Izveštaj o petom domaćem zadatku iz predmeta Mašinsko učenje

Viktor Todosijević 3140/2021

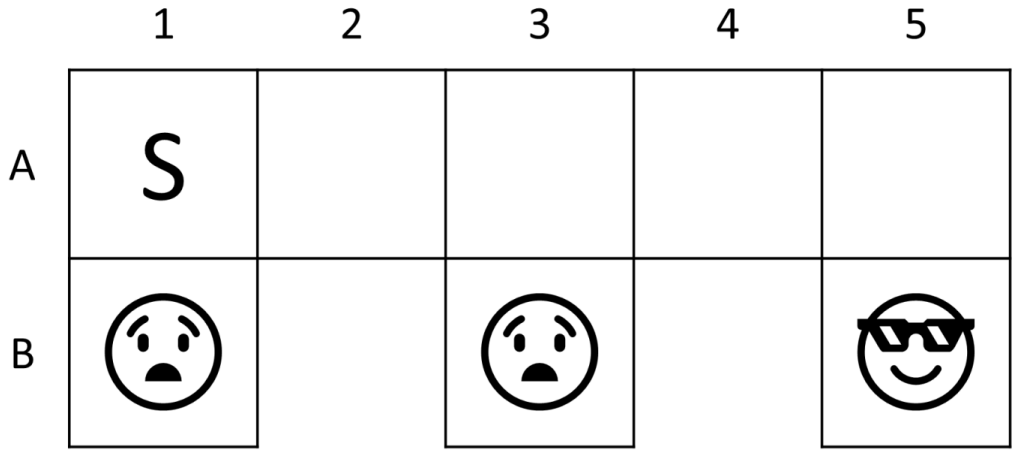
Decembar 2022

1 Uvod

Tema petog domaćeg zadatka iz predmeta Mašinsko učenje je "Učenje podsticanjem". Student je u programskom jeziku Python uspešno implementirao zadato okruženje(odeljak 2) i algoritam Q-učenja(odeljak 3). Student nije uspeo da implementira algoritam REINFORCE.

2 Okruženje

Okruženje u kome će algoritmi biti obučavani (slika 1) je jednostavni dvodimenzioni lavirint sa početnim stanjem $S = (A, 1)$, terminalnim stanjima $T_1 = (B, 1),\ T_2 = (B, 3),\ T_3 = (B, 5)$ i nagradama u tim stanjima $R_1 = -1,\ R_2 = -1,\ R_3 = 3$. Ostala stanja ne donose nagradu. Agent koji se nadje u ovom okruženju ima na raspolaganju četiri moguće akcije: levo, desno, gore i dole. Ako agent načini akciju koja bi ga odvela van okruženja, onda on ostaje u trenutnom stanju. Pored toga, okruženje je stohatično i to na sledeći način: kada se agent odluči za akciju ima 60% šanse da zapravo i načini tu akciju, a 40% da ode u jednom od dva pravca ortogonalna prvobitnoj odluci. U daljem tekstu će stanje u kom se našao agent u koraku t obeležavati sa s, akcija u koraku t sa a, a stanje u koraku $t + 1$ sa s', a nagrada za prelazak iz stanja s u s' sa $R(s)$.



Slika 1: Okruženje.

3 Q-učenje

Q-učenje je algoritam učenja sa posticanjem koji nastoji da nauči optimalnu Q-funkciju $Q^*(s,a)$ okruženja na sledeći način: za svaku uredjenu trojku (s,a,s') dobijenu putem politike ponašanja $\beta(s)$ vrši se korekcija procene

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

gde je $\gamma \in [0,1]$ faktor umanjenja budućih nagrada, a α stopa učenja. Za naučenu optimalnu Q-funkciju $Q^*(s,a)$ optimalna politika je

$$\pi^*(s) = \arg\max_a Q^*(s,a) \tag{2}$$

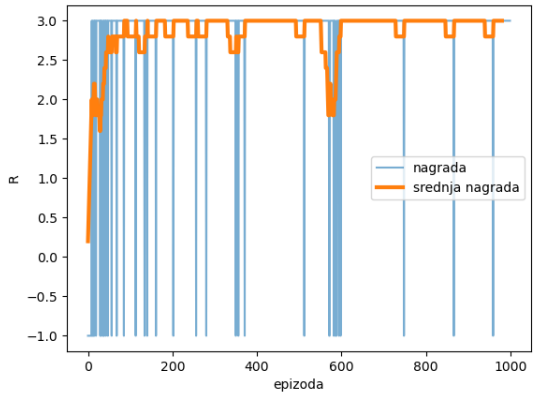
Iz jednačine 2 se vidi da je politika Q-učenja deterministička. Kada bi takvu politiku koristili za istraživanje okruženja veliki deo prostora stanja bi mogao ostati neistražen. Iz tog razloga uvodimo politiku ponašanja $\beta(s)$ koja kaže da u koraku t sa verovatnoćom ϵ_t biramo nasumičnu akciju, a sa verovatnoćom $1-\epsilon_t$ biramo akciju $\pi(s)$ gde je $\epsilon_t = \epsilon_0^t$ za $t \geq 0$.

3.1 Metodologija

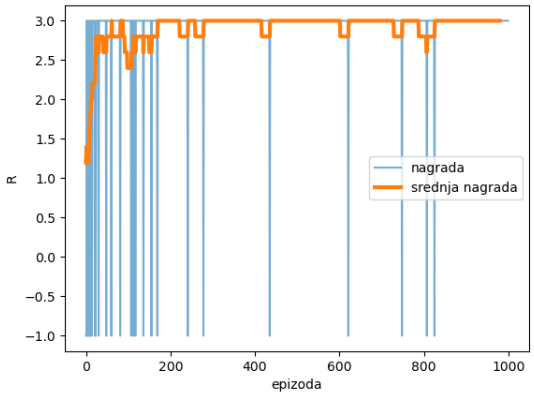
Sprovodimo obučavanje agenta u implementiranom okruženju za dve vrednosti faktora umanjenja budućih nagrada $\gamma \in \{0.9, 0.999\}$ i za dve strategije podešavanja stope učenja α: konstantna stopa učenja i $\alpha = \alpha_e = \frac{ln(e+1)}{e+1}$. U svim eksperimentima koristimo $\epsilon_0 = 0.97$ i obučavamo 1000 epizoda.

3.2 Rezultati

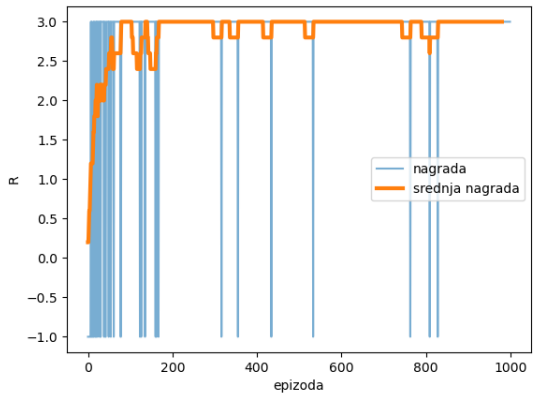
Na slici 2 vidimo da nagrada i srednja nagrada koju agent dobije u svakoj epizodi ne zavisi značajno od γ, dok za stopu učenja koja počinje od 1 i polako se spušta vidimo da se ne spušta dovoljno brzo što unosi šum u obučavanje. Na slici REF vidimo da su za γ bliže vrednosti 1 funkcije vrednosti u različitim stanjima bliže jedna drugoj što može otežati procenu toga da li je jedno stanje bolje od drugog. Za stopu učenja koja počinje od 1 i spušta se tokom obučavanja vidimo da su procene funkcije vrednosti mnogo šumovitije.



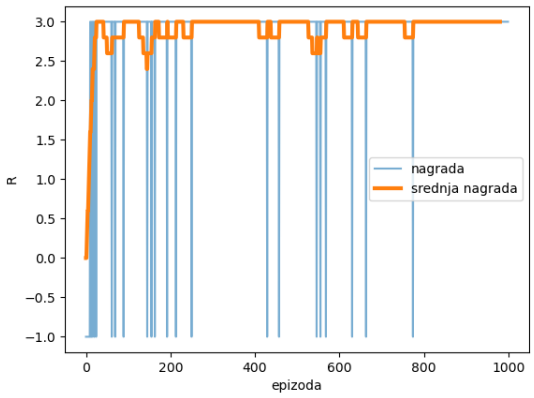
(a) $\gamma = 0.9,\ \alpha_e = \frac{ln(e+1)}{e+1}$



(b) $\gamma = 0.9,\ \alpha = 0.01$

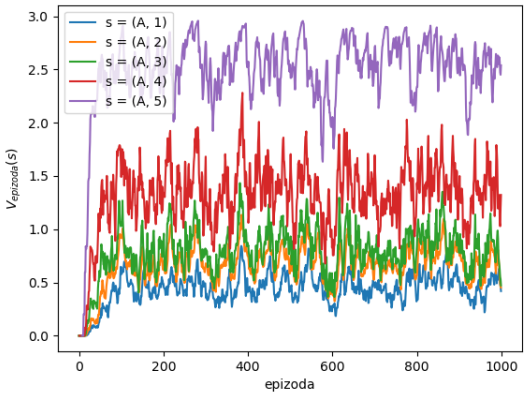


(c) $\gamma = 0.999,\ \alpha_e = \frac{ln(e+1)}{e+1}$

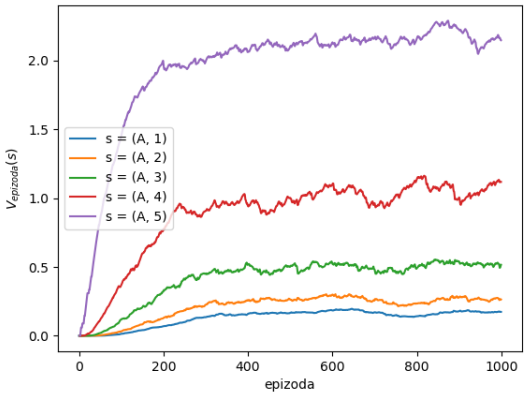


(d) $\gamma = 0.999,\ \alpha = 0.01$

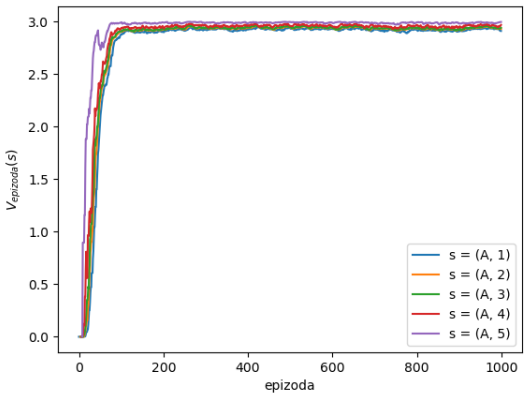
Slika 2: Nagrada i srednja nagrada tokom obučavanja.



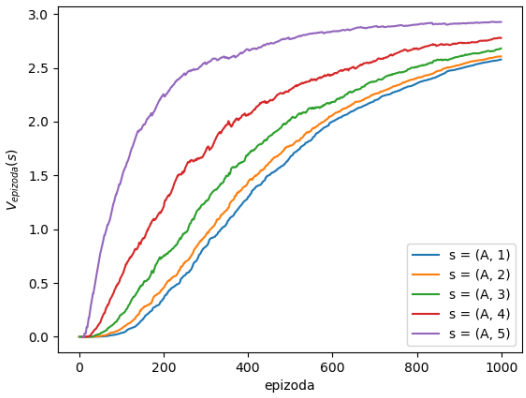
(a) $\gamma = 0.9,\ \alpha_e = \frac{ln(e+1)}{e+1}$



(b) $\gamma = 0.9,\ \alpha = 0.01$



(c) $\gamma = 0.999,\ \alpha_e = \frac{ln(e+1)}{e+1}$



(d) $\gamma = 0.999,\ \alpha = 0.01$

Slika 3: Fukncija vrednost $V(s)$ u funkciji epizode za svako stanje koje nije terminalno.