Агентно-базирани системи

Домашна задача 2

Владимир Христовски – 223030

1. Пакман на прошетка

Состојби:

- (X, Y) - (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3).

- Почетна состојба – (1, 3)

Дозволени акции:

- Exit - Излез

- E – Исток

- W – Запад

- N – Север

- S – Југ

Награди:

- R(RedField) = -80 / -100

- R(GreenField) = +100 / +80 / +25

- R(NormalField) = 0

Фактор на намалување:

$\gamma = 0.5$

Рата на учење:

$\alpha = 0.5$

А:

Q-learning формула за пресметка на Q состојбите за секоја акција

$Q(s,a) = Q(s,a) + \alpha * (R(s') + \gamma * \max Q(s',a') - Q(s,a))$

V* формула за наоѓање на најоптималната Q состојба

$V^*(s)=\max Q(s,a)$

Q(X, Y, Action) = 0 - иницијализација

$V^*(2, 2) = \max\{Q(2, 2, E), Q(2, 2, W), Q(2, 2, N), Q(2, 2, S)\} = \max(0 , 0, -40, -50) = 0$

$Q(2, 2, E) = Q(2, 2, E) + \alpha * (R(3, 2) + \gamma * \max Q(3, 2) - Q(2, 2, E))$

$= 0 + 0.5 * (0 + 0.5 * 0 - 0) = 0 + 0.5 * 0 = 0$

$Q(2, 2, W) = Q(2, 2, W) + \alpha * (R(1, 2) + \gamma * \max Q(1, 2) - Q(2, 2, W))$

$= 0 + 0.5 * (0 + 0.5 * 0 - 0) = 0 + 0.5 * 0 = 0$

$Q(2, 2, N) = Q(2, 2, N) + \alpha * (R(2, 3) + \gamma * \max Q(2, 3) - Q(2, 2, N))$

$= 0 + 0.5 * (-80 + 0.5 * 0 - 0) = 0 + 0.5 * -80 = -40$

$Q(2, 2, S) = Q(2, 2, S) + \alpha * (R(2, 1) + \gamma * \max Q(2, 1) - Q(2, 2, S))$

$= 0 + 0.5 * (-100 + 0.5 * 0 - 0) = 0 + 0.5 * -100 = -50$

$V^*(3, 2) = \max\{Q(3, 2, E), Q(3, 2, W), Q(3, 2, N), Q(3, 2, S)\} = \max(0, 0, 50, 40) = 50$

$Q(3, 2, E) = Q(3, 2, E) + \alpha * (R(3, 2) + \gamma * \max Q(3, 2) - Q(3, 2, E))$

$= 0 + 0.5 * (0 + 0.5 * 0 - 0) = 0 + 0.5 * 0 = 0$

$Q(3, 2, W) = Q(3, 2, W) + \alpha * (R(2, 2) + \gamma * \max Q(2, 2) - Q(3, 2, W))$

$= 0 + 0.5 * (0 + 0.5 * 0 - 0) = 0 + 0.5 * 0 = 0$

$Q(3, 2, N) = Q(3, 2, N) + \alpha * (R(3, 3) + \gamma * \max Q(3, 3) - Q(3, 2, N))$

$= 0 + 0.5 * (+100 + 0.5 * 0 - 0) = 0 + 0.5 * +100 = +50$

$Q(3, 2, S) = Q(3, 2, S) + \alpha * (R(3, 1) + \gamma * \max Q(3, 1) - Q(3, 2, S))$

$= 0 + 0.5 * (+80 + 0.5 * 0 - 0) = 0 + 0.5 * +80 = +40$

$V^*(1, 3) = \max\{Q(1, 3, E), Q(1, 3, W), Q(1, 3, N), Q(1, 3, S)\} = \max(-40, 0, 0, 0) = 0$

$Q(1, 3, E) = Q(1, 3, E) + \alpha * (R(2, 3) + \gamma * \max Q(2, 3) - Q(1, 3, E))$

$= 0 + 0.5 * (-80 + 0.5 * 0 - 0) = 0 + 0.5 * -80 = -40$

$Q(1, 3, W) = Q(1, 3, W) + \alpha * (R(1, 3) + \gamma * \max Q(1, 3) - Q(1, 3, W))$

$= 0 + 0.5 * (0 + 0.5 * 0 - 0) = 0 + 0.5 * 0 = 0$

$Q(1, 3, N) = Q(1, 3, N) + \alpha * (R(1, 3) + \gamma * \max Q(1, 3) - Q(1, 3, N))$

$= 0 + 0.5 * (0 + 0.5 * 0 - 0) = 0 + 0.5 * 0 = 0$

$Q(1, 3, S) = Q(1, 3, S) + \alpha * (R(1, 2) + \gamma * \max Q(1, 2) - Q(1, 3, S))$

$= 0 + 0.5 * (0 + 0.5 * 0 - 0) = 0 + 0.5 * 0 = 0$

Б:

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

$Q(s, a) = (1 - \alpha) * Q(s, a) + \alpha * (R + \gamma * \max Q(s + 1, a'))$

$Q((3, 2), N) = 50$

$Q((1, 2), S) = 0$

$Q((2, 2), E) = 12.5$

Епизода 1:

$Q((1, 3), S) = (1 - \alpha) * Q((1, 3), S) + \alpha * (R + \gamma * \max Q((1, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 0) = 0 + 0.5 * 0 = 0$

$Q((1, 2), E) = (1 - \alpha) * Q((1, 2), E) + \alpha * (R + \gamma * \max Q((2, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 0) = 0 + 0.5 * 0 = 0$

$Q((2, 2), S) = (1 - \alpha) * Q((2, 2), S) + \alpha * (R + \gamma * \max Q((2, 1), a'))$

$= 0.5 * 0 + 0.5 * (-100 + 0.5 * 0) = 0 + 0.5 * -100 = -50$

Епизода 2:

$Q((1, 3), S) = (1 - \alpha) * Q((1, 3), S) + \alpha * (R + \gamma * \max Q((1, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 0) = 0 + 0.5 * 0 = 0$

$Q((1, 2), E) = (1 - \alpha) * Q((1, 2), E) + \alpha * (R + \gamma * \max Q((2, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 0) = 0 + 0.5 * 0 = 0$

$Q((2, 2), E) = (1 - \alpha) * Q((2, 2), E) + \alpha * (R + \gamma * \max Q((3, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 0) = 0 + 0.5 * 0 = 0$

$Q((3, 2), N) = (1 - \alpha) * Q((3, 2), N) + \alpha * (R + \gamma * \max Q((3, 3), a'))$

$= 0.5 * 0 + 0.5 * (+100 + 0.5 * 0) = 0 + 0.5 * 100 = 50$

Епизода 3:

$Q((1, 3), S) = (1 - \alpha) * Q((1, 3), S) + \alpha * (R + \gamma * \max Q((1, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 0) = 0 + 0.5 * 0 = 0$

$Q((1, 2), E) = (1 - \alpha) * Q((1, 2), E) + \alpha * (R + \gamma * \max Q((2, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 0) = 0 + 0.5 * 0 = 0$

$Q((2, 2), E) = (1 - \alpha) * Q((2, 2), E) + \alpha * (R + \gamma * \max Q((3, 2), a'))$

$= 0.5 * 0 + 0.5 * (0 + 0.5 * 50) = 0 + 0.5 * 25 = 12.5$

$Q((3, 2), S) = (1 - \alpha) * Q((3, 2), S) + \alpha * (R + \gamma * \max Q((3, 1), a'))$

$= 0.5 * 0 + 0.5 * (+80 + 0.5 * 0) = 0 + 0.5 * 80 = 40$

в:

f1(s) – x – координати

f2(s) – y – координати

f3(N) = 1

f3(S) = 2

f3(E) = 3

f3(W) = 4

Формула за апроксимативно Q-учење

Q(s, a) = w1 * f1(s) + w2 * f2(s) + w3 * f3(s)

i:

w1 = 0

w2 = 0

w3 = 0

Епизода 1:

f1	f2	f3
f1((1, 3), S) = 1	f2((1, 3), S) = 3	f3((1, 3), S) = 2
f1((1, 2), E) = 1	f2((1, 2), E) = 2	f3((1, 2), E) = 3
f1((2, 2), S) = 2	f2((2, 2), S) = 2	f3((2, 2), S) = 2

Формула за ажурирање на тежините w:

$Wi = wi + \alpha * (R(s') + \gamma * maxQ(s',a') - Q(s,a)) * fi(s, a)$

Транзиција 1:

w1 = w1 + α * (R(1, 2) + γ * maxQ(1, 2) − Q(1, 3)) * f1(1, 3)

= 0 + 0.5 * (0 + 0.5 * 0 - 0) * 1 = 0 + 0.5 * (0 + 0 – 0) * 1= 0 + 0.5 * 0 * 1= 0

w2 = w2 + α * (R(1, 2) + γ * maxQ(1, 2) − Q(1, 3)) * f2(1, 3)

= 0 + 0.5 * (0 + 0.5 * 0 - 0) * 3 = 0 + 0.5 * (0 + 0 – 0) * 3 = 0 + 0.5 * 0 * 3 = 0

w3 = w3 + α * (R(1, 2) + γ * maxQ(1, 2) − Q(1, 3)) * f3(1, 3)

= 0 + 0.5 * (0 + 0.5 * 0 - 0) * 2 = 0 + 0.5 * (0 + 0 – 0) * 2 = 0 + 0.5 * 0 * 2 = 0

Транзиција 2:

w1 = w1 + α * (R(2, 2) + γ * maxQ(2, 2) − Q(1, 2)) * f1(1, 2)

= 0 + 0.5 * (0 + 0.5 * 0 - 0) * 1 = 0 + 0.5 * (0 + 0 – 0) * 1= 0 + 0.5 * 0 * 1= 0

w2 = w2 + α * (R(2, 2) + γ * maxQ(2, 2) − Q(1, 2)) * f2(1, 2)

= 0 + 0.5 * (0 + 0.5 * 0 - 0) * 2 = 0 + 0.5 * (0 + 0 – 0) * 2 = 0 + 0.5 * 0 * 2 = 0

w3 = w3 + α * (R(2, 2) + γ * maxQ(2, 2) − Q(1, 2)) * f3(1, 2)

= 0 + 0.5 * (0 + 0.5 * 0 - 0) * 3 = 0 + 0.5 * (0 + 0 – 0) * 3 = 0 + 0.5 * 0 * 3 = 0

Транзиција 3:

w1 = w1 + α * (R(2, 1) + γ * maxQ(2, 1) − Q(2, 2)) * f1(2, 2)

= 0 + 0.5 * (-100 + 0.5 * 0 - 0) * 2 = 0 + 0.5 * (-100 + 0 – 0) * 2= 0 + 0.5 * -100 * 2= -100

w2 = w2 + α * (R(2, 1) + γ * maxQ(2, 1) − Q(2, 2)) * f2(2, 2)

= 0 + 0.5 * (-100 + 0.5 * 0 - 0) * 2 = 0 + 0.5 * (-100 + 0 – 0) * 2 = 0 + 0.5 * -100 * 2 = -100

w3 = w3 + α * (R(2, 1) + γ * maxQ(2, 1) − Q(2, 2)) * f3(2, 2)

= 0 + 0.5 * (-100 + 0.5 * 0 - 0) * 2 = 0 + 0.5 * (-100 + 0 – 0) * 2 = 0 + 0.5 * -100 * 2 = -100

ii:

w1 = 1

w2 = 1

w3 = 1

Q((2, 2), N) = w1 * f1(2, 2) + w2 * f2(2, 2) + w3 * f3(N) = 1 * 2 + 1 * 2 + 1 * 1 = 2 + 2 + 1 = 5

Q((2, 2), S) = w1 * f1(2, 2) + w2 * f2(2, 2) + w3 * f3(S) = 1 * 2 + 1 * 2 + 1 * 2 = 2 + 2 + 2 = 6

Q((2, 2), E) = w1 * f1(2, 2) + w2 * f2(2, 2) + w3 * f3(E) = 1 * 2 + 1 * 2 + 1 * 3 = 2 + 2 + 3 = 7

Q((2, 2), W) = w1 * f1(2, 2) + w2 * f2(2, 2) + w3 * f3(W) = 1 * 2 + 1 * 2 + 1 * 4 = 2 + 2 + 2 = 8

Q*((2,2), a) = max{ Q((2, 2), N), Q((2, 2), S), Q((2, 2), E), Q((2, 2), W)} = max(5, 6, 7, 8) = 8

Оптимална акција би била акцијата W – запад поради најголемата Q вредност за таа состојба, Q((2, 2), W) = 8.

2. Пакмен во “Вртлог”

Состојби:

- S – почетна состојба

- A - вртлог

- E1 – терминална состојба

- E10 – терминална состојба

Дозволени акции:

Exit – единствена акција за излез од терминалните состојби

Escape – од полето A кон S, E1 или E2 со рамномерно распределена веројатност

R – единствена дозволена акција во состојба S, поместување во десно

Награди:

- R(ExitE1) = +1
- R(ExitE10) = +10
- R(S/A) = 0

Фактор на намалување:

γ = 1

Рата на учење:

α = 0.5

A:

V* = maxQ*(s, a)

Q(s, a) = R(s, a) + γ * sum(P(s' | s, a) * V(s'))

Q*(A, Escape) = R(A, Escape) + 1 * sum(V*(S) / 3 + V*(E1) / 3 + V*(E10) / 3)

= 0 + V*(S) / 3 + V*(E1) / 3 + V*(E10) / 3 = V*(S) / 3 + 1 / 3 + 10 / 3 = V*(S) / 3 + 0.333 + 3.333

= V*(S) / 3 + 3.666

Q*(S, R) = R(S, R) + 1 * sum(1 * V(A)) = 0 + 1 * sum(1 * V*(A)) = 0 + V*(A)

$Q^*(A, Escape) = V^*(A) / 3 + 3.666$

$V^*(A) – V^*(A) / 3 = 3.666$

$2 * V^*(A) / 3 = 3.666$

$V^*(A) = 3.666 * 3 / 2 = 5.5$

$V^*(A) = 5.5$

$V^*(S) = 5.5$

Б:

$Q*(s,a)=s'\sum P(s'|s,a)\cdot[r(s,a,s')+\gamma\cdot a'maxQ*(s',a')]$

Според моето разбирање на задачата, од полето A (“Вртлог”), може да се премине на едно од соседните полиња (S, E1 и E10) со рамномерно распределена веројатност (0,(3) по поле). Во секвенцата T1 има 2 епизоди, едната завршува во E1, другата во E10. Според оваа секвенца агентот нема да знае за можниот премин од A (“Вртлог”) во S. Со ова можеме да заклучиме дека “научените” премини од T1, E1 и E10 од A ќе имаат рамномерно распределена веројатност (0.5 по поле).

Доколку T1 се извршува бесконечно многу пати, Q вредностите за $Q^{T1}(S,R)$ и

$Q^{T1}(A, Escape)$ би биле:

$Q^{T1}(A, Escape) = 0.5 * 1 + 0.5 * 10 = 0.5 + 5 = 5.5$

Бидејќи од состојба S можеме да преземеме само една акција, R – десно кон A со што S директно зависи од акцијата Escape, Q вредноста Q(S, R) = V*(A), пресметано погоре во делот под A

$Q^{T1}(S, R) = V^*(A) = 5.5$

В:

Во секвенцата Т2 има 3 епизоди, една завршува во Е1, останатите две во Е2. Исто како и во барањето под Б, агентот нема да знае дека има можен премин кон S од полето А (“Вртлог”).

Доколку Т2 се извршува бесконечно многу пати, Q вредностите за $Q^{T2}(S,R)$ и

$Q^{T2}(A, Escape)$ би биле:

$Q^{T2}(A, Escape) = 1 * 1 / 3 + 2 * 10 / 3 = 1 / 3 + 20 / 3 = 0,(3) + 6,(6) = 6,(9) = 7$

За Q(S, R) важи истото од решението под Б:

$Q^{T2}(S, R) = V^*(A) = 7$

Г:

Доколку ги анализираме двете решенија доаѓаме до заклучок дека Q^{T1} е добиена со тоа што во секвенцата Т1 распределбата на веројатностите е рамномерна (еднаш заврши во Е1, еднаш во Е10), додека во секвенцата Т2 распределбата не веројатностите не е рамномерна (еднаш заврши во Е1, два пати во Е10), со што соодветно се добиени овие Q вредности за истите.

Оптималната вредност зависи од вистинската очекувана награда, што во случајов бидејќи Q(S, R) има директно дејство од Q(A, Escape), при што Escape има рамномерна распределба кон останатите полиња.

Доколку се водиме според политиката за добивање на поголема добивка, $Q^{T2}(S, R)$ е повеќе оптимална со наклонетост кон Е10.

На крајот се зависи од сознанијата на моделот, односно што тој ќе научи. Според наученото се доаѓа до Q вредност. Ако ние прогласиме една Q вредност за оптимална за модел учен со една секвенца, за истиот модел учен со друга секвенца тој може да биде не оптимален (преценета или потценета Q вредност).

Доколку разгледуваме детерминистички пристап со еднакво распределени веројатности, односно од состојба А со акцијата Escape да се стига до сите три соседни полиња по ист број на пати, оптимална Q вредност би била $Q^{T1}(S, R)$. За тоа е доказ пресметката во решението под А каде што се земени сите 3 можни акции.