Nota: la función de probabilidades de transición P se representará, en general, como un conjunto de tablas representando $P_a(\cdot \mid \cdot)$ 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:

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

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

   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:

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

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

   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:

$P_{a_1}(\cdot \mid \cdot) =$

	s_1	s_2	s_3	s_4
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

$P_{a_2}(\cdot \mid \cdot) =$

	s_1	s_2	s_3	s_4
s_2	0.0	0.5	0.0	0.5

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:

$P_{a_1}(\cdot \mid \cdot) =$

	s_1	s_2	s_3	s_4
s_1	0.0	1.0	0.0	0.0
s_2	1.0	0.0	0.0	0.0

$P_{a_2}(\cdot \mid \cdot) =$

	s_1	s_2	s_3	s_4
s_1	1/3	1/3	1/3	0.0
s_4	1/3	1/3	1/3	0.0

$P_{a_3}(\cdot \mid \cdot) =$

	s_1	s_2	s_3	s_4
s_2	0.75	0.0	0.0	0.25
s_3	0.5	0.5	0.0	0.0

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 - Boletín 5 - Formulario

- Probabilidad inducida

$$\prod P_{\pi(s_i)}(s_{i+1} \mid s_i)$$

- Utilidad inducida

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

- Sistema de ecuaciones (U_π)

s que estás usando y todas las demás

s que estás usando

$$\underset{\downarrow\, s_1, s_2, s_3 \ldots}{U(s)} = R(s, \underbrace{\pi(s)}_{\text{Los valores de la política } \pi \text{ para esa } s}) + \gamma \sum P_{\pi(s)}(s' \mid s)\, U(s')$$

- Ecuaciones de Bellman

$$U(s) = \max \left[R(s, \underbrace{a}_{a_1, a_2, a_3 \ldots}) + \gamma \sum P_a(s' \mid s)\, U(s') \right]$$

IA - Boletim 5

1)

- Probabilidad inducida $\longrightarrow \prod P_{\pi(s_i)}(s_{i+1} | s_i)$

$\langle \underset{0}{S_3}, \underset{1}{S_3}, \underset{2}{S_3}, \underset{3}{S_2}, \underset{4}{S_2} \rangle \qquad \pi(S_1)=a_3 \quad \pi(S_2)=a_3 \quad \pi(S_3)=a_2$

$P_{a_2}(\overset{1}{S_3}|\overset{0}{S_3}) \cdot P_{a_2}(\overset{2}{S_3}|\overset{1}{S_3}) \cdot P_{a_2}(\overset{3}{S_2}|\overset{2}{S_3}) \cdot P_3(\overset{4}{S_2}|\overset{3}{S_2})$

Utilidad inducidad $\longrightarrow \sum \gamma^i R(s_i)$

$R(S_3) + 0'9\, R(S_3) + 0'9^2 R(S_3) + 0'9^3 R(S_2) + 0'9^4 R(S_2)$

- Sist. ecuaciones $U_\pi \longrightarrow U(s) = R(s,\pi(s)) + \gamma \sum P_{\pi(s)}(s'|s)\, U(s')$

El que usas en esa U

← Todos los demas

$U(S_1) = R(S_1,a_3) + 0'9 [P_{a_3}(S_1|S_1) V(S_1) + P_{a_3}(S_2|S_1) U(S_2) + P_{a_3}(S_2|S_1) U(S_3)]$

$U(S_2) = R(S_2,a_3) + 0'9 [P_{a_3}(S_1|S_2) U(S_1) + P_{a_3}(S_2|S_2) U(S_2) + P_{a_3}(S_3|S_2) U(S_3)]$

$U(S_3) = R(S_3,a_2) + 0'9 [P_{a_2}(S_1|S_3) U(S_1) + P_{a_2}(S_2|S_3) U(S_2) + P_{a_2}(S_3|S_2) U(S_3)]$

recompensa

cost=0 ; no se escribe

- Ecuaciones de Bellmam $\longrightarrow U^* = \max(R(s,a) + \gamma \sum P_a(s'|s) U(s'))$

Para sustituir $P(\cdot|\cdot)$, mira la tabla

$U(S_1) = \max [$
$R(S_1,a_1) + 0'9 [\cancelto{0}{P_{a_1}(S_1|S_1)} U(S_1) + \cancelto{0}{P_{a_1}(S_2|S_1)} U(S_2) + \cancelto{0}{P_{a_1}(S_3|S_1)} U(S_3)]$
$R(S_1,a_2) + 0'9 [\cancel{P_{a_2}(S_1|S_1)} U(S_1) + P_{a_2}(S_2|S_1) U(S_2) + P_{a_2}(S_3|S_1) U(S_3)]$
$R(S_1,a_3) + 0'9 [\cancel{P_{a_3}(S_1|S_1)} U(S_1) + P_{a_3}(S_2|S_1) U(S_2) + P_{a_3}(S_3|S_1) U(S_3)]$
$]$

$U(S_1) = \max [R(S_1,a_1) + 0'9(0'3\, U(S_2) + 0'7\, U(S_3)),\ R(S_1,a_1) + 0'9(0'3\, U(S_2) + 0'7\, U(S_3))]$

Fíjate que como el primero queda sólo R(S1), no se usa aquí

- Política óptima

Será aquella política tal que $U_\pi = U^*$

2)

$\pi(s1) = a1 \qquad \pi(s3) = a2$

$\pi(s2) = a3 \qquad \pi(s4) = a1$

- $U_\pi(s)$

$$U(s_1) = R(s1, a1) + 0'5 [P_{a1}(\cancel{s_1}|s_1) U(s1) + P_{a1}(s2|\cancel{s_1}) U(s2) + P_{a1}(s3|s_1) U(s3) + P_{a1}(s4|s_1) U(s4)]$$

$$U(s1) = R(s_1, a_1) + 0'5 [0'2\, U(s3) + 0'8\, U(s4)]$$

$$U(s2) = R(s2, a3) + 0'5 [1\, U(s_1)]$$

$$U(s3) = R(s3, a2) + 0'5 [\tfrac{1}{3} U(s1) + \tfrac{1}{3} U(s2) + \tfrac{1}{3} U(s4)]$$

- Función de utilidad, una iteración de alg. it. valores

V_1

$$U_1(s_1) = \max \begin{cases} R(s_1, a_1) + \gamma [0'2 \cdot V_0(s3) + 0'8\, V_0(s4)] \\ R(s_1, a_2) + \gamma [1/3\, V_0(s_1) + \frac{1}{3} V_0(s2) + \frac{1}{3} V_0(s3)] \end{cases}$$

$$U_1(s_1) = \max \begin{cases} -3 - 2 + 0'5 [0'2 \cdot 1 + 0'8 \cdot 2] = -4'1 \\ \boxed{-3} - 3 + 0'5 [\frac{1}{3} \cdot 2 + \frac{1}{3}(-1) + \frac{1}{3} \cdot 1] = -6'3 \end{cases}$$

recompensa $\qquad$ coste

$$U_1(s2) = R(s2, a3) + \gamma [1 \cdot V_0(s1)] = -2 - 2 - 1 =$$

$$U_1(s3) = 1 - 3 + 0'5 [\tfrac{1}{3} V_0(s_1) + \tfrac{1}{3} V_0(s2) + \tfrac{1}{3} V_0(s4)] = -2'17$$

$$U_1(s4) = 1 - 2 + 0'5 [\tfrac{1}{3} V_0(s3) + \tfrac{1}{3} V_0(s4)] = 0'25$$

3)

- Cantidad de políticas

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

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

Sistema de ecuaciones

$$U(S_1) = R(S_1, a_1) + 0'9 \left[P_{a_1}\cancel{(S_1|S_1)}\, U(S_1) + P_{a_1}(S_2|S_1)\, U(S_2) + P_{a_1}(S_3|S_1)\, U(S_3) + P_{a_1}\cancel{(S_4|S_1)}\, U(S_4) \right]$$

$$U(S_1) = R(S_1, a_1) + 0'9 \left[0'5\, U(S_2) + 0'5\, U(S_3) \right]$$

$$U(S_2) = R(S_2, a_1) + 0'9 \left[P_{a_1}(S_1|S_2)\, U(S_1) + P_{a_1}\cancel{(S_2|S_2)}\, U(S_2) + P_{a_1}\cancel{(S_3|S_2)}\, U(S_3) + P_{a_1}\cancel{(S_4|S_2)}\, U(S_4) \right]$$

$$U(S_2) = R(S_2, a_1) + 0'9 \left[0'5\, U(S_1) \right]$$

$$U(S_3) = R(S_3, a_1) + 0'9 \left[P_{a_1}(S_1|S_3)\, U(S_1) + P_{a_1}(S_2|S_3)\, U(S_2) + P_{a_1}\cancel{(S_3|S_3)}\, U(S_3) + P_{a_1}\cancel{(S_4|S_3)}\, U(S_4) \right]$$

$$U(S_3) = R(S_3, a_1) + 0'9 \left[0'5\, U(S_1) + 1\, U(S_2) \right]$$

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

- Bellman

Como la política usa siempre a_1, Bellman sería $\max(U_\pi)$

4)

- Camtidad de políticas

$\|S\| \times \|A\| = 12$

-

$U(s_1) = \underset{R(s)}{R(s_1, a_1)} + 0'5 [P_{a1}(s_1|s_1) U(s_1) + P_{a1}(s_2|s_1) U(s_2) + P_{a1}(s_3|s_1) U(s_3) + P_{a1}(s_4|s_1) U_{s4}]$

$U(s_2) = R(s_2, a_1) + 0'5 [P_{a1}(s_1|s_2) U(s_1) + P_{a1}(s_2|s_2) U(s_2) + P_{a1}(s_3|s_2) U(s_3) + P_{a1}(s_4|s_2) U(s_3)]$

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

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

- Bellmam

$U(s_1) = \max [$

$R(s_1, a) + 0'5 [P_{a1}(s_1|s_1) U(s_1) + P_{a1}(s_2|s_1) U(s_2) + P_{a1}(s_3|s_1) U(s_3) + P_{a1}(s_4|s_1) U(s_4)]$

$R(s_1, a) + 0's [P_{a2}(s_1|s_2) U(s_2) + \dots$ y así sigue

5)

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

a_1 = invertir
a_2 = no invertir

$$\text{r-c} \xrightarrow{a_2} \begin{cases} 0'5 \rightarrow \text{r-c} \\ 0'5 \rightarrow \text{r-mc} \end{cases}$$

$$\text{r-mc} \xrightarrow{a_2} \begin{cases} 0'5 \rightarrow \text{r-mc} \\ 0'5 \rightarrow \text{p-mc} \end{cases}$$

$$\text{r-mc} \xrightarrow{a_1} \begin{cases} 0'5 \ \text{p-c} \\ 0'5 \ \text{p-mc} \end{cases}$$

$$\text{r-c} \xrightarrow{a_1} \text{p-c}$$

$$\text{p-c} \xrightarrow{a_2} \begin{cases} 0'5 \rightarrow \text{p-mc} \\ 0'5 \rightarrow \text{r-c} \end{cases}$$

$$\text{p-c} \xrightarrow{a_1} \text{p-c}$$

$$\text{p-mc} \xrightarrow{a_2} \text{p-mc}$$

$$\text{p-mc} \xrightarrow{a_1} \begin{cases} 0'5 \rightarrow \text{p-c} \\ 0'5 \rightarrow \text{p-mc} \end{cases}$$

Coste = 10

Π = siempre a_1

$U_n(r\text{-}c) = 10-10 + 0'9\,[1 \cdot U(p\text{-}c)] \longrightarrow U_n(r\text{-}c) = -90$ (2)

$U_n(p\text{-}c) = 0-10 + 0'9\,[1 \cdot U(p\text{-}c)] \longrightarrow U_n(p\text{-}c) = -100$ (1)

$U_n(r\text{-}mc) = 10-10 + 0'9\,[0'5\,U(p\text{-}c) + 0'5\,U(p\text{-}mc)]$

$U_n(p\text{-}mc) = 0-10 + 0'9\,[0'5\,\underbrace{U(p\text{-}c)}_{-100} + 0'5\,U(p\text{-}mc)]$

(3) $-55 + 0'5\,U(p\text{-}mc)$; $U(p\text{-}mc) = -110$

$U(r\text{-}mc) = -100$ (4)