ДОМАШНА ЗАДАЧА 2

АГЕНТНО-БАЗИРАНИ СИСТЕМИ

УЧЕЊЕ СО ПОТТИКНУВАЊЕ

1. Пакман на прошетка

Замислете дека Пакмен ја учи оптималната политика во лавиринтот прикажан на сликата. Состојбите кои се обоени со црвена или зелена боја се терминалните состојби т.е. состојби во кои играта завршува (се изведува акцијата Exit), при што Пакмен ја добива наградата/казната означена во полето. Во останатите состојби (полиња од лавиринтот) Пакман може да ги користи акциите Исток (E), Запад (W), Север (N) и Југ (S), при што детерминистички се поместува за едно поле. Во случај да изведе акција која го води надвор од границите на лавиринтот, Пакмен ќе остане во истата состојба. Факторот на намалување е γ = 0.5, а ратата на учење е 0.5. Почетната позиција на Пакман е состојбата (1,3).

За дадениот проблем за чие решение користите Q-learning, потребно е да одговорите на следните прашања:

а) Да се пресметаат оптималните вредности на состојбите V*(2,2), V*(3,2) и V*(1,3). Образложете ги пресметките преку формулите кои сте ги користеле.

(б) Доколку Пакман започнува во почетната позиција во горниот лев агол и има три епизоди искуство на движење низ лавиринтот, да се пресметаат вредностите на следните три Q-состојби: Q((3,2),N), Q((1,2),S) и Q((2,2),E).

Да се користи алгоритамот за итеративно ажурирање (Q-learning). Секој ред во табелата со епизоди дадени подолу ги претставува торките (s, a, s', r) т.е. моментална состојба во која се наоѓа Пакман, акција која ја презема, состојба во која преминува и награда која ја добива. Образложете ги пресметките преку формулите кои сте ги користеле.

Епизода 1	***Епизода 2***	***Епизода 3***
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
	(3,2), N, (3,3), +100	(3,2), S, (3,1), +80

в) Доколку се користи апроксимативно репрезентација на вредностите на Q-состојбата како линеарна комбинација на следните 3 карактеристики $f_1(s)$ $f_2(s)$ и $f_3(s)$. Првите две карактеристики, $f_1(s)$ и $f_2(s)$ ги претставуваат x- и y-координатите на дадена состојба s, соодветно.

$$Q(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(s)$$

Третата карактеристика добива вредности според акцијата како што е претставено

$$f_3(N)=1 \qquad f_3(S)=2 \qquad f_3(E)=3 \qquad f_3(W)=4$$

i) Доколку сите тежини w_i се иницијално поставено на 0, кои ќе бидат нивните ажурирани вредности после првата епизода. Образложете го вашето решение преку формулите за пресметка.
ii) Доколку векторот на тежини има вредност w = (1,1,1), која акција би била оптимална според апроксимативното Q-учење за состојбата (2,2). Образложете го вашето решение.

2. Пакмен во "Вртлог"

Пакмен се наоѓа во лавиринт со изглед прикажан на сликата. Почетната позиција на Пакман е состојбата **S**, додека терминалните состојби т.е. состојби во кои играта завршува (единствено дозволена акција е **Exit**), се означени со наградите кои Пакмен ги добива по изведување на акцијата за излез. Во состојбата **A**, "вртлог", единствена акција која може да се изведе е **Escape**, која го носи Пакмен во едно од соседните полиња, со рамномерно распределена веројатност. Во состојбата **S** единствено дозволена акција е Десно (**R**), при што Пакмен детерминистички се поместува за едно поле надесно. Факторот на намалување е γ = 1, а ратата на учење е 0.5.

За даденият проблем, потребно е да одговорите на следните прашања:

а) Која е оптималната вредност на состојбите **S** и Вртлог?

Во табелите се дадени две секвенците на премин означени со T1 и T2:

T1

s	***a***	***s'***	***r***
S	R	A	0
A	Escape	E1	0
E1	Exit	-	1
S	R	A	0
A	Escape	E10	0
E10	Exit	-	10

T2

s	***a***	***s'***	***r***
S	R	A	0
A	Escape	E1	0
E1	Exit	-	1
S	R	A	0
A	Escape	E10	0
E10	Exit	-	10
S	R	A	0
A	Escape	E10	0
E10	Exit	-	10

б) Доколку се изведува секвенцата T1 бесконечно, кон кои вредности конвергираат Q-состојбите $Q^{T1}(S, R)$ и $Q^{T1}(A, Escape)$?

в) Доколку се изведува секвенцата T2 бесконечно, кон кои вредности конвергираат Q-состојбите $Q^{T2}(S, R)$ и $Q^{T2}(A, Escape)$?

г) Која вредност на Q-состојбата $Q^{*}(S, R)$ е оптимална: $Q^{T1}(S, R)$ или $Q^{T2}(S, R)$?