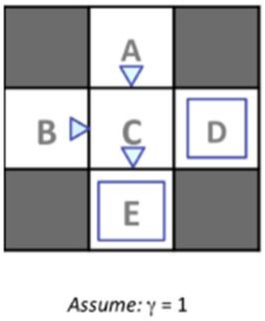
.

CEFET/RJ
Inteligência Artificial (2018.1)
Prof. Eduardo Bezerra (ebezerra@cefet-rj.br)
Lista de exercícios 04

Créditos: essa lista de exercícios contém a tradução dos exercícios disponibilizados na disciplina CS188 - Artificial Intelligence, da Universidade de Berkeley. Os exercícios originais podem ser encontrados em `http://ai.berkeley.edu/section_handouts.html`. São também usados exercícios retirados do livro texto da disciplina (AIMA).

1. **(Aprendizado por Reforço Baseado em Modelo)** Considere o mundo grade e os episódios apresentados abaixo.




Episode 1
A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2
B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3
B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4
A, south, C, -1
C, south, E, -1
E, exit, x, +10

Que modelos seriam aprendidos (i.e., quais as probabilidades das transições) a partir dos episódios observados acima?

(a) $T(A, south, C)$
(b) $T(B, east, C)$
(c) $T(C, south, E)$
(d) $T(C, south, D)$

2. **(Aprendizado por Reforço Baseado em Modelo)** Considere um PDM com 3 estados, A, B e C; e 2 ações Clockwise e CounterClockwise. Não sabemos a função de transição nem a função de recompensa para o PDM. Em vez disso, nos são fornecidas amostras do que um agente experimenta quando ele interage com o meio ambiente (embora nós saibamos que o agente não permanece no mesmo estado depois de tomar uma ação). Neste problema, vamos primeiro estimar o modelo (i.e., a função de transição e a função de recompensa), e então usar o modelo estimado para encontrar as ações ideais.

   Para encontrar as ações ótimas, RL computa as funções de valor ótimas V ou de Q com relação às funções estimadas previamente, T e R. Isto poderia ser feito com qualquer um

dos métodos a seguir: iteração de valor, iteração de política, ou iteração Q-valor. Você já resolveu alguns exercícios que envolveram iteração de valor e iteração de política. Sendo assim, vamos usar o Q-learning neste exercício.

Considere as seguintes amostras que o agente encontrou:

| s | a | s' | r |
|---|---|---|---|
| A | Clockwise | B | 0.0 |
| A | Clockwise | C | -7.0 |
| A | Clockwise | B | 0.0 |
| A | Clockwise | C | -7.0 |
| A | Clockwise | C | -7.0 |
| A | Counterclockwise | C | -6.0 |
| A | Counterclockwise | C | -6.0 |
| A | Counterclockwise | C | -6.0 |
| A | Counterclockwise | C | -6.0 |
| A | Counterclockwise | C | -6.0 |

| s | a | s' | r |
|---|---|---|---|
| B | Clockwise | C | 0.0 |
| B | Clockwise | C | 0.0 |
| B | Clockwise | C | 0.0 |
| B | Clockwise | C | 0.0 |
| B | Clockwise | C | 0.0 |
| B | Counterclockwise | A | 10.0 |
| B | Counterclockwise | A | 10.0 |
| B | Counterclockwise | A | 10.0 |
| B | Counterclockwise | C | 0.0 |
| B | Counterclockwise | A | 10.0 |

| s | a | s' | r |
|---|---|---|---|
| C | Clockwise | A | 0.0 |
| C | Clockwise | B | 1.0 |
| C | Clockwise | B | 1.0 |
| C | Clockwise | B | 1.0 |
| C | Clockwise | B | 1.0 |
| C | Counterclockwise | B | 2.0 |
| C | Counterclockwise | B | 2.0 |
| C | Counterclockwise | B | 2.0 |
| C | Counterclockwise | B | 2.0 |
| C | Counterclockwise | B | 2.0 |

(a) Vamos começar por obter estimativas para as função de transição, $T(s,a,s')$ e para a função de recompensa $R(s,a,s')$ para esse PDM. Preencha os valores faltantes na tabela abaixo para $T(s,a,s')$ e $R(s,a,s')$.

Discount Factor, $\gamma = 0.5$

| s | a | s' | T(s,a,s') | R(s,a,s') |
|---|---|---|---|---|
| A | Clockwise | B | M | N |
| A | Clockwise | C | O | P |
| A | Counterclockwise | C | 1.000 | -6.000 |
| B | Clockwise | C | 1.000 | 0.000 |
| B | Counterclockwise | A | 0.800 | 10.000 |
| B | Counterclockwise | C | 0.200 | 0.000 |
| C | Clockwise | A | 0.200 | 0.000 |
| C | Clockwise | B | 0.800 | 1.000 |
| C | Counterclockwise | B | 1.000 | 2.000 |

| | |
|---|---|
| M | ? |
| N | ? |
| O | ? |
| P | ? |

(b) Agora execute o algoritmo Q-learning usando as funções estimadas T e R. Os valores de $Q_k(s,a)$ são dados da Tabela 1.

(c) Suponha que o Q-learning convirja para a função $Q^*(s,a)$ apresentada na Tabela 2. Qual é a ação ótima (Clockwise ou Counterclockwise), para cada um dos estados (A, B e C)?

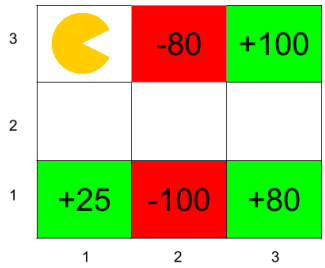
Tabela 1: Valores $Q_k(s,a)$

| | A | B | C |
|---|---|---|---|
| Clockwise | -2,0 | 1,0 | 3,58 |
| Counterclockwise | -5,0 | 6,52 | 6,0 |

Tabela 2: Valores após a convergência do Q-learning.

| | A | B | C |
|---|---|---|---|
| Clockwise | -0,679 | 3,086 | 4,07 |
| Counterclockwise | -2,914 | 8,346 | 6,173 |

3. **(Q-learning)** No mundo grade apresentado abaixo, o Pacman está tentando aprender a política ótima. Se uma ação resulta em um dos estados sombreados a recompensa correspondente é concedida durante essa transição. Todos os estados sombreados são estados terminais, ou seja, o PDM termina uma vez que o agente chega em um estado sombreado. Os outros estados têm as ações disponíveis ***N**orth, **E**ast, **S**outh, **W**est*, que deterministicamente movem o Pacman para o estado vizinho correspondente (ou fazem com que o Pacman permaneça no lugar se a ação tentar sair da grade). Considere que o fator de desconto é $\gamma = 0{,}5$ e que a taxa de aprendizagem do Q-learning é $\alpha = 0{,}5$ para todos os cálculos. O Pacman começa no estado (1, 3).




   (a) Qual o valor ótimo para a função $V^*$ nos seguintes estados?
   $V^*(3,2) =$
   $V^*(2,2) =$
   $V^*(1,3) =$

   (b) Considere que o agente inicia no canto superior esquerdo. Considere também que são dados os seguintes episódios de execuções do agente através do mundo grade, conforme a Tabela 3. Cada linha em um episódio é uma tupla contendo $(s, a, s', r)$.

   Com o uso de atualizações Q-Learning, forneça os Q-valores abaixo após os três episódios acima:
   $Q((3,2),\text{N}) =$
   $Q((1,2),\text{S}) =$
   $Q((2,2),E) =$

Tabela 3: Episódios no mund grade.

| Episódio 1 | Episódio 2 | Episódio 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 |

(c) Considere uma representação baseada em características da função Q:

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

$f_1(s)$ : A coordenada x do estado $\qquad$ $f_2(s)$ : A coordenada y do estado

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

(i) Dado que todos os $w_i$ são inicialmente iguais a 0, quais são seus valores após o primeiro episódio?

$w_1 =$
$w_2 =$
$w_3 =$

(ii) Considere que o vetor de pesos $w$ é igual a $(1, 1, 1)$. Qual a ação recomendada pela função Q no estado (2,2) ?