Guía 2: Procesos de Decisión Markovianos (MDP)

Ejercicio 1

Para un proceso de Markov donde $p_{ss'} = P[S_{t+1} = s' | S_t = s]$ 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:

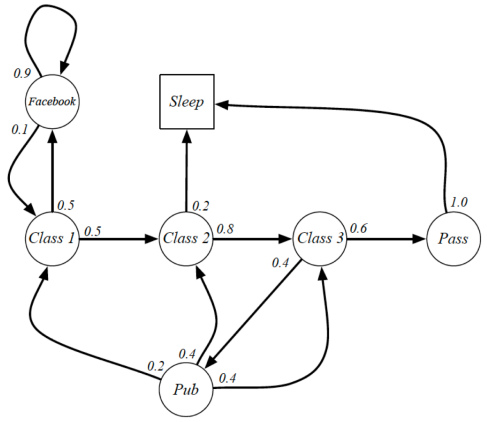
$$P[S_0 = s_0, S_1 = s_1, \ldots, S_t = s_t, S_{t+1} = s_{t+1}] = p_{s_t s_{(t+1)}} p_{s_{(t-1)} s_t} \ldots p_{s_1 s_2} p_{s_0 s_1} p_{s_0},$$

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

Ejercicio 2

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

a) C1 C2 C3 Pass Sleep
b) C1 FB FB C1 C2 Sleep
c) C1 C2 C3 Pub C2 C3 Pass Sleep
d) C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep



Ejercicio 3

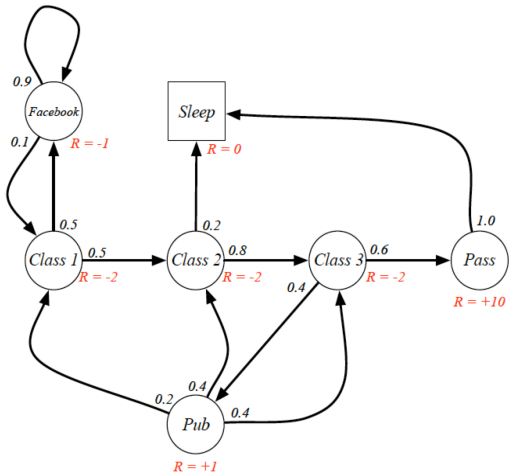
Demostrar que si la recompensa es constante $R_t = R \; \forall t$ y el factor de descuento es $\gamma < 1$, entonces

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

Ejercicio 4

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

a) C1 C2 C3 Pass Sleep
b) C1 FB FB C1 C2 Sleep
c) C1 C2 C3 Pub C2 C3 Pass Sleep
d) C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep



Ejercicio 5

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

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

Ejercicio 6

Demostrar la ecuación de Bellman para MRPs con N estados

$$\mathbf{v} = \mathbf{r} + \gamma \mathbf{P}\mathbf{v},$$

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

Ejercicio 7

[Programación] Resuelva la ecuación de Bellman para el MRP del Ejercicio 4 por los siguientes métodos para $\gamma = 0,\ 0.5, 0.9$:

a) Iterando $\mathbf{v_{n+1}} = \mathbf{r} + \gamma \mathbf{P}\mathbf{v_n}$
b) Usando algún método iterativo para resolver sistemas lineales, por ejemplo, usando eliminación Gaussiana (vía numpy.linalg.solve())
c) Calculando la inversa de la matriz $\mathbf{I} - \gamma \mathbf{P}$

Ejercicio 8

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

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s, a).$$

Ejercicio 9

Demostrar que, dada una política estocástica $\pi(a|s)$, la función de valor del par estado-acción puede escribirse como

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

Ejercicio 10

Demostrar que la función de valor de estado óptima es

$$v_*(s) = max_a \mathrm{q}_*(\mathrm{s}, \mathrm{a}) = max_a \sum_{s', r} p(s', r|s, a)[r + \gamma v_*(s')].$$

Ejercicio 11

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

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

Ejercicio 12

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

$$q_*(s, a) = \sum_{s', r} p(s', r|s, a)[r + \gamma max_{a'} \mathrm{q}_*(\mathrm{s'}, \mathrm{a'})].$$