UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

Ejemplos de problemas tipo examen de Inteligencia Artificial: Especificaciones de Procesos de decisión de Markov (PDM).

Problema 1. Juego de cartas solitario (suma 5).

Se pretende jugar a un juego de cartas donde se utiliza un mazo que cuenta con tres tipos cartas diferentes (el uno, el dos y el tres). El juego consiste en llegar a sumar cinco tirando sobre el mantel la cantidad de cartas que se consideren oportunas. Se puede asumir que el mazo de cartas es infinito y que la proporción de cartas que hay en el mazo es tal que hay el doble de doses que de unos y que de treses (la cantidad de treses es la misma que la de unos). De esta forma, la proporción de cartas que tenemos es la siguiente:

El/la jugador/a empieza el juego tirando una carta y deberá decidir en cada momento si continúa tirando una carta más o si se planta (con lo que el juego termina), teniendo en cuenta que:

- Si se planta sin haber llegado a sumar 5, los puntos que ganará serán la suma del valor de las cartas que haya sobre el mantel. Así, en la siguiente secuencia de cartas de ejemplo obtendría, al acabar el juego, un total de 4 puntos.

- Si el/la jugador/a llega a tirar un total de cartas cuya suma supere los 5 puntos, perderá y el juego se dará por terminado. De este modo, en la siguiente secuencia de ejemplo, el/la jugador/a perderá.

- Y por último, si la tirada de cartas sobre el mantel suma exactamente 5, entonces el/la jugador/a ganará un total de 10 puntos.

1. Formaliza el problema como un proceso de decisión de Markov, identificando de forma clara el conjunto de estados, las acciones disponibles en cada uno de los estados, la función de transición y la función de recompensa.
2. Ejecuta hasta 4 iteraciones del algoritmo de iteración de valores y muestra los Q-valores y los valores de cada estado. Teniendo en cuenta que en los cálculos asumiremos un factor de descuento $\gamma = 1$ y que no se pide que se utilicen más de dos decimales ¿Cuál es la política óptima de juego? (si no te da tiempo puedes ejecutar 3 iteraciones).

UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

3. Considerando: que inicialmente $Q_0(s,a)=0\ \forall s,a$; un factor de aprendizaje $\alpha=1$; un factor de descuento $\gamma = 1$; una función de selección de acciones excesivamente simple que siempre retorna tirar para los tres primeros estados, que responde alternativamente tirar y plantarse en el cuarto estado y plantarse para los siguientes estados en los que no ha perdido; y que en el mazo tiene la siguiente secuencia de cartas: 2, 3, 1, 2, 1, 3, 2, 2, 3, 2, 1, 2. Aplica el algoritmo de Q-learning (Q-aprendizaje) durante 15 iteraciones especificando los Q-values (Q-valores) de los Q-states (Q-estados) que se calculan. Al finalizar, especifica tanto los Valores de los estados como la política que se ha aprendido hasta ese momento.

Pistas pregunta 2:

Fórmula: $Q_{i+1}(s,a)= \sum_{s'} T(s,a,s')[R(s,a,s')+\gamma V_i(s')]$ $\quad V_{i+1}(s)= \max_a Q_{i+1}(s,a)$ tomamos $\gamma=1$

Cálculos: hacer 4 iteraciones. Para cada iteración hace falta calcular primero cada $Q_{i+1}(s,a)$ para luego tener el $V_{i+1}(s)$. El enunciado pide tanto los Q-valores como los V. Como guía se muestra el V y la política aprendida en cada iteración.

Opción 1 (considerando $\{s_i, \mid s_i=\sum c_i, i=1..n, 0<=s_i<=5\} \cup \{perder, ganar\}$**):**

Algunos ejemplos de cálculo:

$Q_1(0, T) =\Sigma_{s'} T(0,T,s')[R(0,T,s')+\gamma V_0(s')] = 0.25\ (0+1*0) + 0.5\ (0+1*0) +0.25(0+1*0)= 0$ $\quad (s'=1,2,3)$

$Q_1(0, P) =\Sigma_{s'} T(0,P,s')[R(0,P,s')+\gamma V_0(s')] = 1(0+1*0) = 0$ $\quad (s'=ganar)$

$V_1(0)= \max_{a=\{T, P\}} Q_1(0,a) = \max (0,0)= 0$, $\arg\max_a = T, P$

$Q_2(3, T) =\Sigma_{s'} T(3,T,s')[R(3,T,s')+\gamma V_1(s')] = 0.25\ (0+1*4) + 0.5\ (0+1*10) +0.25(0+1*0)= 1+5=6$ $\quad (s'= 4, 5, perder)$

$Q_2(3, P) =\Sigma_{s'} T(3,P,s')[R(3,P,s')+\gamma V_1(s')] = 1(3+1*0) = 3$ $\quad (s'=ganar)$

$V_2(3)= \max_{a=\{T, P\}} Q_3(3,a) = \max (6,3)= 6$, $\arg\max_a = T$

$Q_4 (5, T) =\Sigma_{s'} T(5,T,s')[R(5,T,s')+\gamma V_3 (s')] = 1\ (0+1*0) = 0$ $\quad (s'=perder)$

$Q_4 (5, P) =\Sigma_{s'} T(5,P,s')[R(5,P,s')+\gamma V_3 (s')] = 1(10+1*0) = 10$ $\quad (s'=ganar)$

$V_4(5)= \max_{a=\{T, P\}} Q_4(5,a) = \max (0,10)= 10$, $\arg\max_a = P$

s	$V_0(s)$	$V_1(s)$	$\arg\max_a$	$V_2(s)$	$\arg\max_a$	$V_3(s)$	$\arg\max_a$	$V_4(s)$	$\arg\max_a$
0	0	0	T|P	2	**T**	4.875	**T**	5.828125	T
1	0	1	P	3	**T**	5.3125	**T**	5.5	T
2	0	2	P	5.25	**T**	6	**T**	6	T
3	0	3	P	6	**T**	6	**T**	6	T
4	0	4	P	4	**P**	4	**P**	4	P
5	0	10	P	10	**P**	10	**P**	10	P
perder	0	0		0		0		0	
ganar	0	0		0		0		0	

T|P significa que pueden ser ambas, ya que tienen el mismo Q-valor

Opción 2 (considerando $\{s_i, \mid s_i=\sum c_i, i=1..n, 0<=s_i<=5\} \cup \{perder\}$**):**

UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

s	$V_0(s)$	$V_1(s)$	arg max_a	$V_2(s)$	arg max_a	$V_3(s)$	arg max_a	$V_4(s)$	arg max_a
0	0	0	T|P	2	T	4.875	**T**	9.828125	**T**
1	0	1	P	3	T	6.3125	**T**	11.625	**T**
2	0	2	P	5.25	T	10.5	**T**	16.5	**T**
3	0	3	P	6	T|P	12	**T**	18	**T**
4	0	4	P	8	P	12	**P**	16	**P**
5	0	10	P	20	P	30	**P**	40	**P**
perder	0	0		0		0		0	

Opción 3: se podría considerar también otros estados tales como {0, 1, 2, 3, 4, 5, 0F, 1F, 2F, 3F, 4F, 5F, 6F 7F 8F}. Los x serían los estados no terminales (se siguen haciendo acciones) mientras que los xF serían los estados terminales: o bien los terminales a los que se llega después de plantarse; o bien los 6F, 7F, 8F serían los estados terminales en los que se ha perdido. En el caso de los estados terminales, todos sus valores se mantienen a 0 a lo largo de todas las iteraciones y no se calculan sus Q-valores, ya que no se pueden hacer acciones en ellos.

Pistas pregunta 3:

Fórmula: $Q_{i+1}(s,a) = Q_i(s,a) + \alpha\,[R(s,a,s') + \gamma \max_{a'} Q_i(s',a') - Q_i(s,a)]$ con $\alpha=\gamma=1$; $V_{i+1}(s) = \max_a Q_{i+1}(s,a)$

Cálculos: hacer 15 iteraciones. Para cada iteración se calcula un único $Q_{i+1}(s,a)$, el que corresponde al estado y acción de la experiencia, para luego tener el $V_{i+1}(s)$. Como guía se muestra el V y la política aprendida.

Opción 1 (considerando $\{s_i, \mid s_i=\sum c_i,\ i=1..n,\ 0<=s_i<=5\} \cup \{perder, ganar\}$**):**
Algunos ejemplos de cálculo:
Primera experiencia:
Iteración 1: s=0, a=T, s'=2, R(0,T,2)=0, (mazo cartas: 2, 3, 1, 2, 1, 3, 2, 2, 3, 2, 1, 2)

$Q_1(0,T) = Q_0(0,T) + \alpha\,[R(0,T,2) + \gamma \max_{a'} Q_0(2,a') - Q_0(0,T)] = 0+1*[0+1*\max_{a'=\{T,P\}}(0,0)-0]=0$

Iteración 3: s=5, a=P, s'=ganar, R(5,P, ganar)=10 (mazo cartas: 1, 2, 1, 3, 2, 2, 3, 2, 1, 2)

$Q_3(5,P) = Q_2(5,P) + \alpha\,[R(5,P,ganar) + \gamma \max_{a'} Q_2(ganar,a') - Q_2(5,P)] = 0+1*[10+1*\max_{a'=\{\}}()-0]=10$

(ganar es un estado terminal, no le puedo calcular Q values)

(Nota: la experiencia termina una vez que se realiza la acción P, por tanto en la siguiente iteración comienza un juego nuevo)

Segunda experiencia:
Iteración 4: s=0, a=T, s'=1, R(0,T,1)=0 (mazo cartas: 1, 2, 1, 3, 2, 2, 3, 2, 1, 2)

$Q_4(0,T) = Q_3(0,T) + \alpha\,[R(0,T,1) + \gamma \max_{a'} Q_3(1,a') - Q_3(0,T)] = 0+1*[0+1*\max_{a'=\{T,P\}}(0,0)-0]=0$

Iteración 6: s=3, a=T, s'=4, R(3,T,4)=0 (mazo cartas: 1, 3, 2, 2, 3, 2, 1, 2)

$Q_6(3,T) = Q_5(3,T) + \alpha\,[R(3,T,4) + \gamma \max_{a'} Q_5(4,a') - Q_5(3,T)] = 0+1*[0+1*\max_{a'=\{T,P\}}(0,0)-0]=0$

Tercera experiencia:

Iteración 9: s=3, a=P, s'=ganar, R(3,P,ganar)=3 (mazo cartas: 2, 2, 3, 2, 1, 2)

$Q_9(3,P) = Q_8(3,P) + \alpha\,[R(3,P,ganar) + \gamma \max_{a'} Q_8(ganar,a') - Q_8(3,P)] = 0+1*[3+1*\max_{a'=\{\}}()-0]=3$

Quinta experiencia:

Iteración: 14: s=3, a=T, s'=5, R(3,T,5)=0 (mazo cartas: 2, 1, 2)

$Q_{14}(3,T) = Q_{13}(3,T) + \alpha\,[R(3,T,5) + \gamma \max_{a'} Q_{13}(5,a') - Q_{13}(3,T)] = 0+1*[0+1*\max_{a'=\{T,P\}}(0,10)-$

UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

0]=0+[0+10-0]=10

V(0)=max(Q(0,T),Q(0,P))=max(3,0)=3 y π(0)= T

V(3)= $\max_{a=\{T,P\}}$ Q(3,a)= max (10,3)=10, π(3)=T

	Ini.	#It	Q(s,a)	V(s)	π(s)
Q(0,T)	$Q_0(0,T)=0$	1	$Q_1(0,T)=0$	V(0)=3	π(0)=T
		4	$Q_4(0,T)=0$		
		8	$Q_8(0,T)=0$		
		10	$Q_{10}(0,T)=0$		
		13	$Q_{13}(0,T)=3$		
Q(0,P)	$Q_0(0,P)=0$				
Q(1,T)	0	5	$Q_5(1,T)=0$	V(1)=0	π(1)=T|P
Q(1,P)	0				
Q(2,T)	0	2	$Q_2(2,T)=0$	V(2)=4	π(2)=T
		11	$Q_{11}(2,T)=4$		
Q(2,P)	0				
Q(3,T)	0	6	$Q_6(3,T)=0$	V(3)=10	π(3)=T
		14	$Q_{14}(3,T)=10$		
Q(3,P)	0	9	$Q_9(3,P)=3$		
Q(4,T)	0			V(4)=4	π(4)=P
Q(4,P)	0	7	$Q_7(4,P)=4$		
		12	$Q_{12}(4,P)=4$		
Q(5,T)	0			V(5)=10	π(5)=P
Q(5,P)	0	3	$Q_3(5,P)=10$		
		15	$Q_{15}(5,P)=10$		

UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

Opción 2 (considerando $\{s_i, \mid s_i=\sum c_i, i=1..n, 0<=s_i<=5\} \cup \{perder\}$**):**

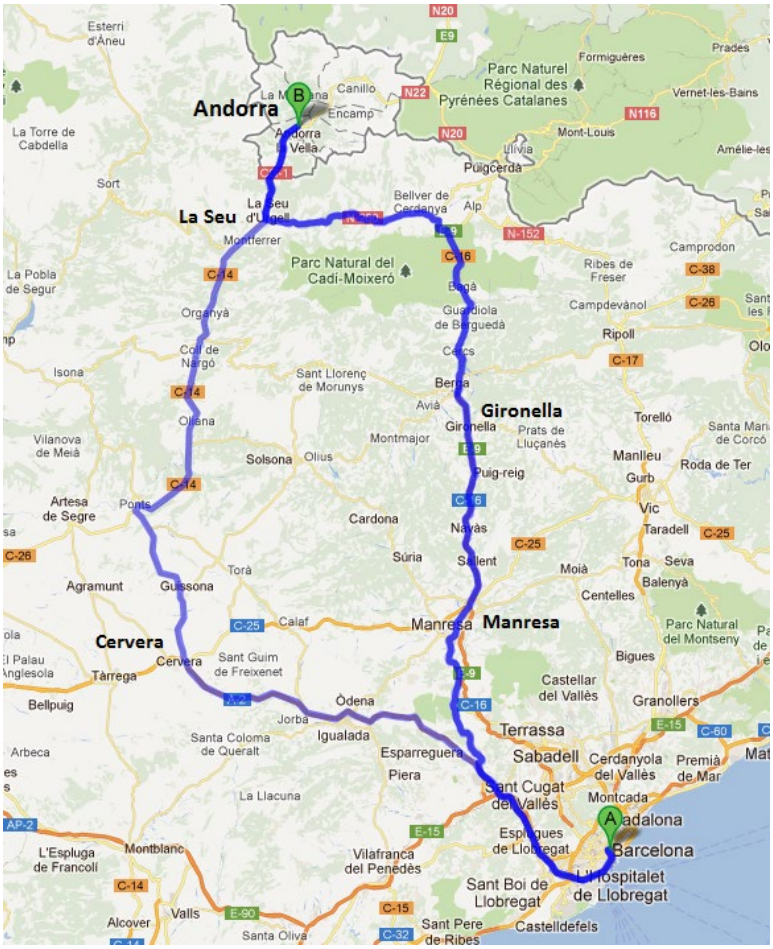
	Ini.	#It	Q(s,a)	V(s)	π(s)
Q(0,T)	$Q_0(0,T)=0$	1	$Q_1(0,T)=0$	V(0)=3	π(0)=T
		4	$Q_4(0,T)=0$		
		8	$Q_8(0,T)=0$		
		10	$Q_{10}(0,T)=0$		
		13	$Q_{13}(0,T)=3$		
Q(0,P)	$Q_0(0,P)=0$				
Q(1,T)	0	5	$Q_5(1,T)=0$	V(1)=0	π(1)=T|P
Q(1,P)	0				
Q(2,T)	0	2	$Q_2(2,T)=0$	V(2)=4	π(2)=T
		11	$Q_{11}(2,T)=4$		
Q(2,P)	0				
Q(3,T)	0	6	$Q_6(3,T)=0$	V(3)=10	π(3)=T
		14	$Q_{14}(3,T)=10$		
Q(3,P)	0	9	$Q_9(3,P)=3$		
Q(4,T)	0			V(4)=8	π(4)=P
Q(4,P)	0	7	$Q_7(4,P)=4$		
		12	$Q_{12}(4,P)=8$		
Q(5,T)	0			V(5)=20	π(5)=P
Q(5,P)	0	3	$Q_3(5,P)=10$		
		15	$Q_{15}(5,P)=20$		

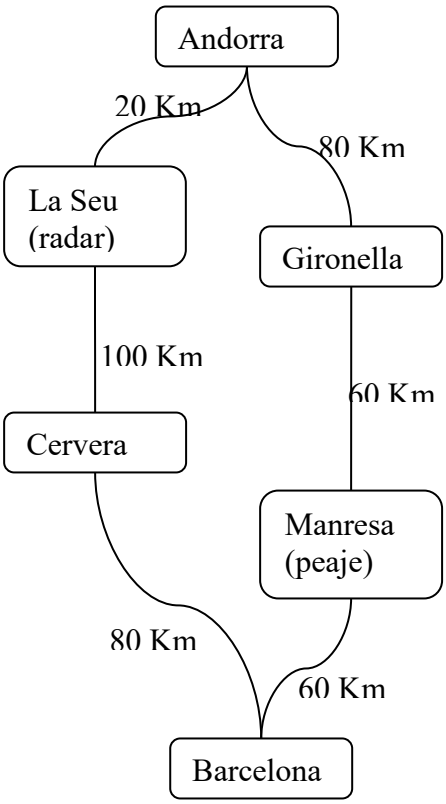
UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

Problema 2. Aprendizaje de la mejor ruta a Andorra.

Nos acabamos de dar cuenta que nos dan 200 euros por un número de lotería que nos regalaron una de las últimas veces que fuimos a Andorra y queremos, analizando la experiencia acumulada de los dos últimos viajes, aprender la mejor ruta en términos económicos para ir de Barcelona a Andorra. En particular recordamos que una ocasión realizamos una ruta a través de la autopista pasando por Manresa y Gironella. En esta primera ruta el importe del peaje ascendió a 20 euros. Por el contrario, la segunda ruta que cogimos no incluyó la autopista, ya que pasamos por Cervera y La Seu d'Urgell. No obstante, en La Seu, un radar detectó que circulamos a mayor velocidad de lo permitido, y tuvimos que pagar una multa de 150 euros. Además, al seguir ambas rutas vimos que la gasolina gastada supone un gasto de 1 euro cada 10 Kilómetros.





1. Formaliza el problema identificando de forma clara el conjunto de estados, las acciones disponibles en cada uno de los estados, la función de transición y la función de recompensa.
2. Escribe el código (preferiblemente en pseudo-código) de Q-Learning.
3. Ejecuta el algoritmo de Q-Learning descrito en el punto 2 asumiendo un factor de descuento $\gamma = 1$ y un factor de aprendizaje fijo a $\alpha=0.5$. Realiza 10 iteraciones incorporando la experiencia de seguir primero las dos rutas completas (desde Barcelona hasta Andorra) y luego repitiendo la primera ruta tantas veces como sea necesario para llegar a calcular las 10 iteraciones. Muestra los Q-values (Q-valores) de cada estado ¿Cuál es la política aprendida hasta entonces?

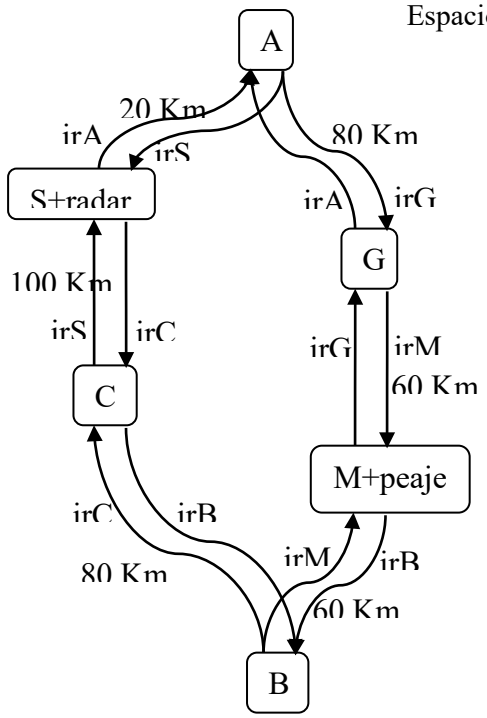
UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

Soluciones: Dos soluciones posibles: una con acciones que permiten ir hacia atrás y otra con acciones que sólo permiten avanzan en la ruta. Ambas son válidas.

Pistas solución 1: considerando acciones que permiten ir hacia atrás en la ruta

Espacio de estados



$V(s)= \max_a Q(s,a)$, $\pi(s) = \arg\max_a Q(s,a)$, inicialmente $Q_0(s,a)=0\ \forall s,a$

	Ini.	#It	Q(s,a)	V(s)	π(s)
Q(B,irM)	Q_0(B,irM)=0	1	Q_1(B,irM)=-13	V(B)=-1	π(B)=irM
		7	Q_7(B,irM)=-19.5		
		10	Q_{10}(B,irM) =-1		
Q(B,irC)	Q_0(B,irC)=0	4	Q_4(B,irC) =-4		
Q(M,irG)	0	2	Q_2(M,irG) = -3	43.5	π(M)=irG
		8	Q_8(M,irG) =43.5		
Q(M,irB)	0				
Q(G,irA)	0	3	Q_3(G,irA) =96	144	π(G)=irA
		9	Q_9(G,irA) =144		
Q(G,irM)	0				
Q(A,irG)	0				
Q(A,irS)	0				
Q(C,irS)	0	5	Q_5(C,irS) =-80	V(C)=0	π(C)=irB
Q(C,irB)	0				
Q(S,irA)	0	6	Q_6(S,irA)=99	V(S)=99	π(S)=irA
Q(S,irC)	0				

Primera experiencia (iteraciones 1, 2, 3, 7, 8, 9, 10), segunda experiencia (iteraciones 4,5,6)

UNIVERSITAT DE BARCELONA

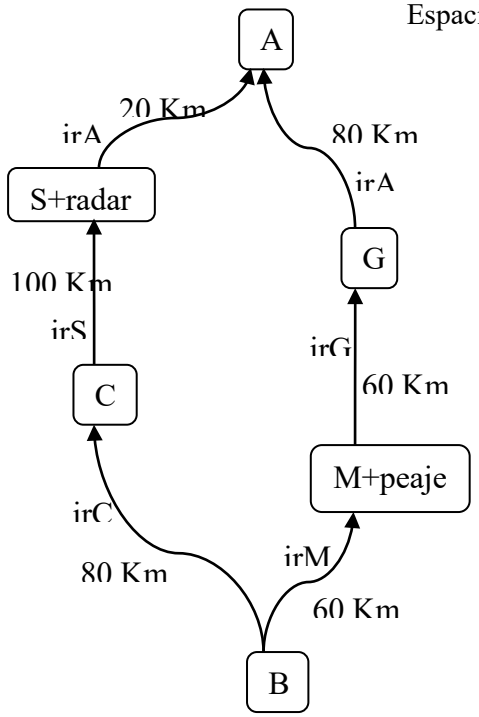
Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

Nota: no calculo V(A) porque Q(A,irG) y Q(A,irS) no se van a evaluar nunca, ya que en A se acaba la experiencia, la dejo en las tablas por recordar que sus Q iniciales son 0, pero se podría quitar, como he hecho en el ejercicio anterior con el perder.

Teniendo en cuenta la política aprendida hasta ahora, de Barcelona se iría a Manresa, de Manresa a Gironella y de Gironella se elige ir a Andorra. (saldría esa ruta, la de Cervera+LaSeu aún se pensaría que en Cervera es mejor ir a Barcelona, aunque en la Seu ya saldría ir a Andorra)

Pistas Solución 2: considerando acciones que sólo permiten avanzan en la ruta

Espacio de estados



	Ini.	#It	Q(s,a)	V(s)	π(s)
Q(B,irM)	Q_0(B,irM)=0	1	Q_1(B,irM)=-13	V(B)=-1.75	π(B)=irM
		7	Q_7(B,irM)=-21		
		10	Q_{10}(B,irM) =-1.75		
Q(B,irC)	Q_0(B,irC)=0	4	Q_4(B,irC) =-4		
Q(M,irG)	0	2	Q_2(M,irG) = -3	43.5	π(M)=irG
		8	Q_8(M,irG) =43.5		
Q(G,irA)	0	3	Q_3(G,irA) =96	144	π(G)=irA
		9	Q_9(G,irA) =144		
Q(C,irS)	0	5	Q_5(C,irS) =-80	V(C)=-80	π(C)=irS
Q(S,irA)	0	6	Q_6(S,irA)=99	V(S)=99	π(S)=irA

UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

Problema 3: Concierto.

Vivimos a las afueras de Barcelona y nos regalan la entrada de un concierto en Madrid. Estamos pensando en el tipo de transporte que vamos a usar para llegar a Madrid. Luego allí en Madrid, tenemos unos colegas que nos vienen a buscar donde lleguemos y nos llevan al concierto directamente, así que en cuanto pisamos suelo madrileño ya no nos preocupamos por nada, sólo disfrutamos del concierto que valoramos con un 10. Analizando dos experiencias anteriores recordamos que, en la primera, fuimos en AVE (que por su precio y tiempo de trayecto valoramos que nos reporta un -3). El AVE sale de la estación de Sants, a la que llegamos en taxi (porque donde vivimos no hay transporte público y en Barcelona no podemos aparcar fácilmente). Esa parte del trayecto resultó cómoda pero cara, así que la valoramos como -4. En cuanto a la segunda experiencia, fuimos en avión. Vivimos cerca del aeropuerto del Prat, así que un vecino nos puede llevar siempre que le arreglamos algo del ordenador (eso lo valoramos con un -2). En el avión, aunque el trayecto fue corto y cómodo, nos perdieron la maleta en la que llevábamos un regalo para nuestros amigos madrileños, por lo que lo valoramos como -6. Consideramos cada uno de estos pasos como unidireccionales (es decir, podemos, por ejemplo, ir en taxi de casa a la estación pero no consideramos que podamos utilizar el taxi para volver a casa desde la estación).

1. Formaliza el problema identificando de forma clara el conjunto de estados, las acciones disponibles en cada uno de los estados y describe usando esta tabla las dos experiencias (incluye la recompensa que valoraremos que tenemos en cada paso).

	Experiencia 1			Experiencia 2		
	Estado (s)	Acción (a)	Recompensa R(s,a,s')	Estado (s)	Acción (a)	Recompensa R(s,a,s')
paso 1						
paso 2						
paso 3						

2. Ejecuta el algoritmo de Q-Learning considerando un factor de descuento $\gamma = 1$ y un factor de aprendizaje fijo a $\alpha=1$. Realiza las iteraciones que sean necesarias para incorporar las dos experiencias (en el orden en el que están descritas). Muestra los Q-valores de cada Q-estado y los Valores de cada estado. Especifica además qué acción marca la política aprendida hasta el momento sobre qué se debe hacer al inicio, cuando estamos en casa.

Pistas solución: (según si se considera como estado terminal estar en Madrid o si se define un estado terminal separado V(M) cambia su valor, pero no la política)

Q-state	Ini.	#It	Q(s,a)	V(s)	π(s)
Q(C,t)	Q_0(C,t)=0	1	Q_1(C,t) = -4	V(C)=-2	π(C)=c
Q(C,c)	Q_0(C,t)=0	4	Q_4(C,c) =-2		
Q(S,a)	0	2	Q_2(S,a) = -3	V(S)=-3	π(S)=a
Q(M,e)	0	3 6	Q_3(M,e) = 10 Q_6(M,e) = 20|10	V(M)=20|10	π(M)=e
Q(P,v)	0	5	Q_5(P,v) = 4	V(P)=4	π(P)=v

Por tanto, en casa (C) deberemos ir en coche (c) al aeropueto.

UNIVERSITAT DE BARCELONA

Departament de Matemàtiques i Informàtica
Gran Via de les Corts Catalanes, 585
08007 Barcelona

Problema 4. Piercings peligrosos.

Luis Ángel, un joven de 14 años se enfrenta al siguiente problema: está enamorado, pero ella (Maria Fernanda) le ignora. El cree saber la razón. María Fernanda únicamente se deja impresionar por chicos que lleven piercings. Por lo tanto Luis Ángel se plantea hacerse unos cuantos piercings. El problema es que Luis Ángel vive en una aldea retirada y no tiene dinero para pagarse un viaje a la ciudad para que le haga los piercings un experto, así que su única opción es pedírselo a su tía Robustiana, de 68 años y que de joven trabajó cosiendo sacos para un almacén de grano. Luis Ángel es un chico muy calculador, así que evalúa con exactitud los riesgos y beneficios inherentes a hacerse o no piercings.

Él sabe que si le pide salir a María Fernanda sin hacerse ningún piercing, la probabilidad de que le diga que sí es de un 10%. En cambio con un piercing es de un 30%, con dos piercings del 80% y con tres piercings del 90%. Luis Ángel cree que Robustiana no vivirá suficiente como para hacerle más de 3 piercings. Luis Ángel sabe que una vez le pida salir a Maria Fernanda, ella no cambiará nunca más de opinión. Por eso, Luis Ángel le da un valor de +500 a conseguir que le diga que si (nunca más volverá a pensar en piercings a partir de ese momento) y un valor de -100 si le dice que no (piensa que se suicidará sin su amor).

Cada vez que Luis Ángel decide hacerse un piercing, tiene un 80% de posibilidades de que salga bien, un 10% de posibilidades de pasar 3 semanas en cama y despertarse sin piercing (a lo que asigna un valor de -20, dado que lo pasa mal y le retrasa en su objetivo de hacerse con el amor de María Fernanda) y un 10% de posibilidades de morir (a lo que asigna un valor de -90, dado que morirá sin saber si ella le hubiera aceptado).

Se solicita:

1. Ayuda a Luis Ángel formalizando el problema como un proceso de decisión de Markov. Identifica con total claridad el conjunto de estados, las acciones disponibles en cada uno de los estados, la función de transición y la función de recompensa.

2. Ejecuta 5 iteraciones del algoritmo de iteración de valores y muestra los Q-valores y los valores de cada estado considerando una γ=1. ¿Cuál es la política óptima para Luis Ángel?

Pistas Solución:
Hacer 5 iteraciones en la línea de lo realizado para el primer ejercicio. Para cada iteración hace falta calcular primero cada $Q_{i+1}(s,a)$ para luego tener el $V_{i+1}(s)$. El enunciado pide tanto los Q-valores como los V. Como guía se muestra el V y la política aprendida en cada iteración.

s	$V_0(s)$	V_1 (s)	π(s)	V_2 (s)	π(s)	V_3(s)	π(s)	V_4 (s)	π(s)	V_5 (s)	π(s)
0	0	-11	P	51.9	P	234.99	P	270.979	P	276,3459	P
1	0	80	S	301	**P**	323.1	P	325.31	P	325,531	P
2	0	380	S	380	S	380	S	380	S	380	S
3	0	440	S	440	S	440	S	440	S	440	S
Si	0	0		0		0		0		0	
No	0	0		0		0		0		0	
mort	0	0		0		0		0		0	

Nota: si se consideran Si, No, mort como estados terminales, todos sus valores se mantienen a 0 a lo largo de toda las iteraciones y no se calculan sus Q-valores