Ejercicio 1

Considere un proceso de decisión de Markov (MDP) con dos estados: uno terminal y otro no-terminal. Existe sólo una acción que lleva del estado no-terminal al estado terminal con probabilidad $1-\rho$ y del estado no-terminal a si mismo con probabilidad ρ. La recompensa es $+1$ en todas las transiciones y el factor de descuento es $\gamma = 1$. Suponga que observa un episodio con 10 iteraciones y un retorno de 10.

- ¿Cuáles son las estimaciones Monte Carlo de primer-visita y de cada-visita del valor del estado no-terminal basadas en ese episodio?
- Compare los valores de estado obtenidos con el teórico si $\rho = 0.9$.

Saque conclusiones.

Solución

Primer-Visita Para la estimación Monte Carlo de primer-visita, es el retorno recolectado al final del episodio después de haber visitado el primer paso.
Suponiendo una inicialización de $G = 0 \;:\; G = 10$

Cada-Visita El estimador Monte Carlo de cada-visita es el promedio de los retornos recibidos en cada estado.
Suponiendo una inicialización de $G = 0 \;:\; G = 10$

$$\begin{aligned} G &= \frac{1+2+3+4+5+6+7+8+9+10}{10} \\ &= \frac{55}{10} \\ &= 5.5 \end{aligned}$$

Por otro lado, el valor teórico del estado no-terminal es $V(S_1) = \sum_{s'} P(s'|S_1) \cdot [R(S_1, s') + \gamma V(s')]$.

$$\begin{aligned} V(S_1) &= \rho \cdot [1 + V(S_1)] + (1-\rho) \cdot 1 [1 + V(S_2)] \\ &= \rho \cdot [1 + V(S_1)] + (1-\rho) \cdot 1 \\ &= \rho + \rho V(S_1) + (1-\rho) \end{aligned}$$

Despejando $V(S_1)$,

$$\begin{aligned} V(S_1) - \rho V(S_1) &= \rho + (1-\rho) \\ V(S_1)(1-\rho) &= 1 \\ V(S_1) &= \frac{1}{1-\rho} \end{aligned}$$

Considerando $\rho = 0.9$,

$$\begin{aligned} V(S_1) &= \frac{1}{1-0.9} \\ &= \frac{1}{0.1} \\ &= 10 \end{aligned}$$

Conclusión

Se observa que el valor teórico del estado no-terminal es 10, mientras que el valor estimado por Monte Carlo de primer-visita es 10 y el valor estimado por Monte Carlo de cada-visita es 5.5.
Por lo tanto, se concluye que el valor estimado por Monte Carlo de primer-visita es igual al valor teórico, mientras que el valor estimado por Monte Carlo de cada-visita es menor al valor teórico.
Por otro lado, la estimación Monte Carlo de cada-visita subestima el valor teórico del estado no-terminal, debido a la naturaleza decreciente de los retornos en cada iteración.

Ejercicio 2

Demostrar que la política ε - *Greedy*, definida de la siguiente manera

$$\pi(a|s) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A(s)|} & si \quad a = a^* \\ \\ \frac{\varepsilon}{|A(s)|} & si \quad a \neq a^* \end{cases} \tag{1}$$

es una distribución de probabilidad válida, donde $|A(s)|$ es el número de acciones para el estado s, $\varepsilon < 1$ es un número positivo pequeño, y a^* es la acción óptima (decisión greedy) para el estado s. ¿Hay que pedir alguna condición sobre ε?

Solución

Para la demostración en cuestión, la política $\pi(a|s)$ es:

$$\pi(a|s) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A(s)|} & si \quad a = a^* \\ \\ \frac{\varepsilon}{|A(s)|} & si \quad a \neq a^* \end{cases} \tag{2}$$

$a = a^*$ La probabilidad corresponde a:

$$\pi\left(a^* \mid s\right) = 1-\varepsilon+\frac{\varepsilon}{|A(s)|}$$

Considerando que $0 \leq \varepsilon < 1$ y $\frac{\varepsilon}{|A(s)|} \geq 0$, se tiene que:

$$\pi\left(a^* \mid s\right) \geq 1-\varepsilon > 0$$

$a \neq a^*$ La probabilidad corresponde a:

$$\pi\left(a \mid s\right) = \frac{\varepsilon}{|A(s)|}$$

Considerando que $\varepsilon \geq 0$ y $|A(s)| > 0$, se tiene que:

$$\pi\left(a \mid s\right) \geq 0$$

Por lo tanto, la política $\pi(a|s) \geq 0$ para toda acción.

Suma de probabilidades La suma de las probabilidades para todas las acciones es:

$$\sum_{a \in A(s)} \pi\left(a \mid s\right) = \pi\left(a^* \mid s\right) + \sum_{a \neq a^*} \pi\left(a \mid s\right)$$

Si $a = a^*$, entonces:

$$\pi\left(a^{*} \mid s\right)=1-\varepsilon+\frac{\varepsilon}{|A(s)|}$$

Para $a \neq a^{*}$, existen $|A(s)|-1$ acciones no óptimas, por lo que:

$$\sum_{a \neq a^{*}} \pi\left(a \mid s\right)=(|A(s)|-1) \cdot \frac{\varepsilon}{|A(s)|}$$

Por lo tanto, la suma de las probabilidades es:

$$\sum_{a \neq a^{*}} \pi\left(a \mid s\right)=1-\varepsilon+\frac{\varepsilon}{|A(s)|}+(|A(s)|-1) \cdot \frac{\varepsilon}{|A(s)|}$$

Al separar el término $\frac{\varepsilon}{|A(s)|}$, se tiene:

$$\sum_{a \in A(s)} \pi\left(a \mid s\right)=1-\varepsilon+\varepsilon$$

$$\sum_{a \in A(s)} \pi\left(a \mid s\right)=1$$

Por lo tanto, la política $\pi(a|s)$ es una distribución de probabilidad válida ya que suman 1.

Condición sobre ε La condición sobre ε es que $\varepsilon \geq 0$ y $\varepsilon < 1$. De tal manera se garantiza:

- Que la probabilidad de la acción óptima a^{*}, $1-\varepsilon+\frac{\varepsilon}{|A(s)|}$ sea mayor a 0.
- Las probabilidades de las acciones no óptimas, $\frac{\varepsilon}{|A(s)|}$, sean mayores a 0.

Ejercicio 3

Se dice que una política π es ε - *soft* si $\pi(a \mid s) \geq \frac{\varepsilon}{|A(s)|}$ $\forall a \neq a^{*}$ y $\forall s$. Demostrar que la política ε - *Greedy* $\pi'(a \mid s) \geq \frac{\varepsilon}{|A(s)|}$ (definida en el Ejercicio 1) es igual o mejor que cualquier política ε - *soft*, es decir $v_{\pi'}(s) \geq v_{\pi}(s)$ $\forall s$.

Solución

El valor de un estado s bajo una política π es:

$$v_{\pi}(s)=\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} R(s_t, a_t) \mid s_0=s\right]$$

donde,
γ: factor de descuento,
a_t: acciones elegidas según la política $\pi(a_t \mid s_t)$,

Política $\varepsilon -$ *soft* Una política π es $\varepsilon -$ *soft* si [1]:

$$\pi(a \mid s) \geq \frac{\varepsilon}{|A(s)|} \ \forall a \neq a^{*} \text{ y } \forall s$$

Política $\varepsilon -$ *Greedy* La política $\varepsilon -$ *Greedy* es:

$$\pi'(a \mid s) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|A(s)|} & si \quad a = a^* \\ \\ \frac{\varepsilon}{|A(s)|} & si \quad a \neq a^* \end{cases} \tag{3}$$

La política $\pi'(a \mid s)$ favorece la acción *Greedy*, y las demás acciones tienen una probabilidad uniforme $\frac{\varepsilon}{|A(s)|}$.

Comparación de políticas Para comparar las políticas π y π', se considera el valor de un estado s bajo la política π y π'.

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a \mid s) \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) v_\pi(s') \right]$$

Se analizarán los valores de $v_{\pi'}(s)$ y $v_\pi(s)$ a fin de comparar si π' o π maximizan el valor, respectivamente.

Valor de $v_{\pi'}(s)$ La política π' favorece la acción *Greedy*, que maximiza la recompensa inmediata y el valor futuro,

- Para $a = a^*$, $\pi'(a \mid s) = 1 - \varepsilon + \frac{\varepsilon}{|A(s)|}$, es mayor que la probabilidad mínima de una política $\varepsilon -$ *soft*.
- Para $a \neq a^*$, $\pi'(a \mid s) = \frac{\varepsilon}{|A(s)|}$, es igual a la probabilidad mínima de una política $\varepsilon -$ *soft*.

Para cualquier estado s, el valor de $v_{\pi'}(s)$ es mayor o igual al valor de $v_\pi(s)$ ya que:

- La recompensa inmediata asociada con a^* es mínimamente el valor el valor de cualquier otra acción.
- El valor futuro descontado con a^* es mayor que el de cualquier otra acción, ya que a^* es *Greedy*.

Si se comparan las ecuaciones de Bellman para $v_{\pi'}(s)$ y $v_\pi(s)$:

$$v_{\pi'}(s) = \sum_{a \in A(s)} \pi'(a \mid s) \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) v_{\pi'}(s') \right]$$
$$v_\pi(s) = \sum_{a \in A(s)} \pi(a \mid s) \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) v_\pi(s') \right]$$

Teniendo en cuenta que $\pi'(a^* \mid s) > \pi(a^* \mid s) \geq \pi(a \mid s)$ para $a \neq a^*$ y que a^* maximiza la recompensa y el valor futuro, se concluye que:

$$v_{\pi'}(s) \geq v_\pi(s) \ \forall s$$

Se requieren las siguientes condiciones para ε:

- π debe ser $\varepsilon -$ *soft*, es decir, $\pi(a \mid s) \geq \frac{\varepsilon}{|A(s)|} \forall a$
- $\varepsilon \geq 0$ y corresponde a un valor pequeño: $0 \leq \varepsilon < 1$.

Ejercicio 4

Dada una trayectoria de acciones y estados $A_t,\ S_t,\ A_{t+1},\ S_{t+1},\ldots,A_T,\ S_T$, en un proceso de decisión Markoviano (MDP) bajo la política $\pi(a \mid s)$, demostrar que la probabilidad conjunta de esa trayectoria se puede escribir como:

$$P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T] = P[S_t] \prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)$$

Nota: considere $P[S_t] = 1$.

Solución

Sea,
$\pi(a \mid s)$: política que determina la probabilidad de tomar una acción $A_k = a$ en el estado $S_k = s$,
$p(S_{k+1} \mid S_k, A_k)$: Modelo de transición que determina la probabilidad de pasar al estado S_{k+1} desde el estado S_k al tomar la acción A_k,
$P[S_t]$: Probabilidad de estar en el estado S_t al inicio de la trayectoria. En este caso, se considera $P[S_t] = 1$.
Sea una trayectoria de acciones y estados $A_t,\ S_t,\ A_{t+1},\ S_{t+1},\ldots,A_T,\ S_T$ en un proceso de decisión Markoviano (MDP) bajo la política $\pi(a \mid s)$. Se expande la probabilidad conjunta de la trayectoria como:

$$P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T] = P[S_t]{\cdot}P[A_t \mid S_t]{\cdot}P[S_{t+1} \mid S_t, A_t]{\cdot}P[A_{t+1} \mid S_{t+1}] \ldots P[S_T \mid S_{T-1}, A_{T-1}]P[S_T]$$

En el proceso de decisión Markoviano (MDP), se tiene que la pólitica y el modelo de transición corresponden a, respectivamente:

$$P[A_K \mid S_K] = \pi(A_K \mid S_K)$$
$$P[S_{t+1} \mid S_k, A_k] = p(S_{k+1} \mid S_k, A_k)$$

Al sustituir la política y el modelo de transición en la probabilidad conjunta de la trayectoria, se obtiene:

$$P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T] = P[S_t] \cdot \pi(A_t \mid S_t) \cdot p(S_{t+1} \mid S_t, A_t) \cdot \pi(A_{t+1} \mid S_{t+1}) \ldots$$

Se expresa como un producto:

$$P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T] = P[S_t] \prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)$$

Teniendo en cuenta que $P[S_t] = 1$, la probabilidad conjunta se simplifica a:

$$P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T] = \prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)$$

Ejercicio 5

Usando el resultado del *Ejercicio 4*, demostrar que el *importance-sampling ratio* correspondiente a la aplicación del método *off-policy* es:

$$\rho_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$$

Solución

Sea,
$\pi(a \mid s)$: Política objetivo a evaluar,
$p(S_{k+1} \mid S_k, A_k)$: Modelo de transición del entorno,
$b(A_k \mid S_k)$: Política de comportamiento utilizada para generar la trayectoria,

Del *Ejercicio 4* se tiene que la probabilidad conjunta de una trayectoria bajo la política $\pi(a \mid s)$ es:

$$P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T] = P[S_t] \prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)$$

Por otro lado, la probabilidad conjunta de la trayectoria bajo la política $b(a \mid s)$ es:

$$P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T] = P[S_t] \prod_{k=t}^{T-1} b(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)$$

El *importance-sampling ratio* mide la relación entre la probabilidad de una trayectoria bajo la política objetivo π y la políticia de comportamiento b.

$$\rho_{t:T-1} = \frac{P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T]}{P[S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T, A_T]}$$

Tras sustituir las expresiones de las probabilidades conjuntas se obtuvo:

$$\rho_{t:T-1} = \frac{P[S_t] \prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)}{P[S_t] \prod_{k=t}^{T-1} b(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)}$$

Tras simplificar el término $P[S_t]$ se obtuvo:

$$\rho_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)}$$

Ejercicio 6

Demostrar la fórmula de la implementación incremental de un promedio ponderado. Es decir, el promedio ponderado

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

puede calcularse incrementalmente con la fórmula:

$$V_{n+1} = V_n + \frac{W_n}{C_n} \left[G_n - V_n \right],$$

con $C_{n+1} = C_n + W_{n+1}$

Solución

Considerando $C_n = \sum_{k=1}^{n-1} W_k$ se escribe V_n como:

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{C_n}$$

Para V_{n+1}, se tiene:

$$V_{n+1} = \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k}$$

De la expresión anterior, el numerador y el demoninador corresponden a, respectivamente:

$$\sum_{k=1}^{n} W_k G_k = \sum_{k=1}^{n-1} W_k G_k + W_n G_n$$
$$\sum_{k=1}^{n} W_k = \sum_{k=1}^{n-1} W_k + W_n$$

Sustituyendo en V_{n+1}, se obtiene:

$$V_{n+1} = \frac{\sum_{k=1}^{n-1} W_k G_k + W_n G_n}{\sum_{k=1}^{n-1} W_k + W_n}$$

El denominador se puede escribir como $C_n + W_n$, por lo que:

$$V_{n+1} = \frac{\sum_{k=1}^{n-1} W_k G_k + W_n G_n}{C_n + W_n}$$

Se remplaza en el numerador por $\sum_{k=1}^{n-1} W_k G_k = V_n C_n$:

$$V_{n+1} = \frac{V_n C_n + W_n G_n}{C_n + W_n}$$

Tras dividir el numerador entre $C_n + W_n$, se obtuvo:

$$V_{n+1} = \frac{C_n V_n}{C_n + W_n} + \frac{W_n G_n}{C_n + W_n}$$

Se simplifica la expresión:

$$V_{n+1} = V_n \frac{C_n}{C_n + W_n} + G_n \frac{W_n}{C_n + W_n}$$

En el primer término $\frac{C_n}{C_n+W_n}$ corresponde a:

$$\frac{C_n}{C_n + W_n} = 1 - \frac{W_n}{C_n + W_n}$$

Tras sustituir en V_{n+1}, se obtiene:

$$V_{n+1} = V_n \left(1 - \frac{W_n}{C_n + W_n}\right) + G_n \frac{W_n}{C_n + W_n}$$

Se simplifica la expresión:

$$V_{n+1} = V_n - V_n \frac{W_n}{C_n + W_n} + G_n \frac{W_n}{C_n + W_n}$$

Se agrupa los términos que contienen $\frac{W_n}{C_n+W_n}$:

$$V_{n+1} = V_n + \frac{W_n}{C_n + W_n} (G_n - V_n)$$

Del enunciado del ejercicio, se tiene que $C_{n+1} = C_n + W_{n+1}$, por lo que:

$$C_{n+1} = C_n + W_{n+1}$$

Por lo tanto, la fórmula de la implementación incremental de un promedio ponderado es:

$$V_{n+1} = V_n + \frac{W_n}{C_n}\left[G_n - V_n\right]$$

Conclusión El promedio ponderado:

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

Puede calcularse incrementalmente con la fórmula:

$$V_{n+1} = V_n + \frac{W_n}{C_n}\left[G_n - V_n\right],$$

con $C_{n+1} = C_n + W_{n+1}$.

Referencias

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* The MIT Press, second ed., 2018.