Ejercicio 1

Para un proceso de Markov donde $p_{ss'} = P[S_{t+1} = s' \mid S_t = s]$ son las probabilidades de transición, demostrar que la probabilidad de un episodio se puede escribir como el producto de las probabilidades de transición y la condición inicial, es decir:

$$P[S_0 = s_0, S_1 = s_1, \ldots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{(t+1)}} p_{s_{(t-1)} s_t} \cdots p_{s_1 s_2} p_{s_0 s_1} p_{s_0},$$

donde $s_n \in S$ y $p_{s_0} P[S_0 = s_0]$.

Solución

Sea,
S: Secuencia de estados

Se aplica la regla de la cadena para la probabilidad de una secuencia de eventos, para un proceso de Markov,

$$\begin{aligned} P[S_0 = s_0, S_1 = s_1, \ldots, S_{t+1} = s_{t+1}] &= P[S_{t+1} = s_{t+1}, S_t = s_t, \ldots, S_0 = s_0] \\ &= P[S_{t+1} = s_{t+1} \mid S_t = s_t, S_{t-1} = s_{t-1}, S_0 = s_0]\ P[S_t = s_t, \ldots, S_0 = s_0] \end{aligned}$$

Por la propiedad de Markov, la probabilidad de transición de un estado a otro depende solo del estado actual, por lo que la probabilidad de transición de un estado a otro depende solo del estado actual, matemáticamente se expresa como,

$$\begin{aligned} P[S_{t+1} = s_{t+1} \mid S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0] &= P[S_{t+1} = s_{t+1} \mid S_t = s_t] \\ &= p_{s_t s_{t+1}} \end{aligned}$$

Por lo tanto, la probabilidad de una secuencia de eventos se puede escribir como,

$$P[S_0 = s_0, S_1 = s_1, \ldots, S_t = s_t, S_{t+1} = s_{t+1}] = P[S_{t+1} \mid S_t = s_t]\ P[S_t = s_t, \ldots, S_0 = s_0]$$

Se aplica recursivamente la regla de la cadena para obtener la probabilidad conjunta en términos de las probabilidades de transición,

$$P[S_0 = s_0, S_1 = s_1, \ldots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{t+1}}\ P[S_t = s_t, \ldots, S_0 = s_0]$$

Se aplica de nuevo la regla de la cadena para obtener la probabilidad conjunta en términos de las probabilidades de transición,

$$P[S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0] = p_{s_{t-1} s_t}\ P[S_{t-1} = s_{t-1}, \ldots, S_0 = s_0]$$

Tras aplicar la propiedad de Markov se llegó a que,

$$P[S_1 = s_1, S_0 = s_0] = p_{s_0 s_1}\ P[S_0 = s_0]$$

La probabilidad conjunta de la secuencia completa de estados es,

$$P[S_0 = s_0, S_1 = s_1, \ldots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{t+1}}\ p_{s_{t-1} s_t} \cdots p_{s_1 s_2}\ p_{s_0 s_1}\ p_{s_0} P[S_0 = s_0]$$

Si se define $p_{s_0} = P[S_0 = s_0]$, la expresión se simplifica a,

$$P[S_0 = s_0, S_1 = s_1, \ldots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{t+1}}\ p_{s_{t-1} s_t} \cdots p_{s_1 s_2}\ p_{s_0 s_1}\ p_{s_0}$$

Se concluye que la probabilidad de un episodio se puede escribir como el producto de las probabilidades de transición y la condición inicial p_{s_0}.

Ejercicio 2

Para el proceso de Markov, de la figura 1, calcular la probabilidad de cada uno de los siguientes episodios condicionados al estado inicial *C1*:

1. C1 C2 C3 Pass Sleep
2. C1 FB FB C1 C2 Sleep
3. C1 C2 C3 Pub C2 C3 Pass Sleep
4. C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

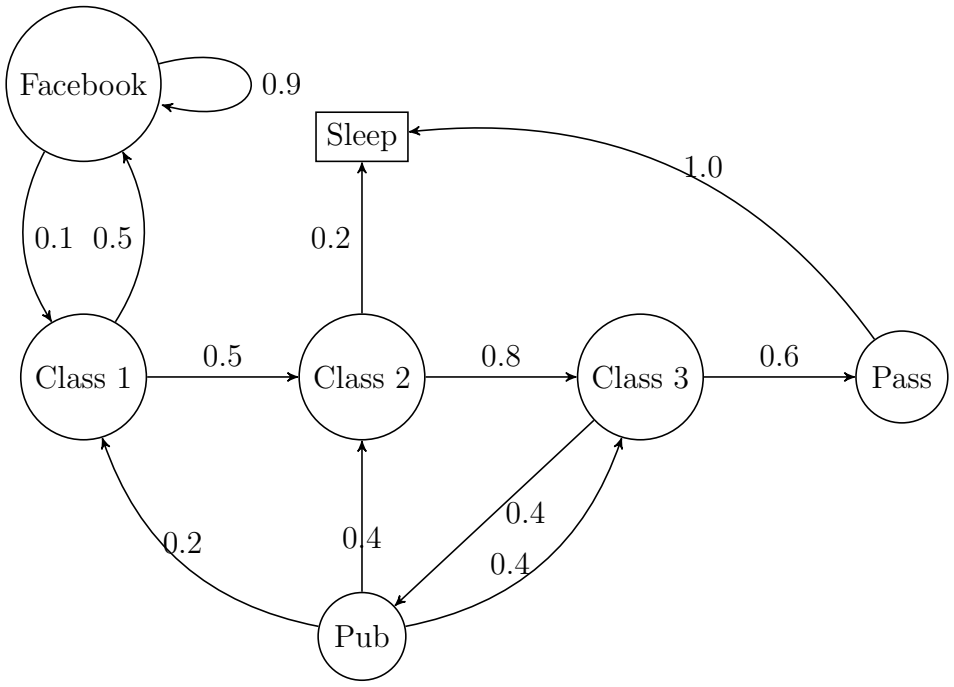

Figura 1: Grafo de transición de estados

Solución

Ejercicio 3

Demostrar que si la recompensa es constante $R_t = R \; \forall t$ y el factor de descuento $\gamma < 1$, entonces,

$$G_t = \frac{R}{1-\gamma}$$

Solución

El retorno G_t es la suma de las recompensas futuras descontadas a partir del tiempo t, es decir,

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$$

Dado que la recompensa es constante, $R_t = R \; \forall t$, entonces, $R_{t+1} = R$, $R_{t+2} = R$, ..., por lo que,

$$G_t = R + \gamma R + \gamma^2 R + \gamma^3 R + \dots$$

Tras factorizar toma la forma,

$$G_t = R(1 + \gamma + \gamma^2 + \dots)$$

La serie geométrica $1 + \gamma + \gamma^2 + \dots$ es infinita. Sin embargo, dado que $\gamma < 1$, la serie converge a,

$$\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

Por lo tanto,

$$1 + \gamma + \gamma^2 + \dots = \frac{1}{1-\gamma}$$

Que al remplazar en la expresión de G_t se obtiene,

$$G_t = R\left(\frac{1}{1-\gamma}\right)$$
$$G_t = \frac{R}{1-\gamma}$$

Ejercicio 4

Para el proceso de Recompensas Markoviano (MRP) de la figura 2, el retorno G_0 de cada uno de los siguientes episodios con estado inicial C1 y $\gamma = 0.5$:

1. C1 C2 C3 Pass Sleep
2. C1 FB FB C1 C2 Sleep
3. C1 C2 C3 Pub C2 C3 Pass Sleep
4. C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

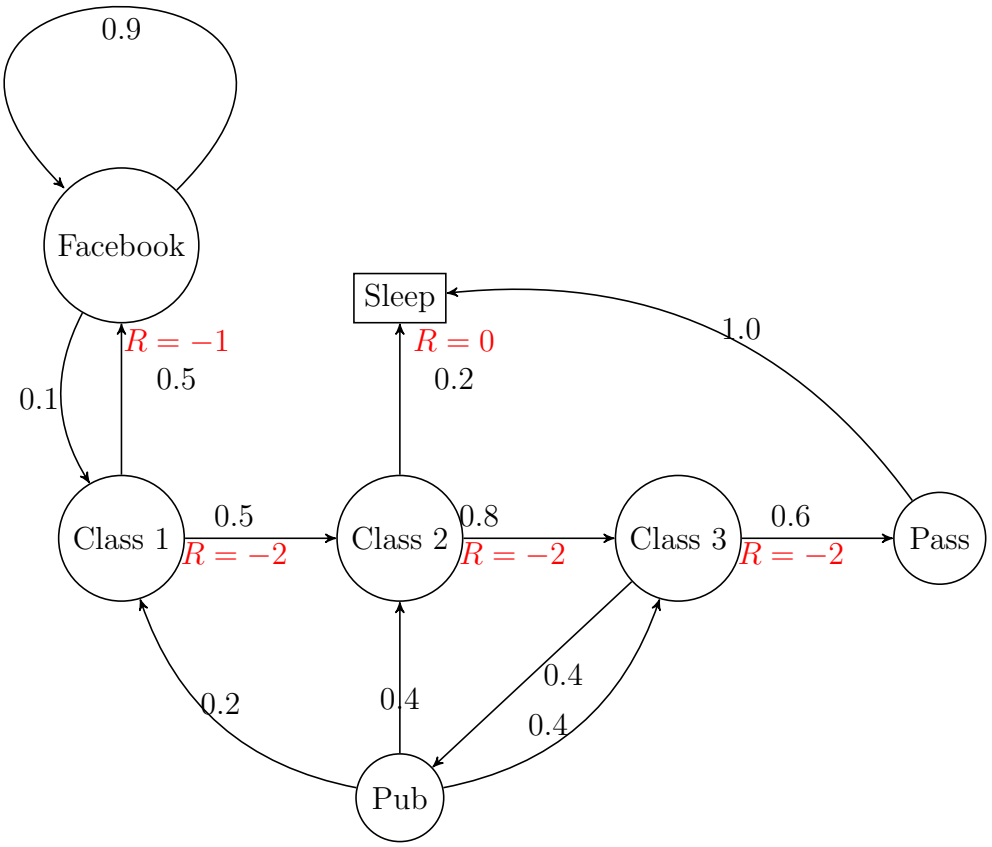

Figura 2: Grafo de transición de estados

Solución

Ejercicio 5

Demostrar que si $\gamma = 0$ y $R_t = R(S_t)$ (la recompensa depende del estado), entonces

$$v(s) = \sum_{s' \in S} p(S_{t+1} = s' | S_t = s) R(S_{t+1})$$

Solución

Sea,
$v(s)$: Valor esperado del retorno a partir del estado s.
$R(S_t)$: Recompensa en el estado S_t.
G_t: Retorno a partir del tiempo t y $S_t = s$.
Por definición, el valor esperado del retorno a partir del estado s es,

$$v(s) = E[G_t | S_t = s]$$

Donde, G_t está dado por,

$$G_t = R_t + \gamma R(_{t+1} + \gamma^2 R_{t+2} + \dots$$

Si $\gamma = 0$ y $R_t = R(S_t)$, entonces el agente no considera el futuro y la recompensa depende del estado actual. Por lo tanto, el retorno G_t se reduce a,

$$G_t = R_t = R(S_t)$$

Por otro lado, $v(s)$ se puede expresar en términos del siguiente estado S_{t+1} y la recompensa $R(S_{t+1})$ como,

$$v(s) = E[G_t|S_t = s] = E[R(S_{t+1})|S_t = s]$$

Por lo tanto, el valor esperado del retorno a partir del estado s es la suma de las recompensas esperadas en el siguiente estado ponderadas por la probabilidad de transición al siguiente estado, es decir,

$$v(s) = \sum_{s' \in S} p(S_{t+1} = s'|S_t = s)R(S_{t+1})$$

En otras palabras, el valor de un estado es la expectativa de la recompensa inmediata en el siguiente estado ponderada por la probabilidad de transición al siguiente estado.

Ejercicio 6

Demostrar la ecuación de *Bellman* para MRPs con N estados

$$v = r + \gamma P v,$$

donde $v \in \mathbb{R}^N$ es el vector de valores de estados, $r \in \mathbb{R}^N$ es el vector de recompensas medias partiendo de cada estado, $P \in \mathbb{R}^{N \times N}$ es la matriz de transiciones y γ es el factor de descuento.

Solución

Ejercicio 7

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Solución

Ejercicio 8

Demostrar que, dada una política estocástica $\pi(a \mid s)$, la función de valor de estado puede escribirse como

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

Solución

Ejercicio 9

Demostrar que, dada una política estocástica $\pi(a \mid s)$, la función de valor de estado puede escribirse como

$$q_\pi(s, a) = \sum_{s', r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]$$

Solución

Ejercicio 10

Demostrar que la función de valor de estado óptima es

$$v_*(s) = \max_a q_*(s,a) = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')]$$

Solución

Ejercicio 11

Demostrar que la función de valor esado-acción verifica la siguiente ecuación:

$$q_\pi(s,a) = \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \sum_{a'} q_\pi(s', a')\pi(a' \mid s') \right]$$

Solución

Ejercicio 12

Demostrar que la función de valor estado-acción óptima es:

$$q_*(s,a) = \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right]$$

Solución

Referencias