CMPUT 366/609 Assignment 2: Markov Decision Process 1

Ricardo Holguin Esquer
Tpicos avanzados en I.A.

October 15, 2018

Ejercicio 1. Trajectories, returns, and values.

a) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_1:

s	a	s'	$p(s'\|s,a)$	$r(s,a,s')$
X	left	X	1	0
Y	right	goal	1	4

$$(X, left, 0), (X, left, 0), \ldots$$

b) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_2:

s	a	s'	$p(s'\|s,a)$	$r(s,a,s')$
X	right	X	3/4	1
X	right	Y	1/4	-1
Y	right	goal	1	4

$$(X, right, 1), (X, right, 1), (X, right, 1), (X, right, -1), (Y, right, 4), (Goal, nothing, nothing)$$

c) Assuming the discount-rate parameter is $\gamma = 0.5$, what is the return from the initial state for the second trajectory?

$$G_0 = 1 + \gamma(1 + \gamma(1 + \gamma(-1 + \gamma(4))))$$
$$G_0 = 1 + \gamma(1) + \gamma^2(1) + \gamma^3(-1) + \gamma^4(4)$$
$$G_0 = 1.875$$

d) Assuming $\gamma = 0.5$, what is the value of state Y under policy π_1 ?

$v_{\pi_1}(\mathrm{Y}) = 1[0.5(4)] = 2$

e) Assuming $\gamma = 0.5$, what is the action-value of X,left under policy π_1 ?

$q_{\pi_1}(\text{X}, left) = 0$

f) Assuming $\gamma = 0.5$, what is the value of state X under policy π_2?

$v_{\pi_2}(\text{X}) = \frac{1}{4}[-1 + 0.5(1(4))] = 0.25$

Ejercicio 2. Questions from the book and others that are not.

a) Exercise 3.1 - 3 ejemplos de aplicaciones de un MDP.

1) Blackjack: Los estados pueden ser las cartas en la mano del jugador, la suma de estos, la mano del delaer y la suma de este. La recompensa puede ser simple, donde 1 sea si se gana un juego, 0 si no, y -1 si se pierde. Las acciones pueden ser Tomar una carta o no tomar una carta.
2) Uno: Los estados pueden ser la mano del jugador y la ultima carta que se encuentra en el montn de cartas ya jugadas. Las acciones puede ser jugar una carta, o jalar n cartas (hasta que se tenga una carta que jugar). La recompensa puede ser 1 si se puede jugar una carta y -1 si se jala una carta
3) Ajedrez: Los estados son las posiciones de las piezas en el tablero. Las acciones son los movimientos de las piezas, resultado 1 si se gana o -1 si se pierde.

b) Exercise 3.6 - El robot que no aprende a salir del laberinto.
Tal vez la recompensa no esta bien definida, donde no se castiga los caminos sin salida o no se recompensa las salidas del laberinto. Otro problema puede ser que en los episodios que se dejo aprendiendo nunca llego a una salida por lo que la poltica no reconoce un camino correcto para la salida. Otra solucin puede ser aumentar el nmero de episodios para que as los estados donde ya haya visitado valgan menos y menos hasta que eventualmente el robot quiera moverse a otro lado.

c) Exercise 3.8 - Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = 1, R_2 = 2, R_3 = 6, R_4 = 3, and R_5 = 2$, with $T = 5$. What are $G_0, G_1, ..., G_5$? Hint: Work backwards.

Usando $G_t = R_{t+1} + \gamma G_{t+1}$

$$G_5 = 0$$

$$G_4 = R_5 + \gamma G_5 = 2 + (0.5)(0) = https://www.overleaf.com/project/5bc3ee90b43798777eecc2e22$$

$$G_3 = R_4 + \gamma G_4 = 3 + (0.5)(2) = 4$$

$$G_2 = R_3 + \gamma G_3 = 6 + (0.5)(4) = 8$$

$$G_1 = R_2 + \gamma G_2 = 2 + (0.5)(8) = 6$$

$$G_0 = R_1 + \gamma G_1 = -1 + (0.5)(6) = 2$$

https://www.overleaf.com/project/5bc3ee90b43798777eecc2e2

d) Exercise 3.9 - Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0 ?

$$G_1 = 2 + \sum_{i=1}^{\infty} 0.9^i \times 7 = 2 + 63 = 65$$

$$G_0 = 0 + 9 \times 2 + \sum_{i=2}^{\infty} 0.9^i \times 7 = 0 + 1.8 + 63 = 64.8$$

e) Exercise 3.11 - Bellman equation.

f) Exercise 3.12 - Give an equation for v in terms of v_π and π