Markov Decision Process Exercises

Belen Isabel Chavarría Silvestre

Octubre 2018

Ejercicio 3.1 Cree tres tareas de ejemplo propias que se ajusten al marco de MDP, identificando para cada uno sus estados, acciones y recompensas. Haz los tres ejemplos tan diferentes entre sí como sea posible. El marco es abstracto y flexible y se puede aplicar de muchas maneras diferentes. Estire sus límites de alguna manera en al menos uno de sus ejemplos.

Ejercicio 3.7Imagina que estás diseñando un robot para ejecutar un laberinto. Decides darle una recompensa de +1 por escapar del laberinto y una recompensa de cero en cualquier otro momento. La tarea parece dividirse naturalmente en episodios, los sucesivos recorren el laberinto, por lo que decide tratarla como una tarea episódica, donde el objetivo es maximizar la recompensa total esperada (3.7). Después de ejecutar el agente de aprendizaje durante un tiempo, encontrará que no muestra ninguna mejora en escapar del laberinto. ¿Qué va mal? ¿Ha comunicado efectivamente al agente lo que desea que logre?

Respuesta Para el robot le es lo mismo equivocarse 10 veces que una. Puede estar lejos de la meta y recorrer el mismo camino 20 veces hasta encontrar la ruta correcta que encontrarla desde un inicio. Las recompensas están dadas para que salga en algún momento no para que lo haga lo más pronto posible.

Ejercicio 3.8 Supongamos $\gamma = 0.5$ y se recibe la siguiente secuencia de recompensas $R_1 = 1, R_2 = 2, R_3 = 6, R_4 = 3 y R_5 = 2$, con $T = 5$. ¿Cuáles son $G_0, G_1, ..., G_5$?
Respuesta

$$\begin{aligned} G_0 &= 1 + 0.5 * 2 + 0.5^2 * 6 + 0.5^3 * 3 + 0.5^4 * 2 = 4 \\ G_1 &= 2 + 0.5 * 6 + 0.5^2 * 3 + 0.5^3 * 2 = 6 \\ G_2 &= 6 + 0.5 * 3 + 0.5^2 * 2 = 8 \\ G_3 &= 3 + 0.5 * 2 = 4 \\ G_4 &= 2 \\ G_5 &= 0 \end{aligned} \tag{1}$$

Ejercicio 3.9 Supongamos $\gamma = 0.9$ y la secuencia de recompensa es $R_1 = 2$ seguida de una secuencia infinita de 7s. ¿Qué son G_1 y G_0?

Respuesta

$$\begin{aligned} G_0 &= 2 + 0.9 * 2 + 0.9^2 * 2 + \dots \\ &= 2 * \sum_{i=0}^{\infty} (0.9)^i \\ &= 2 * \frac{1}{1 - 0.9} \\ &= 20 \end{aligned}$$

Ejercicio 3.11 Si el estado actual es S_t, y las acciones se seleccionan de acuerdo con la política estocástica π, ¿cuál es la esperanza de R_{t+1} en términos de π y la función de cuatro argumentos p (3.2)?
Respuesta

$$\sum_{r \in R} \sum_{s' \in S} p(s', r | s_t, \pi(s_t))$$

Ejercicio 3.12 Da una ecuación para v_π en terminos de q_π y π.
Respuesta

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a|s) * q_\pi(s, a)$$

Ejercicio 3.13 Da una ecuación para q_π en terminos de v_π y los cuatro argumentos p.
Respuesta

$$q_\pi(s, a) = \sum_{s' \in S} \sum_{r \in R} p(s', r | s, a)[r + \gamma E_\pi[G_{t+1} | S_{t+1} = s', A_t = a]]$$

Ejercicio 3.14 La ecuación de Bellman (3.14) debe mantenerse para cada estado para la función de valor v_π que se muestra en la Figura 3.2 (derecha) del Ejemplo 3.5. Muestre numéricamente que esta ecuación es válida para el estado central, valorada en +0.7, con respecto a sus cuatro estados vecinos, valorada en +2.3, +0.4, 0.4 y +0.7. (Estos números son precisos sólo para un lugar decimal.)

Respuesta

$$
\begin{aligned}
v_\pi(s_c) &= \frac{1}{4}\sum_{s'\in S} p(s, right, s')[r + \gamma v_\pi(s')] \\
&\quad + \frac{1}{4}\sum_{s'\in S} p(s_c, left, s')[r + \gamma v_\pi(s')] \\
&\quad + \frac{1}{4}\sum_{s'\in S} p(s_c, north, s')[r + \gamma v_\pi(s')] \\
&\quad + \frac{1}{4}\sum_{s'\in S} p(s_c, south, s')[r + \gamma v_\pi(s')] \\
v_\pi(s_c) &= \frac{1}{4}p(s_c, right, s_r)[0 + 0.9 * v_\pi(s_r)] \\
&\quad + \frac{1}{4}p(s_c, left, s_l)[0 + 0.9 * v_\pi(s_l)] \\
&\quad + \frac{1}{4}p(s_c, north, s_n)[0 + 0.9 * v_\pi(s_n)] \\
&\quad + \frac{1}{4}p(s_c, south, s_s)[0 + 0.9 * v_\pi(s_s)] \\
v_\pi(s_c) &= \frac{1}{4}(0.9)(0.4 + 0.7 + 2.3 - 0.4) \\
&= 0.675 \\
&\approx 0.7
\end{aligned}
$$

Ejercicio 3.15 En el ejemplo de gridworld, las recompensas son positivas para los objetivos, negativas para correr hacia el borde del mundo y cero en el resto del tiempo. ¿Son importantes los signos de estas recompensas, o solo los intervalos entre ellas? Demuestre, usando (3.8), que agregar una constante c a todas las recompensas agrega una constante, v_c, a los valores de todos los estados, y por lo tanto no afecta los valores relativos de ningún estado bajo ninguna política. ¿Qué es v_c en términos de c y γ?
Respuesta

$$
\begin{aligned}
G_t &= R_{t+1} + c + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \dots \\
&= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \\
&= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c \\
&= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \frac{c}{1-\gamma}
\end{aligned}
$$

$$
\begin{aligned}
v_\pi^c(s) &= E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \frac{c}{1-\gamma} | s_t = s] \\
&= E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s] + E_\pi[\frac{c}{1-\gamma} | s_t = s] \\
&= E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s] + \frac{c}{1-\gamma} \\
&= v_\pi(s) + v_c
\end{aligned}
$$

Pregunta 3
Respuesta

$$
\begin{aligned}
v_\pi(s_1) &= \sum_{a \in A(s_1)} \pi(a|s) \sum_{s' \in S} p(s, a, s')[r + \gamma v_\pi(s')] \\
&= 0.5(0.8(3 + 0.8 * 2) + 0.2(-6 + 0.8 * -1)) + 0.5(0.25(-3 + 0.8 * -1) + 0.75(4 + 0.8 * v(s_1)) \\
\frac{7}{10} v_\pi(s_1) &= \frac{113}{40} \\
v_\pi(s_1) &\approx 4.035
\end{aligned}
$$

De acuerdo a los algoritmos de programación dinámica $v_*(s_1) = a_2$ donde a_2 es la acción que puede llevarte de nuevo a s_1