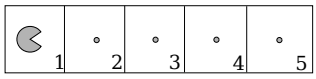
Inteligência Artificial — Lista de exercícios: MDP e Aprendizado por reforço
Prof. Hendrik Macedo

1. Consiere o MDP abaixo onde S = {1, 2, 3, 4, 5, Fim}. Para 1 até 4, as ações (determinísticas) possíveis são *Direita (D)* ou *Encerrar (E),* que encerra o jogo. Na casa 5 só é possível *E*. Comer um ponto dá recompensa de 10. *E* dá recompensa de 20. Considere 3 políticas:
π0(s)=*E*;
π1(s≤3)=*D*, π1(s>3)=*E*;
π2(s≤4)=*D*, π2(s>4)=*E*.



(a) Para γ = 1.0, quais os valores de Vπ0(1), Vπ1(1), Vπ2(1), V*(1), respectivamente?
(b) Existe um valor para γ onde π0 seja estritamente melhor que π2? Marque a alternativa que representaria uma possibilidade.

○ $\gamma = \frac{1}{2}$ ○ $\gamma = \frac{1}{4}$ ○ qualquer $\gamma < 1$ ○ ∄

2. Considere um agente Q-learning com α fixo e fator de desconto γ, que estava no estado 34, realizou ação 7, recebeu recompensa 3 e terminou no estado 65. Que valores foram atualizados? Dê uma expressão para o compto do novo valor (seja tão específico quanto possível).

3. Considere um jogo de blackjack onde você suspeita que as cartas não aparecem com probabilidades iguais (como deveria ser) e decide usar *Q-learning* ao invés de *Value Iteration*. Considere a tabela de Q-values inicial (abaixo) e o novo episódio observado (ao lado).

s	a	$Q(s,a)$
19	hit	−2
19	stay	5
20	hit	−4
20	stay	7
21	hit	−6
21	stay	8
bust	stay	−8

s	a	r	s	a	r	s	a	r
19	hit	0	21	hit	0	bust	stay	−10

Marque a alternativa que corresponde à veracidade da atualização da tabela de Q-values abaixo. Considere $\alpha = 0.5$ e $\gamma = 1.0$.

s	a	$Q(s,a)$
19	hit	*a = 3*
19	stay	
20	hit	
20	stay	
21	hit	*b = -7*
21	stay	
bust	stay	*c = -9*

○ a, b e c estão corretos;
○ a, b e c estão errados;
○ a e b estão corretos e c errado;
○ a e c estão corretos e b errado;
○ b e c estão corretos e a errado;
○ apenas uma letra está correta.

4. Considere um MDP com 6 estados: Bathroom, Kitchen, Bedroom, Dining Room, Under-Attack and Dead. Isso representa um domínio de um rato robótico procurando alimentos em uma casa com quatro cômodos (Bathroom, Kitchen, Bedroom, Dining Room). Nos estados Bathroom, Kitchen, Bedroom, Dining Room, existem três ações disponíveis: stay in place (S), move horizontally (H) e move vertically (V). Enquanto se alimenta, o rato pode ser atacado por um gato robótico que também habita a casa, o que faz com que o rato entre no estado Under-Attack. Do estado Under-Attack, há apenas uma ação, Die. No estado Dead há apenas uma ação, Stay Dead. As recompensas e probabilidades de transição são as seguintes:

T(s,a,s')

	s'					
s, a	Bathroom	Kitchen	Bedroom	Dining Room	Under-Attack	Dead
Bathroom, H	0	0.6	0.4	0	0	0
Bathroom, V	0	0.4	0.6	0	0	0
Bathroom, S	0.75	0	0	0	0.25	0
Kitchen, H	0.6	0	0	0.4	0	0
Kitchen, V	0.4	0	0	0.6	0	0
Kitchen, S	0	0.75	0	0	0.25	0
Bedroom, H	0.4	0	0	0.6	0	0
Bedroom, V	0.6	0	0	0.4	0	0
Bedroom, S	0	0	0.75	0	0.25	0
Dining Room, V	0	0.6	0.4	0	0	0
Dining Room, H	0	0.4	0.6	0	0	0
Dining Room, S	0	0	0	0.75	0.25	0
Under-Attack, Die	0	0	0	0	0	1.0
Dead, Stay Dead	0	0	0	0	0	1.0

s	$R(s)$
Bathroom	+4
Kitchen	+10
Bedroom	0
Dining Room	+2
Under-Attack	-50
Dead	0

(a) Qual o número total de Políticas possíveis?
(b) Desenvolva a iteração de valor manualmente neste problema. Inicialize o valor de cada estado com 0 (zero). Use γ = 0.5. Dê os valores para todos os estados após cada iteração. Você pode parar após 6 iterações.
(c) Dados os valores calculados, qual a Política ótima para cada estado?