Ejercicios de programación dinámica para Tópicos Avanzados de Inteligencia Artificial

Ivan Alejandro Moreno Soto

14 de octubre de 2018

1. Ejercicios de CMPUT 366/609 Assignment 2: Markov Decision Processes 1

1.1. Pregunta 1

(a) *Muestre una trayectoria típica de X para la política π_1.*

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

(b) *Muestre una trayectoria típica de X para la política π_2.*

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

(c) *Asumiendo que el factor de descuento es $\gamma = 0{,}5$, ¿Cuál es la recompensa del estado inicial para la segunda trayectoria?*

$$r_0 = 1 + \gamma(1) + \gamma^2(1) - \gamma^3(-1) + \gamma^4(4) = 1{,}875$$

(d) *Asumiendo $\gamma = 0{,}5$, ¿Cuál es el valor del estado Y con la política π_1?*

$$V^{\pi_1}(Y) = 4$$

(e) *Asumiendo $\gamma = 0{,}5$, ¿Cuál es el valor de acción de $X, left$ con la política π_1?*

$$q(X, left) = 0$$

(f) *Asumiendo $\gamma = 0{,}5$, ¿Cuál es el valor del estado X con la política π_2?*

$$V^{\pi_2}(X) = 1 + \gamma(1) + \gamma^2(1) + \gamma^3(-1) + \gamma^4(4) = 1{,}875$$

1.2. Pregunta 2

(a) *Ejercicio 3.1 de Sutton-Barto*

- Jugar blackjack. Cada posible valor de cartas son los estados, mientras que las acciones son pedir cartas o quedarse. Las recompensas se calculan únicamente cuando el jugador gane o pierda.

- Jugar un videojuego de combate por turnos. Los puntos de vida de cada personaje y de cada enemigo, junto con los objetos disponibles son los estados. Las acciones son cada ataque o movimiento que los personajes puedan realizar en el turno actual. Las recompensas pueden ser calculadas tomando en cuenta el daño hecho a los enemigos, el recibido, y si estos fueron derrotados.

- Navegar un bosque. Cada estado es la posición actual en el bosque. Las acciones pueden ser caminar en 4 direcciones, o incluso en 8. Las recompensas pueden ser calculadas respecto a la distancia que existe de la salida.

(b) *Ejercicio 3.7 de Sutton-Barto* No le fue comunicado bien el objetivo. Como no obtiene mejores recompensas por ninguna ruta, no distingue entre las potencialmente mejores.

(c) *Ejercicio 3.8 de Sutton-Barto*

$$G_5 = 2, G_0 = 1$$

$$G_4 = 3 + (0{,}5)(2) = 4$$

$$G_3 = 6 + (0{,}5)(3) + (0{,}5)^2(2) = 8$$

$$G_2 = 2 + (0{,}5)(6) + (0{,}5)^2(3) + (0{,}5)^3(2) = 6$$

$$G_1 = -1 + (0{,}5)(2) + (0{,}5)^2(6) + (0{,}5)^3(3) + (0{,}5)^4(2) = 2$$

(d) *Ejercicio 3.9 de Sutton-Barto*

Para G_1 tenemos

$$G_1 = 2 + \sum_{i=1}^{\infty} 0{,}9^i \times 7 = 2 + 70 = 72$$

Para G_0 tenemos

$$G_0 = 0 + 1{,}8 + \sum_{i=2}^{\infty} 0{,}9^i \times 7 = 1{,}8 + 70 = 71{,}8$$

(e) *Ejercicio 3.11 de Sutton-Barto*

$$r(s, a, s') = \sum_{a \in A} \pi(a|s) \sum_{r \in R} r \frac{p(s', r|s, a)}{p(s'|s, a)}$$

(f) *Ejercicio 3.12 de Sutton-Barto*

$$v_\pi(s) = q_\pi(s, a) + v_\pi(s')$$

(g) *Ejercicio 3.13 de Sutton-Barto*

$$q_\pi(s, a) = p(s', r|s, a) v_\pi(s)$$

(h) *Ejercicio 3.14 de Sutton-Barto* Si calculamos la ecuación de Bellman recordando que solo tenemos una cifra de precisión y que las transiciones son deterministas, tenemos

$$\begin{aligned} v_\pi(centro) =& (0{,}25 \times 0{,}9 \times 2{,}3) + (0{,}25 \times 0{,}9 \times 0{,}4) \\ & (0{,}25 \times 0{,}9 \times 0{,}7) + (0{,}25 \times 0{,}9 \times -0{,}4) \\ =& (0{,}25 \times 0{,}9)(2{,}3 + 0{,}4 + 0{,}7 - 0{,}4) \\ =& 0{,}675 = 0{,}7 \end{aligned}$$

(i) *Ejercicio 3.15 de Sutton-Barto*

Solo importan los intervalos entre las recompensas, porque estas distancias son las que determinan las acciones que son mejores.

Demostración. Si agregamos una constante c a las recompensas de cada estado tenemos por la ecuación 3.8

$$\begin{aligned}G_t &= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c\end{aligned}$$

Sustituyendo en la ecuación de Bellman tenemos

$$\begin{aligned}v_\pi(s) &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma(E[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}] + E[\sum_{k=0}^{\infty} \gamma^k c])] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma(v_\pi(s) + v_c)] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s)] + \gamma v_c \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s)] + v_c\end{aligned}$$

□

1.3. Pregunta 3

Tenemos que $\gamma = 0{,}8$. Si calculamos el valor v_π para el nodo correspondiente t tenemos

$$\begin{aligned}v_\pi(t) =& \sum_a \pi(a|t) \sum_{s',r} \rho(s', r|t, a)[r + \gamma v_\pi(s')] \\ =&(0{,}5)[0{,}8(3 + 0{,}8 \times 2) + 0{,}2(-6 + 0{,}8 \times 7)]+ \\ &(0{,}5)[0{,}25(-3 + 0{,}8 \times -1) + 0{,}75(4 + 0{,}8 \times 0)] \\ =&2{,}825\end{aligned}$$

Es fácil ver que la política óptima siempre escoge irse por la rama izquierda, además de que la política óptima siempre es determinista. Así, v_* se calcula como sigue

$$
\begin{aligned}
v_*(t) =& \sum_{s',r} \rho(s', r|t, a)[r + \gamma v_*(s')] \\
=& 0{,}8(3 + 0{,}8 \times 2) + 0{,}2(-6 + 0{,}8 \times 7) \\
=& 3{,}6
\end{aligned}
$$