,C i dvije akcije ”$+$” i ”$-$”. Pretpostavimo da je $\gamma = 0.5$ i $\alpha = 0.5$ i da je trenutna procjena Q vrijednosti dana tablicom:

	A	B	C
$-$	0.43	1.061	-0.035
$+$	0.951	1.484	8.434

Dobijemo dva nova primjera $A, +, B, 0$ i $B, -, C, 0$. Odredite q-vrijednosti.

Zadatak 4.

Pretpostavite da poznati Pacman želi naučiti optimalnu strategiju izlaska iz mreže. Ako traganje za izlaskom završi u nekom od obojanih stanja slijedi navedena nagrada. Sva obojana stanja su ciljna (tj. Pacman završava svoje traganje ako dođe u jedno od tih stanja). Sva ostala stanja dozvoljavaju akcije iz skupa $\{Sjever, Istok, Jug, Zapad\}$ koje će deterministički prebaciti Pacmana u jedno od susjednih stanja (ili ga ostaviti u mreži ukoliko akcija pokušava Pacmana izbaciti van). Pretpostavite da je faktor umanjenja $\gamma = 0.5$ i da je faktor učenja $\alpha = 0.5$. Pacman kreće u potragu iz stanja $(1, 3)$.



Dane su sljedeće epizode prolazaka kroz mrežu. Svaka primjer (redak epizode) predstavljen je uređenom četvorkom (s, a, s', r).

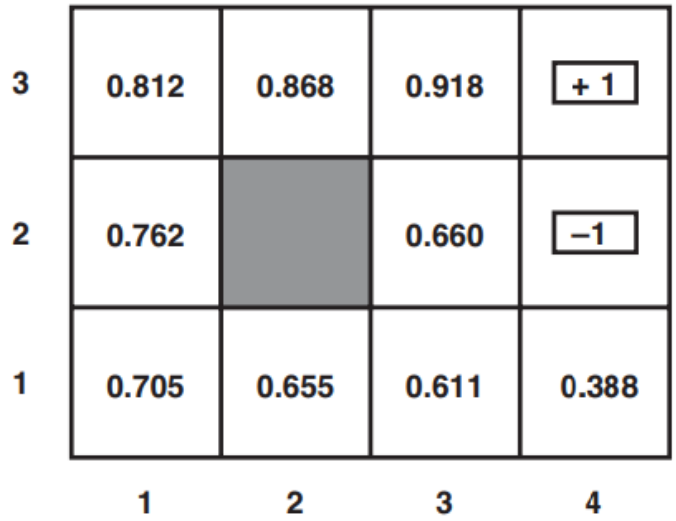
Epizoda 1	Epizoda 2	Epizoda 3
(1,3), J, (1,2), 0	(1,3), J, (1,2), 0	(1,3), J, (1,2), 0
(1,2), I, (2,2), 0	(1,2), I, (2,2), 0	(1,2), I, (2,2), 0
(2,2), J, (2,1), -100	(2,2), I, (3,2), 0	(2,2), I, (3,2), 0
	(3,2), S, (3,3), +100	(3,2), J, (3,1), +80

Koristeći Q-učenje, odredite sljedeće Q-vrijednosti nakon prethodno navedene tri epizode: $Q((3,2),S)$, $Q((1,2),J)$ i $Q((2,2),I)$.

Zadatak 5.

Domaća zadaća

Pretpostavimo da je zadana mreža sa sljedećim vrijednostima:



(a) Za $\gamma = 0.7$ i $\alpha = 0.6$, odredite Q-vrijednosti za $Q((2,1), \rightarrow)$ i $Q((4,1), \uparrow)$ nakon izvršenja epizode: $(1,1), \rightarrow, (2,1), -0.4$; $(2,1), \rightarrow, (3,1), -0.2$; $(3,1), \rightarrow, (4,1), -0.2$; $(4,1), \uparrow, (4,2), 0.1$.

(b) Koristeći TDL i α i γ kao u (a) dijelu zadatka, odredite vrijednosti u stanjima nakon izvršavanja primjera iz (a) dijela zadatka.

Zadatak 6.

Promotrimo sljedeću mrežu (nagrade su prikazane lijevo, a imena stanja desno).

Dan je sljedeći niz primjera, gdje X predstavlja stanje završetka igre:

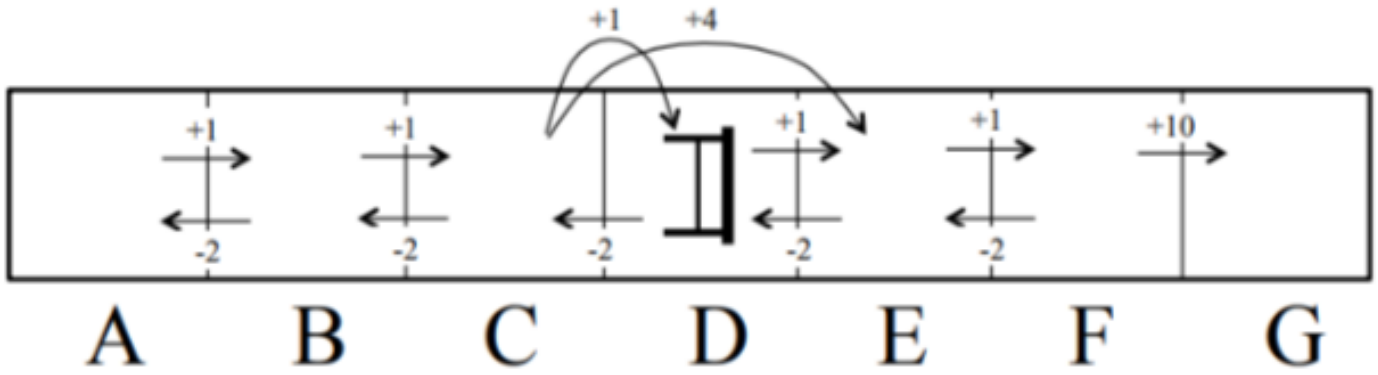
s	a	s'	r	s	a	s'	r
A	$\rightarrow$	R	0	A	$\rightarrow$	R	0
R	izlaz	X	16	R	izlaz	X	16
A	$\leftarrow$	L	0	A	$\leftarrow$	L	0
L	izlaz	X	4	L	izlaz	X	4

(a) Uz faktor umanjenja $\gamma = 1$ i faktor učenja $\alpha = 0.5$, koju bi vrijednost za A vratio TDL?

(b) Nakon izvršenja navedenih primjera, odredite Q vrijednost za $Q(A, \rightarrow)$ uz $\alpha = 0.5$ i $\gamma = 1$.

Zadatak 7.

Promotrimo MPO prikazan na slici koji modelira stazu sa preprekama. Jedina prepreka se nalazi na polju D, dok je ciljno stanje stanje G. Dozvoljene akcije su lijevo i desno. U polju C se ne može izvršiti akcija desno, nego akcija skoči koja rezultira ili uspješnim preskokom na polje E ili padom na prepreku na polju D. Nagrade su prikazane na slici.



Opis akcija: desno - deterministički se pomakni udesno; lijevo - deterministički se pomakni ulijevo; skoči - dozvoljena samo iz C i vrijedi $T(C, \text{skoči}, D) = 0.5$ i $T(C, \text{skoči}, E) = 0.5$.
Neka je dana epizoda:

s	a	s'	r
C	skoči	E	4
E	desno	F	1
F	lijevo	E	-2
E	desno	F	1

Uz $\gamma = 1$ i $\alpha = 0.5$, odredite sve Q vrijednosti nakon izvršenja primjera iz epizode.

Aproksimativno Q-učenje

- koristeći reprezentaciju pomoću značajki možemo zapisati q funkciju (ili funkciju vrijednosti) za svako stanje koristeći nekoliko težina:

$$V(s) = w_1 f_1(s) + w_2 f_2(s) + \cdots + w_n f_n(s)$$
$$Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a) + \cdots + w_n f_n(s,a)$$

- q učenje s linearnom q funkcijom:

- dobijemo novi primjer (s, a, s', r)
- izračunamo razliku $[r + \gamma \max_{a'} Q(s', a')] - Q(s, a)$
- za aproksimativno q-učenje mijenjamo težine:
  $w_i \leftarrow w_i + \alpha[razlika] f_i(s, a)$

Zadatak 8.

Pretpostavimo da se u špilu nalaze karte 2 - 10 i J, Q, K i A te da sve imaju jednaku vjerojatnost da budu izvučene. Svaka od numeriranih karata vrijedi onoliko bodova koji broj piše na njoj, J, Q i K vrijede 10, a A 11. Na svakom potezu može se odabrati akcija povuci (dodatnu kartu) ili ostani (engl. stay). Prilikom akcije povuci se ne dobiva nikakva nagrada, dok se prilikom akcije ostani nagrada računa po pravilu: 0, ako je zbroj bodova na igračevim kartama točno 15; +10, ako je zbroj bodova iz $\langle 15, 21]$ i -10 inače. Nakon izvršenja akcije ostani, igra prelazi u terminalno stanje kraj i završava. Ako je zbroj bodova na igračevim kartama 22 ili više, igrač je u stanju neuspjeh i iz njega može napraviti samo akciju ostani. Skup stanja ove igre je: $\{0, 2, 3, 4, \cdots, 21, neuspjeh, kraj\}$.

(a) Uz danu tablicu početnih Q vrijednosti, faktor umanjenja $\gamma = 1$ i faktor učenja $\alpha = 0.5$, odredite Q vrijednosti za stanja i akcije koje se promijene nakon izvršenja epizode dane u tablici.

s	a	$Q(s, a)$
19	povuci	-2
19	ostani	5
20	povuci	-4
20	ostani	7
21	povuci	-6
21	ostani	8
neuspjeh	ostani	-8

s	a	s'	r
19	povuci	21	0
21	povuci	neuspjeh	0
neuspjeh	ostani	kraj	-10

(b) Sada želimo prikazati $Q(s,a)$ kao linearnu kombinaciju značajki, tj. $Q(s,a) = \sum_i w_i f_i(s,a)$ za neke funkcije značajki $f_i(s,a)$. Neka je zadano

$$f_1(s,a) = \begin{cases} 0, & \text{ako je } a = \text{ostani.} \\ +1, & \text{ako je } a = \text{povuci i } s \geq 15. \\ -1, & \text{ako je } a = \text{povuci i } s < 15. \end{cases}$$

$$f_2(s,a) = \begin{cases} 0, & \text{ako je } a = \text{ostani.} \\ +1, & \text{ako je } a = \text{povuci i } s \geq 18. \\ -1, & \text{ako je } a = \text{povuci i } s < 18. \end{cases}$$

Koju od sljedećih strategija je moguće nedvosmisleno odabrati iz prikaza Q vrijednosti kao linearne kombinacije značajki?

i)

s	$\pi(s)$
14	povuci
15	povuci
16	povuci
17	povuci
18	povuci
19	povuci

ii)

s	$\pi(s)$
14	ostani
15	povuci
16	povuci
17	povuci
18	ostani
19	ostani

iii)

s	$\pi(s)$
14	povuci
15	povuci
16	povuci
17	povuci
18	ostani
19	ostani

iv)

s	$\pi(s)$
14	povuci
15	povuci
16	povuci
17	povuci
18	povuci
19	ostani

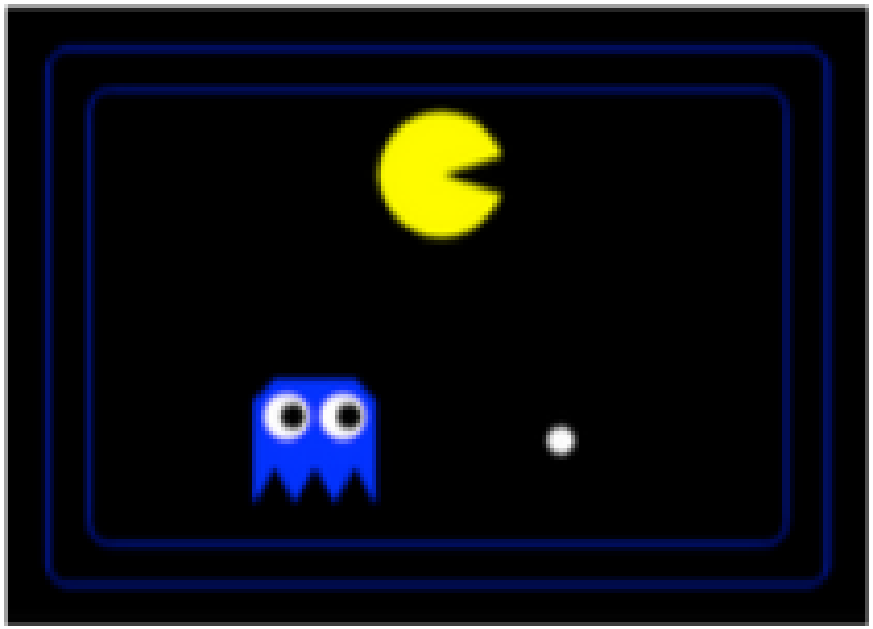
v)

s	$\pi(s)$
14	povuci
15	povuci
16	povuci
17	ostani
18	povuci
19	ostani

Zadatak 9.

Domaća zadaća

Pretpostavite da Pacman u igri želi pojesti hranu i izbjeći duha. Pri tome može izvršiti akcije: ostani (na trenutnoj poziciji), desno, lijevo i dolje. Trenutno stanje igre je prikazano na slici. Mreža je veličine 3×5.



Pretpostavimo da se Q vrijednost aproksimira formulom $Q(s,a) = w_0 f_0(s,a) + w_1 f_1(s,a)$, gdje je:

$$f_0 = 1/(\text{Manhattan udaljenost do najbliže hrane + 1}) \text{ i}$$

$$f_1 = 1/(\text{Manhattan udaljenost do najbližeg duha + 1}).$$

Pacman odabire akciju na temelju $\arg\max_a Q(s,a) = \arg\max_a w^T f(s,a)$, gdje je w vektor težina.

(a) Uz vektor težina $w = [0.2, 0.5]$ koju bi od dozvoljenih akcija Pacman odabrao ako kreće iz stanja sa slike?

(b) Uz vektor težina $w = [0.2, -1]$ koju bi od dozvoljenih akcija Pacman odabrao ako kreće iz stanja sa slike?

Zadatak 10.

Domaća zadaća

Neka je dano stanje kao na slici i neka je Q vrijednost reprezentirana kao: $Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a)$, gdje je:

$$f_1(s,a) = 1/(\text{Manhattan udaljenost do najbliže hrane}) \text{ i}$$

$$f_2(s,a) = \text{Manhattan udaljenost do najbližeg duha.}$$

(a) Uz težine $w_1 = 1$ i $w_2 = 10$, koju bi akciju Pacman odabrao: dolje ili lijevo?

(b) Nakon izvršenja akcije iz (a) dijela zadatka nagrada je $r = 9$. Iz tog novog stanja izračunajte Q vrijednost pomicanja desno i lijevo, zatim vrijednost danog primjera te sa svim dobivenim informacijama nove težine nakon izvršenja akcije iz (a) dijela zadatka uz faktor umanjenja $\gamma = 1$ i faktor učenja $\alpha = 0.5$.