

Nota: la función de probabilidades de transición P se representará, en general, como un conjunto de tablas representando $P_a(\cdot | \cdot)$ para cada acción a . En esas tablas, cada fila corresponde a un estado en el que la acción es ejecutable y en cada columna se indica la probabilidad de transitar al estado correspondiente.

1. Consideremos el proceso de decisión de Markov tal que $S = \{s_1, s_2, s_3\}$, $A = \{a_1, a_2, a_3\}$ y P viene dado por:

$$P_{a_1}(\cdot | \cdot) = \begin{array}{c|ccc} & s_1 & s_2 & s_3 \\ \hline s_2 & 0.4 & 0.1 & 0.5 \\ s_3 & 0.5 & 0.0 & 0.5 \end{array} \quad P_{a_2}(\cdot | \cdot) = \begin{array}{c|ccc} & s_1 & s_2 & s_3 \\ \hline s_1 & 0.0 & 0.3 & 0.7 \\ s_3 & 0.0 & 0.5 & 0.5 \end{array}$$

$$P_{a_3}(\cdot | \cdot) = \begin{array}{c|ccc} & s_1 & s_2 & s_3 \\ \hline s_1 & 0.0 & 0.3 & 0.7 \\ s_2 & 0.8 & 0.2 & 0.0 \end{array}$$

Consideremos $R(s_1) = -1$, $R(s_2) = -0.04$ y $R(s_3) = 1$ como recompensas de los estados, 0 como coste de aplicar las acciones y 0.9 como factor de descuento.

Dada la política $\pi(s_1) = a_3$, $\pi(s_2) = a_3$ y $\pi(s_3) = a_2$, se pide lo siguiente:

- ¿Cuál es la probabilidad inducida por π de la historia (parcial) $\langle s_3, s_3, s_3, s_2, s_2 \rangle$?
¿Cuál es la utilidad inducida por π de esa historia?
- Plantear el sistema de ecuaciones que caracteriza U_π .
- Plantear las ecuaciones de Bellman que caracterizan U^* .
- Supongamos que hemos resuelto las ecuaciones anteriores y que conocemos U^* . Describir cómo podríamos obtener una política óptima.

2. Consideremos el proceso de decisión de Markov tal que $S = \{s_1, s_2, s_3, s_4\}$, $A = \{a_1, a_2, a_3\}$ y P viene dado por:

$$P_{a_1}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_1 & 0.0 & 0.0 & 0.2 & 0.8 \\ s_4 & 0.0 & 0.0 & 0.5 & 0.5 \end{array} \quad P_{a_2}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_1 & 1/3 & 1/3 & 1/3 & 0.0 \\ s_3 & 1/3 & 1/3 & 0.0 & 1/3 \end{array}$$

$$P_{a_3}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_2 & 1.0 & 0.0 & 0.0 & 0.0 \\ s_3 & 1/3 & 1/3 & 0.0 & 1/3 \end{array}$$

Consideremos $R(s_1) = -3$, $R(s_2) = -2$, $R(s_3) = 1$ y $R(s_4) = 1$ como recompensas de los estados, $C(s_1, a_1) = C(s_4, a_1) = 2$, $C(s_1, a_2) = C(s_3, a_2) = C(s_3, a_3) = 3$ y $C(s_2, a_3) = 1$ como costes de aplicar las acciones y 0.5 como factor de descuento.

Dada la política $\pi(s_1) = a_1$, $\pi(s_2) = a_3$, $\pi(s_3) = a_2$ y $\pi(s_4) = a_1$, se pide lo siguiente:

- Calcular $U_\pi(s)$ para cada $s \in S$, planteando y resolviendo el sistema de ecuaciones que caracteriza U_π .

- Considerando $U_0(s_1) = -2, U_0(s_2) = -1, U_0(s_3) = 1, U_0(s_4) = 2$ como función de utilidad inicial, calcular la función de utilidad que se obtiene al ejecutar una iteración del algoritmo de iteración de valores.
 - Describir hasta cuando el algoritmo anterior seguiría realizando iteraciones y cómo se obtendría entonces una política óptima.
3. Consideremos el proceso de decisión de Markov tal que $S = \{s_1, s_2, s_3, s_4\}, A = \{a_1, a_2\}$ y P viene dado por:

$$P_{a_1}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_1 & 0.0 & 0.5 & 0.5 & 0.0 \\ s_2 & 0.0 & 0.0 & 1.0 & 0.0 \\ s_3 & 1.0 & 0.0 & 0.0 & 0.0 \\ s_4 & 0.0 & 0.0 & 0.0 & 1.0 \end{array} \quad P_{a_2}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_2 & 0.0 & 0.5 & 0.0 & 0.5 \end{array}$$

Consideremos $R(s_1) = 1, R(s_2) = 2, R(s_3) = 3$ y $R(s_4) = 10$ como recompensas de los estados, 0 como coste de aplicar las acciones y 0.9 como factor de descuento.

Se pide lo siguiente:

- Determinar cuántas políticas distintas es posible especificar para este sistema.
 - Dada la política π que aplica la acción a_1 en cada estado, plantear y resolver el sistema de ecuaciones que caracteriza U_π .
 - Plantear las ecuaciones de Bellman que caracterizan U^* .
 - Considerando como función de utilidad inicial la que asocia 0 a cada estado, calcular la función de utilidad que se obtiene al ejecutar dos iteraciones del algoritmo de iteración de valores.
4. Consideremos el proceso de decisión de Markov tal que $S = \{s_1, s_2, s_3, s_4\}, A = \{a_1, a_2, a_3\}$ y P viene dado por:

$$P_{a_1}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_1 & 0.0 & 1.0 & 0.0 & 0.0 \\ s_2 & 1.0 & 0.0 & 0.0 & 0.0 \end{array} \quad P_{a_2}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_1 & 1/3 & 1/3 & 1/3 & 0.0 \\ s_4 & 1/3 & 1/3 & 1/3 & 0.0 \end{array}$$

$$P_{a_3}(\cdot | \cdot) = \begin{array}{c|cccc} & s_1 & s_2 & s_3 & s_4 \\ \hline s_2 & 0.75 & 0.0 & 0.0 & 0.25 \\ s_3 & 0.5 & 0.5 & 0.0 & 0.0 \end{array}$$

Consideremos $R(s_1) = 3, R(s_2) = 0, R(s_3) = 0$ y $R(s_4) = 2$ como recompensas de los estados, $C(s_1, a_1) = C(s_2, a_1) = 1, C(s_1, a_2) = C(s_4, a_2) = 2$ y $C(s_2, a_3) = C(s_3, a_3) = 3$ como costes de aplicar las acciones y 0.5 como factor de descuento.

Se pide lo siguiente:

- Determinar cuántas políticas distintas es posible especificar para este sistema.
- Dada la política $\pi(s_1) = a_1, \pi(s_2) = a_1, \pi(s_3) = a_3$ y $\pi(s_4) = a_2$, plantear y resolver el sistema de ecuaciones que caracteriza U_π .

- Plantear las ecuaciones de Bellman que caracterizan U^* .
 - Considerando π como política inicial, calcular la política que se obtiene al ejecutar una iteración del algoritmo de iteración de políticas.
5. A lo largo de su vida, una empresa pasa por situaciones muy distintas que, por simplificar, resumiremos en que al inicio de cada campaña puede estar rica o pobre y ser conocida o desconocida. Para ello puede decidir en cada momento o bien invertir en publicidad, o bien optar por no hacer publicidad. Estas dos acciones no tienen siempre un resultado fijo, aunque podemos describirlo de manera probabilística:
- Si la empresa es rica y conocida y no invierte en publicidad, seguirá rica, pero existe la posibilidad del 50 % de que se vuelva desconocida. Si gasta en publicidad, con toda seguridad seguirá conocida, pero pasará a ser pobre.
 - Si la empresa es rica y desconocida y no gasta en publicidad, seguirá desconocida y, además, existe un 50 % de posibilidad de que se vuelva pobre. Si gasta en publicidad, se volverá pobre, pero existe un 50 % de posibilidad de que se vuelva conocida.
 - Si la empresa es pobre y conocida y no invierte en publicidad, pasará a ser pobre y desconocida con un 50 % de probabilidad, y rica y conocida en caso contrario. Si gasta en publicidad, con toda seguridad seguirá en la misma situación.
 - Si la empresa es pobre y desconocida y no invierte en publicidad, seguirá en la misma situación con toda seguridad. Si gasta en publicidad, seguirá pobre, pero con un 50 % de posibilidad de pasar a ser conocida.

Supondremos que la recompensa en las campañas en la que la empresa es rica es de 10 y de 0 en las que es pobre. El objetivo es conseguir la mayor recompensa acumulada a lo largo del tiempo, aunque penalizaremos las ganancias obtenidas en campañas muy lejanas en el tiempo, introduciendo un factor de descuento de 0.9.

Se pide lo siguiente:

- Representar lo anterior como un proceso de decisión de Markov.
- Si π es la política que consiste en invertir siempre en publicidad, plantear y resolver el sistema de ecuaciones que caracteriza U_π .
- Plantear las ecuaciones de Bellman que caracterizan U^* .
- Considerando π como política inicial, calcular la política que se obtiene al ejecutar una iteración del algoritmo de iteración de políticas.

IA - Boleín 5 - Formulario

► Probabilidad inducida

$$\prod P_{ncs_i}(s_{i+1} | s_i)$$

► Utilidad inducida

$$\sum \gamma^i R C s_i)$$

► Sistema de ecuaciones (U_{π})

$$U(s) = R(s, \pi(s)) + \gamma \sum P_{ncs}(s' | s) U(s')$$

$s_1, s_2, s_3 \dots$

Los valores de la
política π para
esa s

s que estás usando y todas las
demás

s que estás usando

► Ecuaciones de Bellman

$$U(s) = \max \left[R(s, a) + \gamma \sum P_a(s' | s) U(s') \right]$$

\downarrow
 $a_1, a_2, a_3 \dots$

1)

■ Probabilidad inducida $\rightarrow \prod P_n(s_i) (s_{i+1} | s_i)$
 $\langle s_3, s_3, s_3, s_2, s_2 \rangle$ $P(s_1)=a_3$ $P(s_2)=a_3$ $P(s_3)=a_2$
 $P_{a_2}(s_3|s_3) \cdot P_{a_2}(s_3|s_3) \cdot P_{a_2}(s_2|s_3) \cdot P_3(s_2|s_2)$

Utilidad inducida $\rightarrow \sum \delta^i R(s_i)$

$R(s_3) + \delta^1 q R(s_3) + \delta^1 q^2 R(s_3) + \delta^1 q^3 R(s_2) + \delta^1 q^4 R(s_2)$

■ Sist. ecuaciones $U_\pi \rightarrow U(s) = R(s, \pi(s)) + \delta \sum P_n(s') U(s')$
 El que usa en esa U
 Todas las demás

$U(s_1) = R(s_1, a_3) + \delta^1 q [P_{a_3}(s_1|s_1) U(s_1) + P_{a_3}(s_2|s_1) U(s_2) + P_{a_3}(s_3|s_1) U(s_3)]$

$U(s_2) = R(s_2, a_3) + \delta^1 q [P_{a_3}(s_1|s_2) U(s_1) + P_{a_3}(s_2|s_2) U(s_2) + P_{a_3}(s_3|s_2) U(s_3)]$

$U(s_3) = \underbrace{R(s_3, a_2)}_{\text{recursivo}} + \delta^1 q [P_{a_2}(s_1|s_3) U(s_1) + P_{a_2}(s_2|s_3) U(s_2) + P_{a_2}(s_3|s_3) U(s_3)]$
 cost=0, no se escribe

■ Ecuaciones de Bellman $\rightarrow U^* = \max (R(s, a) + \delta \sum P_e(s'|s) U(s'))$

Para sustituir $P(s, i)$, miras la tabla

$U(s_1) = \max [$
 $R(s_1, a_1) + \delta^1 q [P_{a_1}(s_1|s_1) U(s_1) + P_{a_1}(s_2|s_1) U(s_2) + P_{a_1}(s_3|s_1) U(s_3)]$
 $R(s_1, a_2) + \delta^1 q [P_{a_2}(s_1|s_1) U(s_1) + P_{a_2}(s_2|s_1) U(s_2) + P_{a_2}(s_3|s_1) U(s_3)]$
 $R(s_1, a_3) + \delta^1 q [P_{a_3}(s_1|s_1) U(s_1) + P_{a_3}(s_2|s_1) U(s_2) + P_{a_3}(s_3|s_1) U(s_3)]$
 $]$

$U(s_1) = \max [R(s_1, a_1) + \delta^1 q (\delta^1 U(s_2) + \delta^1 U(s_3)) , R(s_1, a_2) + \delta^1 q (\delta^1 U(s_2) + \delta^1 U(s_3)) + \delta^1 U(s_3)]$

Fixate que como el primero queda solo $R(s_1)$, me se usa solo

■ Políticas óptimas

Señal aquellas políticas tal que $U_\pi = U^*$

2)

$$\begin{aligned} \pi(s_1) &= a_1 & \pi(s_3) &= a_2 \\ \pi(s_2) &= a_3 & \pi(s_4) &= a_1 \end{aligned}$$

■ $U_n(s)$

$$U(s_1) = R(s_1, a_1) + 0.5 [P_{a_1}(s_1|s_1)U(s_1) + P_{a_1}(s_2|s_1)U(s_2) + P_{a_1}(s_3|s_1)U(s_3) + P_{a_1}(s_4|s_1)U(s_4)]$$

$$U(s_1) = R(s_1, a_1) + 0.5 [0.2 U(s_3) + 0.8 U(s_4)]$$

$$U(s_2) = R(s_2, a_3) + 0.5 [1 U(s_1)]$$

$$U(s_3) = R(s_3, a_2) + 0.5 \left[\frac{1}{3} U(s_1) + \frac{1}{3} U(s_2) + \frac{1}{3} U(s_4) \right]$$

■ Funções de utilidade, uma iteração de alg. it. valores(U₁)

$$U_1(s_1) = \max \begin{cases} R(s_1, a_1) + \delta [0.2 \cdot U_0(s_3) + 0.8 U_0(s_4)] \\ R(s_1, a_2) + \delta \left[\frac{1}{3} U_0(s_1) + \frac{1}{3} U_0(s_2) + \frac{1}{3} U_0(s_3) \right] \end{cases}$$

$$U_1(s_1) = \max \begin{cases} -3 - 2 + 0.5 [0.2 \cdot 1 + 0.8 \cdot 2] = -4.1 \\ \boxed{-3} - 3 + 0.5 \left[\frac{1}{3} \cdot 2 + \frac{1}{3} (-1) + \frac{1}{3} \cdot 1 \right] = -6.3 \end{cases}$$

recesso ↗ ↘ custo

$$U_1(s_2) = R(s_2, a_3) + \delta [1 \cdot U_0(s_1)] = -2 - 2 - 1 =$$

$$U_1(s_3) = 1 - 3 + 0.5 \left[\frac{1}{3} U_0(s_1) + \frac{1}{3} U_0(s_2) + \frac{1}{3} U_0(s_4) \right] = -2.17$$

$$U_1(s_4) = 1 - 2 + 0.5 \left[\frac{1}{3} U_0(s_3) + \frac{1}{3} U_0(s_4) \right] = 0.25$$

3)

■ Cantidad de políticas

$$\|S\| \times \|A\| \rightarrow 4 \cdot 2 = 8$$

$$\pi(S_1) = \pi(S_2) = \pi(S_3) = \pi(S_4) = a_1$$

Sistema de ecuaciones

$$U(S_1) = R(S_1, a_1) + \gamma [P_{a_1}(S_1|S_1) U(S_1) + P_{a_1}(S_2|S_1) U(S_2) + P_{a_1}(S_3|S_1) U(S_3) + P_{a_1}(S_4|S_1) U(S_4)]$$

$$U(S_1) = R(S_1, a_1) + \gamma [0.5 U(S_2) + 0.5 U(S_3)]$$

$$U(S_2) = R(S_2, a_1) + \gamma [P_{a_1}(S_1|S_2) U(S_1) + P_{a_1}(S_2|S_2) U(S_2) + P_{a_1}(S_3|S_2) U(S_3) + P_{a_1}(S_4|S_2) U(S_4)]$$

$$U(S_2) = R(S_2, a_1) + \gamma [0.5 U(S_1)]$$

$$U(S_3) = R(S_3, a_1) + \gamma [P_{a_1}(S_1|S_3) U(S_1) + P_{a_1}(S_2|S_3) U(S_2) + P_{a_1}(S_3|S_3) U(S_3) + P_{a_1}(S_4|S_3) U(S_4)]$$

$$U(S_3) = R(S_3, a_1) + \gamma [0.5 U(S_1) + 1 U(S_2)]$$

con $U(S_4)$ igual, me da palo hacerlo

■ Bellman

Como la política usa siempre a_1 , Bellman resuelve $\max (U_{\pi})$

4)

■ Cantidad de políticas

$$||S|| \times ||\lambda|| = 12$$

■

$$U(s_1) = R(s_1, a_1) + \gamma \sigma' S \left[P_{q1}(s_1|s_1) U(s_1) + P_{q1}(s_2|s_1) U(s_2) + P_{q1}(s_3|s_1) U(s_3) + P_{q1}(s_4|s_1) U(s_4) \right]$$

$$U(s_2) = R(s_2, a_1) + \gamma \sigma' S \left[P_{q1}(s_1|s_2) U(s_1) + P_{q1}(s_2|s_2) U(s_2) + P_{q1}(s_3|s_2) U(s_3) + P_{q1}(s_4|s_2) U(s_4) \right]$$

$$U(s_3) = R(s_3, a_3) + \dots$$

$$U(s_4) = R(s_4, a_2) + \dots$$

■ Bellman

$$U(s_1) = \max \left[\right.$$

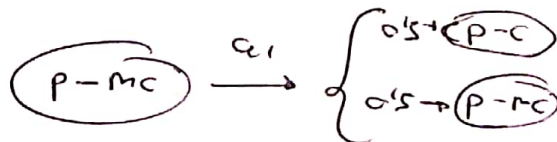
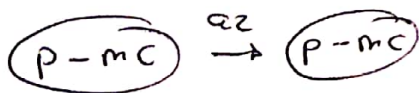
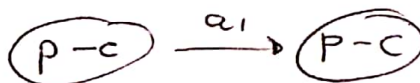
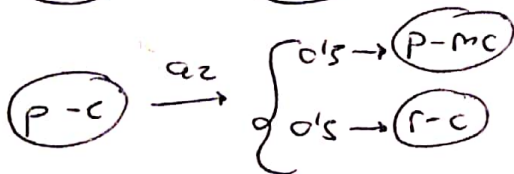
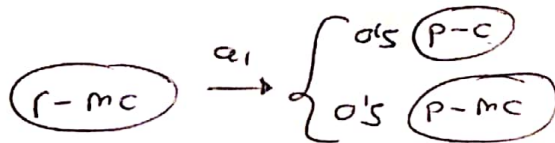
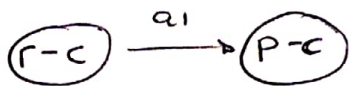
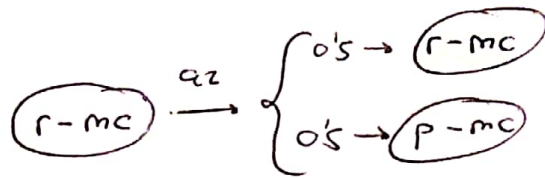
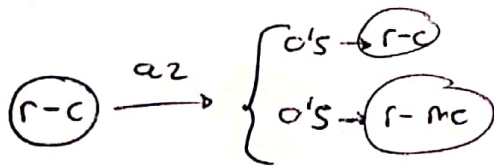
$$R(s_1, a) + \gamma \sigma' S \left[P_{q1}(s_1|s_1) U(s_1) + P_{q1}(s_2|s_1) U(s_2) + P_{q1}(s_3|s_1) U(s_3) + P_{q1}(s_4|s_1) U(s_4) \right]$$

$$R(s_1, a) + \gamma \sigma' S \left[P_{q2}(s_1|s_2) U(s_2) + \dots \right] \text{ y así sigue}$$

5)

	rica	pobre
como	$r-c$	$p-c$
no como	$r-mc$	$p-mc$

$a_1 = \text{invertir}$
 $a_2 = \text{no invertir}$



$Cost = 10$

$\pi = \text{Siempre } a_1$

$$U_n(r-c) = 10 - 10 + 0.9 [1 - U(p-c)] \rightarrow U_n(r-c) = -90 \quad (2)$$

$$U_n(p-c) = 0 - 10 + 0.9 [1 - U(p-c)] \rightarrow U_n(p-c) = -100 \quad (1)$$

$$U_n(r-mc) = 10 - 10 + 0.9 [0.5 U(p-c) + 0.5 U(p-mc)]$$

$$U_n(p-mc) = 0 - 10 + 0.9 [0.5 \underbrace{U(p-c)}_{-100} + 0.5 U(p-mc)]$$

(3)
 $-55 + 0.5 U(p-mc);$
 $U(p-mc) = -110$

(4)
 $U(r-mc) = -100$