Lista de Exercícios de Planejamento Probabilístico

Nome: _____

Questões extraídas do livro "Artificial Intelligence: A Modern Approach" de Stuart Russell e Peter Norvig.

1. Considere o mundo 4×3 da Figura 1 (visto em aula) e compute quais quadrantes podem ser atingidos desde $(1, 1)$ seguindo a seqüência de ações $[Up, Up, Right, Right, Right]$, e com quais probabilidades.

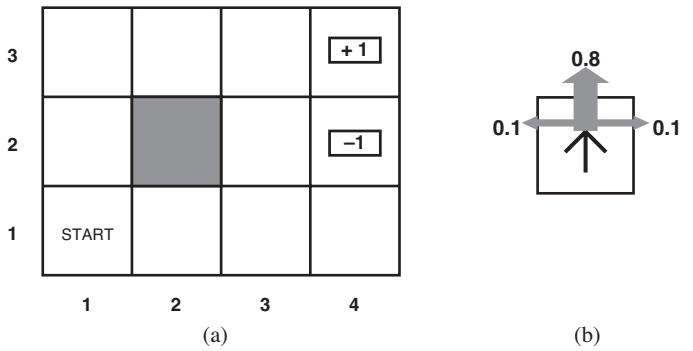

Figura 1: Sequential decision-making problem

Solution:
Para resolver esta questão, é mais eficiente criar um "mapa de ocupação", indicando, em cada ponto do tempo, qual a probabilidade de o agente estar em cada quadrante (você saberá que está certo se as linhas somarem um total de 1), por exemplo, depois de executar o primeiro Up, teremos as probabilidades abaixo (cmo boa parte da tabela a completar).

		Up	Up	Right	Right	Right
(1,1)	1	0.1				
(1,2)		0.8				
(1,3)						
(2,1)		0.1				
(2,3)						
(3,1)						
(3,2)						
(3,3)						
(4,1)						
(4,2)						
(4,3)						

2. Em alguns casos, MDPs são formuladas com uma função de recompensa $R(s, a)$ que depende na ação tomada, ou com uma função de recompensa $R(s, a, s')$ que também depende no estado resultante.

   (a) Escreva as equações de Bellman para estas formulações.

   Solution: Boa sorte.

(b) Mostre como gerar funções de recompensa $R(s,a)$ e $R(s,a,s')$ a partir de uma função de recompensa $R(s)$ e de u,a função de transição $P(s' \mid a,s)$

Solution:

$$R(s,a) = R(s) + \gamma \sum_{s'} P(s' \mid a,s) R(s')$$

$$R(s,a,s') = R(s) + \gamma P(s' \mid a,s) R(s')$$

(c) (Difícil) Mostre como um MDP com função de recompensa $R(s,a,s')$ pode ser transformada em um MDP diferente com função de recompensa $R(s,a)$, tal que as políticas ótimas do novo MDP correspondam exatamente às políticas ótimas do MDP original.

Solution: Utilize as definições formais de MDP para gerar estas equações.

(d) (Difícil) Faça o mesmo para converter MDPs com $R(s,a)$ em MDPs com $R(s)$.

Solution: Boa sorte.

3. O que pode acontecer com uma política gerada para uma MDP com horizonte finito?

Solution: Sendo uma MDP com horizonte finito existe algum fator limitante (tempo, número de passos), logo é comum a política mudar as ações para cada estado de acordo com o fator limitante, *i.e.* uma política *nonstationary*.

4. Como o algoritmo de avaliação de políticas $U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t)\right]$ pode ser utilizado para calcular a perda esperada de utilidade por parte de um agente utilizando um conjunto de estimativas de utilidade U e um modelo estimado P, quando comparado com um agente utilizando os valores corretos?

Solution: Boa sorte.