فرهاد امان ۹۹۳۱۰۰۶

۱-

الف)

Wt < 0 : i
Wt > : iii

ب)

Wt < 0 : iv

ج)

Wt < 0 : v

د)

f=|y-y0|
Wt < 0 : ii
Wt > 0 : iv

۲-

V1(X0) = max(0.5 * (0 + 0)), 1 * (6 + 0) + 0.5 * (6 + 0)) = 9

V1(X1) = max(1 * (0 + 0) + 0.5 * (0 + 0) + 0.5 * (0 + 0), 0.5 * (0 + 0), 0.5 * (0 + 0)) = 0

V1(X2) = max(1 * (5 + 0) + 0.5 * (5 + 0) + 0.5 * (5 + 0), 0.5 * (0 + 0), 0.5 * (0 + 0)) = 10

V1(X3) = max(1 * (24 + 0) + 0.5 * (24 + 0), 0.5 * (0 + 0)) = 36

V2(X0) = max(1 * (6 + 0.5 * 9) + 0.5 * (6 + 0.5 * 9), 0.5 * (0 + 0)) = 15.75

V2(X1) = max(1 * (0 + 0) + 0.5 * (0 + 0) + 0.5 * (0 + 0), 0.5 * (0 + 0.5 * 9), 0.5 * (0 + 0.5 * 10)) = 2.5

V2(X2) = max(1 * (5 + 0.5 * 10) + 0.5 * (5 + 0.5 * 10) + 0.5 * (5 + 0.5 * 10), 0.5 * (0 + 0), 0.5 * (0 + 0.5 * 36)) = 40

V2(X3) = max(1 * (24 + 0.5 * 36) + 0.5 * (24 + 0.5 * 36), 0.5 * (0 + 0.5 * 10)) = 63

۳-

الف)

Vπ(si) = i, i = 3, 4, 5, 6

V1π(s1) = V1π(s2) = (1/6) * (-1 + 0) + (1/6) * (-1 + 0) + (1/6) * (-1 + 0) + (1/6) * (-1 + 0) + (1/6) * (-1 + 0) + (1/6) * (-1 + 0) = (1/6) * (-6)

V2π(s1) = V2π(s2) = (1/6) * (-1 + (1/6) * (-6)) + (1/6) * (-1 + (1/6) * (-6)) + (1/6) * (-1 + 3) + (1/6) * (-1 + 4) + (1/6) * (-1 + 5) + (1/6) * (-1 + 6) = (1/6 + 2/6^2) * (-6) + (1/6) * 18

V3π(s1) = V3π(s2) = (1/6) * (-1 + (1/6 + 2/6^2) * (-6) + (1/6) * 18) + (1/6) * (-1 + (1/6 + 2/6^2) * (-6) + (1/6) * 18) + (1/6) * (-1 + 3) + (1/6) * (-1 + 4) + (1/6) * (-1 + 5) + (1/6) * (-1 + 6) = (1/6 + 2/6^2 + 2^2/6^3) * (-6) + (1/6 + 2/6^2) * 18

=> Vπ(s1) = Vπ(s2) = 12/4 = 3

ب) ارزش مورد انتظار پس از انجام عملیات تاس برای هر حالت s:

3 = ((1+6-) + (1+5-) + (1+4-) + (1+3-) + (1+3-) + (1+3-))/6

حالت	s1	s2	s3	s4	s5	s6
pi0	تاس	تاس	تمام	تمام	تمام	تمام
pi1	تاس	تاس	تمام/تاس	تمام	تمام	تمام

ج) بله همگرا می شوند کافیست برای حالت ۳ اتمام انتخاب شود.

۴-
الف)
نادرست
چون ممکن است در حالتی چند اکشن ما را به حداکثر ارزش برساند و دو سیاست مختلف هردو بهینه باشند ولی متفاوت چون در این حالت اکشن متفاوتی از بین این اکشن ها انجام دهند.

ب)
درست
برای مثال در یک MDP که دو حالت ترمینال A و B دارد، به طوری که پاداش لحظه‌ای رفتن به A برابر 1 و پاداش لحظه‌ای رفتن به B برابر 10 باشد و بقیه پاداش‌های لحظه‌ای صفر باشند، اگر فاصله A از مبدأ یک قدم به شمال و فاصله B از مبدأ دو قدم به جنوب باشد، در صورتی که گاما (γ) کمتر از 0.1 باشد، سیاست بهینه عامل را به شمال و اگر بیشتر باشد، عامل را به جنوب می‌برد.

ج)
بله
چون پس از گذر زمان، فرایند یادگیری بهتر شده و exploration دیگر مهم نیست.

د)
عامل: رباتی که در یک ماز در حال حرکت است.
محیط: ماز با دیوارها و موانع.
توضیح: در این مثال بهتر است که عامل مبتنی بر مدل باشد. عامل با داشتن یک مدل دقیق از ماز و موانع آن، می تواند مسیر خود را با کارایی بیشتری برنامه ریزی کند و از برخوردهای غیر ضروری جلوگیری کند. با یک مدل، عامل می‌تواند سناریوهای مختلف را در ماز مجازی شبیه‌سازی کند، از آنها یاد بگیرد و عملکرد خود را بدون نیاز به کاوش فیزیکی در ماز واقعی بهبود بخشد.

عامل: یک بازیکن هوشمند در بازی شطرنج. محیط: محیط بازی شطرنج با قوانین مشخص. توضیح: در این مثال، بهتر است عامل بر پایه مدل نباشد و بر اساس تجربه عمل کند. چرا که بازی شطرنج قوانین دقیق و قابل پیش‌بینی دارد و استفاده از مدل نمی‌تواند اطلاعات بیشتری در اختیار عامل قرار دهد. به جای آن، بازیکن می‌تواند از تجربه خود و تجربه حریفانش در بازی شطرنج برای ارزیابی حرکات و انتخاب بهترین راهبرد استفاده کند.

عامل: ربات خودران در محیط شهری پرترافیک. محیط: محیط شهری با خیابان‌ها، تقاطع‌ها، خودروها و پیاده‌روها. توضیح: در این مثال، بهتر است عامل بر پایه مدل باشد. چرا که در یک محیط پیچیده مانند

شهر با ترافیک شلوغ، داشتن مدل دقیق از قوانین ترافیک و رفتار خودروها و پیاده‌روها به عامل کمک می‌کند تا تصمیمات بهتری درباره حرکت و رانندگی اتخاذ کند. همچنین، با استفاده از مدل، عامل می‌تواند تجربیات مجازی را جمع‌آوری کرده و راهبردهای جدید را برای حل مسائل ترافیکی در شهر آزمایش کند.

۵-

.V(A): 0 + 0.5 * - 1 + 0 = -0.5
.V(B): 0 + 0.5 * - 1 + 0 = -0.5
.V(C): 0 + 0.5 * 32 + 0 = 16

.V(A): 0.5 * - 0.5 + 0.5 * - 1 - 0.5 = -1
.V(B): 0.5 * - 0.5 + 0.5 * - 99 + 0 = -49.75

.Q(A, East): 0 + 0.5 * - 1 + 0 = -0.5
.Q(B, East): 0 + 0.5 * - 1 + 0 = -0.5
.Q(C, East): 0 + 0.5 * 32 + 0 = 16

.Q(A, East): 0.5 * - 0.5 + 0.5 * - 1 + 0 = -0.75
.Q(B, East): 0.5 * - 0.5 + 0.5 * - 99 + 16 = -41.75

.Q(A, South) = 0
.Q(A, East) = -0.75
.Q(B, East) = -41.75

۶-
الف) مقدار بهینه حالت (3، 2) به صورت زیر تعیین می شود: حداکثر مقدار را بین 0 به اضافه 0.5 برابر مقدار حالت (2، 2)، 80 به علاوه 0.5 برابر مقدار حالت (3، 1) و 100 بگیرید. به اضافه 0.5 برابر مقدار حالت (3، 3). حداکثر مقدار 100 است.

مقدار بهینه حالت (2، 2) به صورت زیر تعیین می شود: حداکثر مقدار را بین 0 به اضافه 0.5 برابر مقدار حالت (1، 2)، -100 به اضافه 0.5 برابر مقدار حالت (2، 1) بگیرید، - 80 به علاوه 0.5 برابر مقدار حالت (2، 3)، و 0 به علاوه 0.5 برابر مقدار حالت (3، 2). حداکثر مقدار 50 است.

مقدار بهینه حالت (1، 2) به صورت زیر تعیین می شود: حداکثر مقدار را بین 25 به اضافه 0.5 برابر مقدار حالت (1، 1)، 0 به اضافه 0.5 برابر مقدار حالت (1، 3) و 0 بگیرید. به اضافه 0.5 برابر مقدار حالت (2، 2). حداکثر مقدار 25 است.

ب)
در قسمت اول:
Q-value عمل S را در حالت (1، 3) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [0 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده 0 است.
Q-value عمل E را در حالت (1، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [0 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده 0 است.
Q-value عمل S را در حالت (2، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [-100 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده -50 است.

در قسمت دوم:
Q-value عمل S را در حالت (1، 3) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [0 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده 0 است.
Q-value عمل E را در حالت (1، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [0 به علاوه 0.5 برابر حداکثر Q-value فعلی، که 50- است] اضافه کنید. مقدار به روز شده -12.5 است.
Q-value عمل E را در حالت (2، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [0 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده 0 است.
Q-value عمل N را در حالت (3، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [+100 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده +50 است.

در قسمت سوم:
Q-value عمل S را در حالت (1، 3) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [0 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده 0 است.

Q-value عمل E را در حالت (1، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [-12.5 به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده -6.25 است.
Q-value عمل E را در حالت (2، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [0 به علاوه 0.5 برابر حداکثر Q-value فعلی، که 50+ است] اضافه کنید. مقدار به روز شده +12.5 است.
Q-value عمل S را در حالت (3، 2) به صورت زیر به روز کنید: 0.5 برابر مقدار قبلی را بگیرید و 0.5 برابر [80+ به اضافه 0.5 برابر حداکثر Q-value فعلی، که 0 است] اضافه کنید. مقدار به روز شده +40 است.
بنابراین، مقادیر Q به شرح زیر است:
Q((3، 2)، N) = 50+، Q((1، 2)، S) = 0، Q((2، 2)، E) = 12.5+ ، Q((3، 2)، N) = 50+.

بیایید وزن ها را به صورت w1، w2 و w3 نشان دهیم. می توانیم تابع Q را به صورت زیر بیان کنیم:

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

در مرحله اول، به روز رسانی وزن به صورت زیر محاسبه می شود:

w1 ← 0 + 0.5 * ((-100 + 0) - 0) * f1((2، 2)، S) = -50 * 2 = -100
w2 ← 0 + 0.5 * ((-100 + 0) - 0) * f2((2، 2)، S) = -50 * 2 = -100
w3 ← 0 + 0.5 * ((-100 + 0) - 0) * f3((2، 2)، S) = -50 * 2 = -100

در مرحله دوم وزن های (w1, w2, w3) = (-1, 1, 2) را داریم. برای یافتن عمل انتخاب شده توسط تابع Q در حالت (1، 2)، Qf(s، a) را برای هر عمل a ارزیابی می کنیم:

Qf((1، 2)، N) = -1 * f1((1، 2)، N) + 1 * f2((1، 2)، N) + 2 * f3((1، 2)، N) = -1 + 2 + 1 = 2
Qf((1، 2)، S) = -1 * f1((1، 2)، S) + 1 * f2((1، 2)، S) + 2 * f3((1، 2)، S) = -1 + 2 + 2 = 3
Qf((1، 2)، E) = -1 * f1((1، 2)، E) + 1 * f2((1، 2)، E) + 2 * f3((1، 2)، E) = -1 + 2 + 3 = 4
Qf((1، 2)، W) = -1 * f1((1، 2)، W) + 1 * f2((1، 2)، W) + 2 * f3((1، 2)، W) = -1 + 2 + 4 = 5

بنابراین، عمل انتخاب شده، اقدامی است که مقدار Q را به حداکثر می رساند که در این مورد W است.