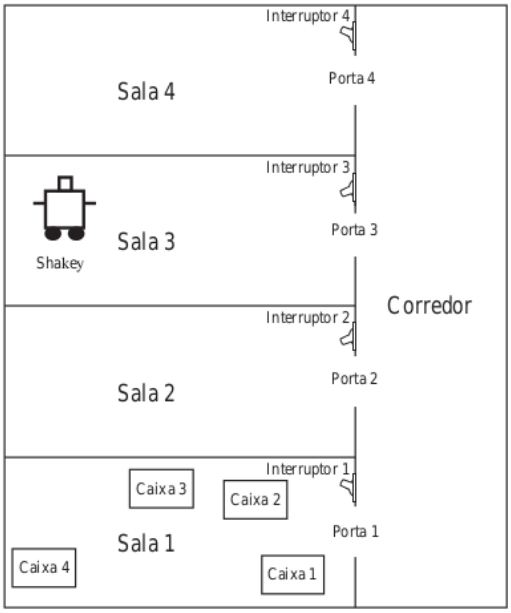
Inteligência Artificial - 2023
Lista 3
Entrega: dia 29 de junho

1. O planejador STRIPS original foi projetado para controlar o robô Shakey. A figura abaixo mostra uma versão do mundo de Shakey que consiste em quatro salas dispostas ao longo de um corredor, onde cada sala tem uma porta e um interruptor de luz. As ações no mundo de Shakey incluem movimentar-se de um lugar para outro, empurrar objetos móveis (como caixas), subir e descer de objetos rígidos (como caixas) e ligar e desligar interruptores. O robô propriamente dito nunca chegou a conseguir subir em uma caixa ou acionar um interruptor, mas o planejador de STRIPS era capaz de descobrir e imprimir planos que estavam além das habilidades do robô.



As seis ações de Shakey são as seguintes:

- `Ir(x, y, r)`, que exige que Shakey esteja em `x` e que `x` e `y` sejam posições na mesma sala `r`. Por convenção, **uma porta entre duas salas está em ambas as salas**.
- Empurrar uma caixa `b` da posição `x` para a posição `y` dentro da mesma sala: `Empurrar(b, x, y, r)`. Precisaremos do predicado `Caixa` e de constantes relativas às caixas.
- Subir em uma caixa de posição `x`: `Subir(x, b)`; descer de uma caixa para a posição `x`: `Descer(b, x)`. Precisaremos do predicado `Sobre` e da constante `Piso`.
- Ligar ou desligar um interruptor: `Ligar(s, b)`; `Desligar(s, b)`. Para ligar ou desligar um interruptor de luz, Shakey tem de estar em cima de uma caixa na posição do interruptor de luz.

Descreva as seis ações de Shakey e o estado inicial da figura. Construa um plano para Shakey colocar $\texttt{Caixa}_2$ em $\texttt{Sala}_2$.

2. Considere um MDP 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 problema. Em vez disso, são fornecidas amostras do que acontece quando o agente interage com o meio ambiente. 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.

   Considere as seguintes amostras que o agente encontrou:

s	a	s'	r
A	Clockwise	B	0.0
A	Clockwise	C	1.0
A	Clockwise	C	1.0
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Counterclockwise	C	9.0
A	Counterclockwise	C	9.0
A	Counterclockwise	B	0.0
A	Counterclockwise	B	0.0
A	Counterclockwise	B	0.0

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	-2.0
B	Counterclockwise	A	-2.0
B	Counterclockwise	A	-2.0
B	Counterclockwise	A	-2.0
B	Counterclockwise	A	-2.0

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

(a). Vamos começar obtendo estimativas para as funções de transição T(s,a,s') e de recompensa R(s,a,s') para esse MDP. Preencha os valores que faltam na tabela abaixo:

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	B	0.600	0.000
A	Counterclockwise	C	0.400	9.000
B	Clockwise	C	1.000	0.000
B	Counterclockwise	A	1.000	-2.000
C	Clockwise	A	0.800	0.000
C	Clockwise	B	0.200	-6.000
C	Counterclockwise	B	1.000	6.000

(b). Agora execute uma iteração do algoritmo Q-learning usando as funções estimadas T e R, calculando os valores de Q_{k+1}(s,a). Os valores de Q_k(s,a) são:

	A	B	C
Clockwise	1.6	3.0	0.24
Counterclockwise	4.8	-0.2	6.0

3. No mundo apresentado na figura, o Pacman está tentando aprender a política ótima. Todos os estados coloridos são estados terminais, ou seja, neles, a única ação possível é Sair, o agente recebe a recompensa correspondente, vai para um estado D fora do tabuleiro e o MDP termina. Os outros estados têm as ações disponíveis North, East, South, West, que deterministicamente movem o Pacman para o estado vizinho correspondente (ou fazem com que o Pacman permaneça no lugar se não houver a posição vizinha). 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). Considere que o agente começa no canto superior esquerdo. Considere também que são dados os seguintes episódios de execuções, em que cada linha em um episódio é uma tupla contendo (s, a, s', r):

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(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	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), E, (3,2), 0	(2,2), S, (2,1), 0	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
(3,2), N, (3,3), 0	(2,1), Exit, D, -100	(3,2), S, (3,1), 0	(3,2), N, (3,3), 0	(3,2), S, (3,1), 0
(3,3), Exit, D, +50		(3,1), Exit, D, +30	(3,3), Exit, D, +50	(3,1), Exit, D, +30

Dê os valores abaixo após os 5 episódios:
Q((3,2), N) =
Q((3,2), S) =
Q((2,2), E) =

(b). Considere uma representação baseada em características:

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

onde
$f_1(s)$: A coordenada x do estado
$f_2(s)$: 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?

(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) ?