Procesos de Decisión de Markov

Aprendizaje Automático

Ingeniería Informática

Fernando Fernández Rebollo y Daniel Borrajo Millán

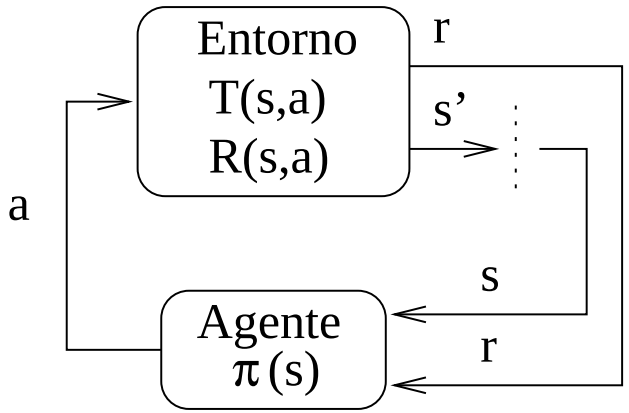
Grupo de Planificación y Aprendizaje (PLG)
Departamento de Informática
Escuela Politécnica Superior
Universidad Carlos III de Madrid

27 de febrero de 2009

En Esta Sección:

Aprendizaje por Refuerzo

- El aprendizaje por refuerzo consiste en aprender a decidir, ante una situación determinada, qué acción es la más adecuada para lograr un objetivo.
- Elementos principales:
  - Proceso iterativo de prueba y error
  - Aprendizaje a través de señales de refuerzo



Métodos de Resolución

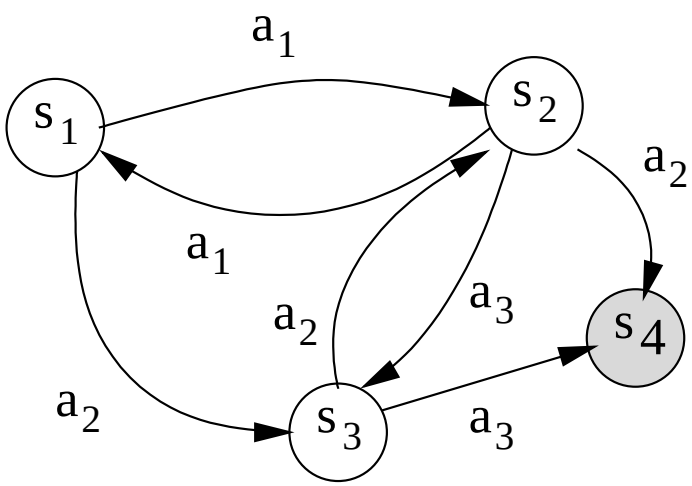
- Si se conoce el modelo (funciones de transición de estado y refuerzo): Programación Dinámica
- Si no se conoce el modelo:
  - Aprender el modelo: métodos basados en el modelo o *Programación Dinámica*
  - Aprender las funciones de valor directamente: métodos libres de modelo o Aprendizaje por Refuerzo

Definición de un MDP

- Un MDP se define como una tupla $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$, tal que:
  - $\mathcal{S}$ es un conjunto de estados
  - $\mathcal{A}$ un conjunto de acciones
  - $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S})$, donde un miembro de $P(\mathcal{S})$ es una distribución de probabilidad sobre el conjunto $\mathcal{S}$; es decir, transforma estados en probabilidades. Se dice que $T(s, a, s')$ es la probabilidad de que se realice una transición desde s hasta s' ejecutando la acción a
  - $\mathcal{R} : \mathcal{S}\mathrm{x}\mathcal{A} \rightarrow \Re$, que para cada par estado-acción proporciona su refuerzo. Se dice que $\mathcal{R}(s, a)$ es el refuerzo recibido tras ejecutar la acción a desde el estado s.

Ejemplo de MDP Determinista

La ejecución de una acción desde un estado siempre produce la misma transición de estado y el mismo refuerzo/coste



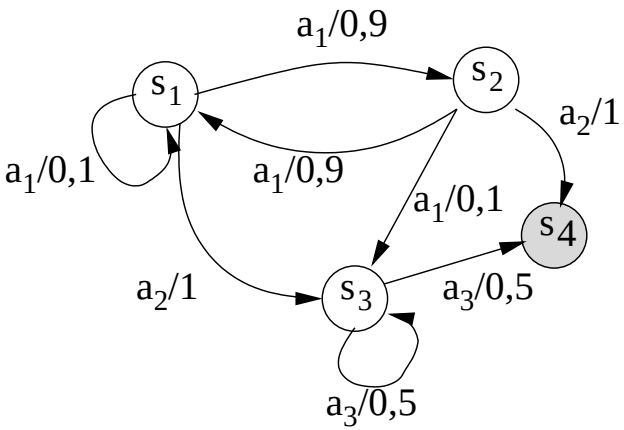
	s_i			
	s_1	s_2	s_3	s_4
$\mathcal{T}(s_1, a1, s_i)$	0	1	0	0
$\mathcal{T}(s_1, a2, s_i)$	0	0	1	0
$\mathcal{T}(s_1, a3, s_i)$	1	0	0	0
$\mathcal{T}(s_2, a1, s_i)$	1	0	0	0
$\mathcal{T}(s_2, a2, s_i)$	0	0	0	1
$\mathcal{T}(s_2, a3, s_i)$	0	1	0	0
$\mathcal{T}(s_3, a1, s_i)$	0	0	1	0
$\mathcal{T}(s_3, a2, s_i)$	0	1	0	0
$\mathcal{T}(s_3, a3, s_i)$	0	0	0	1
$\mathcal{T}(s_4, a1, s_i)$	0	0	0	1
$\mathcal{T}(s_4, a2, s_i)$	0	0	0	1
$\mathcal{T}(s_4, a3, s_i)$	0	0	0	1

$R(s_i, a, s_4) = 1$ si $s_i = \{1, 2, 3\}$.

$R(s_i, a, s_j) = 0$ e.c.o.c.

Ejemplo de MDP Estocástico

Las transiciones de estado y la función de refuerzo son funciones estocásticas, por lo que la misma situación puede producir distintos resultados



	s_i			
	s_1	s_2	s_3	s_4
$\mathcal{T}(s_1, a1, s_i)$	0,1	0,9	0	0
$\mathcal{T}(s_1, a2, s_i)$	0	0	1	0
$\mathcal{T}(s_1, a3, s_i)$	1	0	0	0
$\mathcal{T}(s_2, a1, s_i)$	0,9	0	0,1	0
$\mathcal{T}(s_2, a2, s_i)$	0	0	0	1
$\mathcal{T}(s_2, a3, s_i)$	0	1	0	0
$\mathcal{T}(s_3, a1, s_i)$	0	0	1	0
$\mathcal{T}(s_3, a2, s_i)$	0	0	1	0
$\mathcal{T}(s_3, a3, s_i)$	0	0	0,5	0,5
$\mathcal{T}(s_4, a1, s_i)$	0	0	0	1
$\mathcal{T}(s_4, a2, s_i)$	0	0	0	1
$\mathcal{T}(s_4, a3, s_i)$	0	0	0	1

$R(s_i, a, s_4) = 1$ si $s_i = \{1, 2, 3\}$.

$R(s_i, a, s_j) = 0$ e.c.o.c.

Propiedad de Markov

- Propiedad de Markov:
  El estado anterior y la última acción realizada son suficientes para describir el estado actual y el refuerzo recibido

$$Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t\} =$$
$$Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0, \}$$

- Consecuencia: la acción a ejecutar sólo depende del estado actual

Políticas y Optimalidad

- Objetivo de planificación:
  - Encontrar una política, π, que para cada estado $s \in S$, decida cuál es la acción, $a \in A$, que debe ser ejecutada, de forma que se maximice alguna medida de refuerzo a largo plazo.
- Criterio de optimalidad de horizonte infinito descontado:

$$\sum_{k=0}^{\infty} \gamma^k r_k \tag{44}$$

donde $0 \leq \gamma \leq 1$

Funciones de Valor y Políticas

- El cálculo de las políticas óptimas, $\pi^*(s)$, se basa en las funciones de valor:
  - Función de valor-estado (dada una política π):

$$V^\pi(s) = E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}$$

  - Función de valor-acción (dada una política π):

$$Q^\pi(s,a) = E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\}$$

  - Relación entre las funciones de valor:

$$V^\pi(s) = \max_{a \in \mathcal{A}} Q^\pi(s,a)$$

Funciones de Valor Óptimas

- Función de valor-estado óptima:

$$V^*(s) = \max_\pi V^\pi(s) \forall s \in \mathcal{S}$$

- Función de valor-acción óptima:

$$Q^*(s,a) = \max_\pi Q^\pi(s,a) \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

- Relación entre las funciones de valor óptimas:

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a), \forall s \in \mathcal{A}$$

Funciones de Valor Óptimas

- Definición de una política óptima, $\pi^*(s)$, en función de $Q^*(s,a)$:

$$\pi^*(s) = \arg_{a \in \mathcal{A}} \max Q^*(s,a)$$

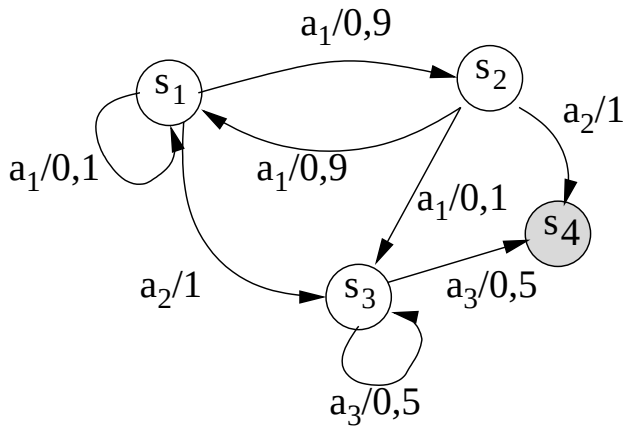
- Ecuaciones de optimalidad de Bellman:

$$Q^*(s,a) = \sum_{s'} \mathcal{T}(s,a,s')[R(s,a) + \gamma V^*(s'))]$$

$$Q^*(s,a) = \sum_{s'} \mathcal{T}(s,a,s')[R(s,a) + \gamma \max_{a'} Q^*(s',a'))]$$

$$V^*(s) = \max_{a \in \mathcal{A}} \sum_{s'} \mathcal{T}(s,a,s')[\mathcal{R}(s,a) + \gamma V^*(s')]$$

Ejemplo



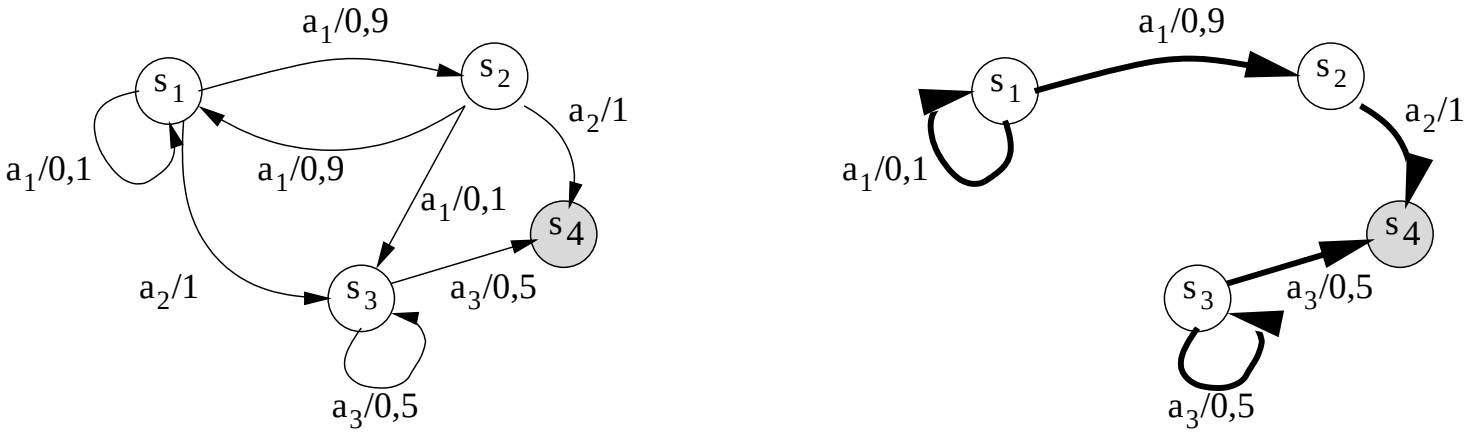
$$
\begin{aligned}
Q^*(s_4, a_i) = \quad & 0 \\
Q^*(s_2, a_2) = \quad & \mathcal{T}(s_2, a_2, s_1)[\mathcal{R}(s_2, a_2, s_1) + \gamma \, \text{máx}_{a'} \, Q^*(s_1, a')] + \\
& \mathcal{T}(s_2, a_2, s_2)[\mathcal{R}(s_2, a_2, s_2) + \gamma \, \text{máx}_{a'} \, Q^*(s_2, a')] + \\
& \mathcal{T}(s_2, a_2, s_3)[\mathcal{R}(s_2, a_2, s_3) + \gamma \, \text{máx}_{a'} \, Q^*(s_3, a')] + \\
& \mathcal{T}(s_2, a_2, s_4)[\mathcal{R}(s_2, a_2, s_4) + \gamma \, \text{máx}_{a'} \, Q^*(s_4, a')] \\
& 1 + \gamma \times 0 = 1 \\
Q^*(s_2, a_1) = \quad & ¿? \\
\end{aligned}
$$

Evaluación de la Política

- Recibir la política π a evaluar
- Inicializar $V(s) = 0, \forall s \in \mathcal{S}$
- Repetir
  - $\Delta \leftarrow 0$
  - Para cada $s \in S$
    1. $v \leftarrow V(s)$
    2. $V(s) \leftarrow \sum_{s'} \mathcal{T}(s, \pi(s), s')[\mathcal{R}(s, \pi(s)) + \gamma V(s')]$
    3. $\Delta \leftarrow max(\Delta, |v - V(s)|)$
- Hasta que $\Delta < \theta$ (un número positivo entero)

Ejemplo: Evaluación de la Política

- Suponemos que la política inicial es: $\pi(s_1) = a_1$, $\pi(s_2) = a_2$, $\pi(s_3) = a_3$ y $\pi(s_4) = a_1$



MDP $\longrightarrow$ Sistema de Markov

Ejemplo: Evaluación de la Política

$$V(s_1) = \mathcal{T}(s_1, a_1, s_1)[\mathcal{R}(s_1, a_1, s_1) + \gamma V(s_1)] + \mathcal{T}(s_1, a_1, s_2)[\mathcal{R}(s_1, a_1, s_2) + \gamma V(s_2)] + \mathcal{T}(s_1, a_1, s_3)[\mathcal{R}(s_1, a_1, s_3) + \gamma V(s_3)] + \mathcal{T}(s_1, a_1, s_4)[\mathcal{R}(s_1, a_1, s_4) + \gamma V(s_4)] = 0,1[0 + 0] + 0,9[0 + 0] + 0,0[0 + 0] + 0,0[0 + 0] = 0$$

$$V(s_2) = \mathcal{T}(s_2, a_2, s_1)[\mathcal{R}(s_2, a_2, s_1) + \gamma V(s_1)] + \mathcal{T}(s_2, a_2, s_2)[\mathcal{R}(s_2, a_2, s_2) + \gamma V(s_2)] + \mathcal{T}(s_2, a_2, s_3)[\mathcal{R}(s_2, a_2, s_3) + \gamma V(s_3)] + \mathcal{T}(s_2, a_2, s_4)[\mathcal{R}(s_2, a_2, s_4) + \gamma V(s_4)] = 0[0 + 0] + 0[0 + 0] + 0[0 + 0] + 1[1 + 0] = 1$$

$$V(s_3) = \mathcal{T}(s_3, a_3, s_1)[\mathcal{R}(s_3, a_3, s_1)) + \gamma V(s_1)] + \mathcal{T}(s_3, a_3, s_2)[\mathcal{R}(s_3, a_3, s_2) + \gamma V(s_2)] + \mathcal{T}(s_3, a_3, s_3)[\mathcal{R}(s_3, a_3, s_3) + \gamma V(s_3)] + \mathcal{T}(s_3, a_3, s_4)[\mathcal{R}(s_3, a_3, s_4) + \gamma V(s_4)] = 0[0 + 0] + 0[0 + 0] + 0,5[0 + 0] + 0,5[1 + 0] = 0,5$$

$$V(s_4) = 0$$

Ejemplo: Evaluación de la Política

	Iter. 1	Iter. 2	Iter. 3	Iter. 4
$V(s_1)$	0	0,9	0,99	0,999
$V(s_2)$	1	1	1	1
$V(s_3)$	0,5	0,75	0,875	0,937
$V(s_4)$	0	0	0	0

Mejora de la Política

- Recibir la política a mejorar π
- Recibir la función de valor $V(s)$ de la política π
- *política_estable* $\leftarrow$ *cierto*
- Para cada $s \in S$
  - $b \leftarrow \pi(s)$
  - $\pi(s) \leftarrow \arg\max_a \sum_{s'} \mathcal{T}(s, a, s')[\mathcal{R}(s, a) + \gamma V(s')]$
  - Si $b \neq \pi(s)$, entonces *política_estable* $\leftarrow$ *falso*

Ejemplo: Mejora de la Política

$$\begin{aligned}
\pi(s_3) = \arg_{a_i} \max(\textstyle\sum_{s'} \mathcal{T}(s_3, a_i, s')[\mathcal{R}(s_3, a_i, s') + \gamma V(s')] = \\
\arg_{a_i} \max(\textstyle\sum_{s'} \mathcal{T}(s_3, a_1, s')[\mathcal{R}(s_3, a_1, s') + \gamma V(s')], \\
\textstyle\sum_{s'} \mathcal{T}(s_3, a_2, s')[\mathcal{R}(s_3, a_2, s') + \gamma V(s')], \\
\mathcal{T}(s_3, a_3, s_1)[\mathcal{R}(s_3, a_3, s_1)) + \gamma V(s_1)] + \\
\mathcal{T}(s_3, a_3, s_2)[\mathcal{R}(s_3, a_3, s_2) + \gamma V(s_2)] + \\
\mathcal{T}(s_3, a_3, s_3)[\mathcal{R}(s_3, a_3, s_3) + \gamma V(s_3)] + \\
\mathcal{T}(s_3, a_3, s_4)[\mathcal{R}(s_3, a_3, s_4) + \gamma V(s_4)]) = \\
\arg_{a_i} \max(0 + 0 + 1[0 + \gamma 0'937] + 0, \\
0 + 0 + 1[0 + \gamma 0'937] + 0, \\
0[0 + 0'999] + 0[0 + 1] + 0'5[0 + 0'937] + 0'5[1 + 0]) = \\
\arg_{a_i} \max(0'843, 0, 843, 0'968) = a_3
\end{aligned}$$

Algoritmo de Iteración de la Política

Permite calcular la función de valor óptima de un MDP:

- Inicialización: $V(s) \in \Re$ y $\pi(s) \in \mathcal{A}$ arbitrarios para todo $s \in \mathcal{S}$
- Repetir
  1. Evaluación de la Política
  2. Mejora de la Política
- Hasta que $pol\acute{i}tica\_estable = cierto$

Algoritmo de Iteración de Valor

Iteración de Valor

- Inicializar $V(s)$ arbitrariamente. Por ejemplo, $V(s) = 0, \forall s \in \mathcal{S}$
- Repetir
  - $\Delta \leftarrow 0$
  - Para todo $s \in S$
    - $v \leftarrow V(s)$
    - $V(s) \leftarrow \text{máx}_a \sum_{s'} \mathcal{T}(s, a, s')[\mathcal{R}(s, a) + \gamma V(s')]$
    - $\Delta \leftarrow max(\Delta, |v - V(s)|)$
- Hasta que $\Delta < \theta$ (un número positivo entero)
- Dar como salida una política π tal que $\pi(s) = \arg \text{máx}_a \sum_{s'} \mathcal{T}(s, a, s')[\mathcal{R}(s, a) + \gamma V(s')]$