CMPUT 366/609 Assignment 2: Markov Decision Processes 1

Adrian Emilio Vazquez Icedo

14/10/2018

1 Question 1

Consider the MDP above, in which there are two states, X and Y, two actions, right and left, and the deterministic rewards on each transition are as indicated by the numbers. Note that if action right is taken in state X, then the transition may be either to X with a reward of +1 or to Y with a reward of -1. These two possibilities occur with probabilities 3/4 (for the transition to X) and 1/4 (for the transition to state Y). Consider two deterministic policies, π_1 and π_2:

$$\pi_1(\text{X}) = \text{left } \pi_2(\text{X}) = \text{right}$$
$$\pi_1(\text{Y}) = \text{right } \pi_2(\text{Y}) = \text{right}$$

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

$$(X, left, 0), (X, left, 0), (X, left, 0), ...$$

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

$$(X, right, 1), (X, right, 1), (X, right, 1), (X, right, -1), (Y, right, 4)$$

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

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

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

$$\upsilon_{\pi_1}(Y) = 4$$

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

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

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

$$\upsilon_{\pi_2}(X) = 1 + \gamma * 1 + \gamma^2 * 1 + \gamma^3 * (-1) + \gamma^4 * 4 = 1.875$$

2 Question 2

(a) Exercise 3.1

Domino: Los estados son las fichas que cada jugador tiene, las acciones serian colocar ficha o pasar y la recompenza se puede calcular tomando en cuenta las veces que se coloca una ficha y cuando se pasa sin colocar. 21: Los estados son la suma de las cartas que tiene el jugador y el dealer y las acciones son pedir una carta o quedarse. Las recompensa seria si gana o no gana el juego.

(b) Exercise 3.7

Al no tener un objetivo para el agente se puede considerar que cualquir recompenza es 0, por lo que al realizar cualquier accion no se puede considerar mejor o peor.

(c) Exercise 3.8

$$\gamma = .5$$
$$G_5 = 2, G_0 = 1$$
$$G_4 = 3 + r * 2 = 4$$
$$G_3 = 6 + r * 3 + r^2 * 2 = 8$$
$$G_2 = 2 + r * 6 + r^2 * 3 + r^3 * 2 = 6$$
$$G_2 = -1 + r * 2 + r^2 * 6 + r^3 * 3 + r^4 * 2 = 6$$

(d) Exercise 3.9

(e) Exercise 3.11

(f) Exercise 3.12

(g) Exercise 3.13

(h) Exercise 3.14

(i) Exercise 3.15

(j) Figure 3.6 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.7) to express this value symbolically, and then to compute it to three decimal places. Hint: Equation (3.9) is also relevant.

3 Question 3

4 Question 4