Introducción al Aprendizaje Reforzado

Conceptos básicos y métodos tabulares

Julio Waissman Vilanova

Departamento de Matemáticas
Universidad de Sonora

Universidad Autónoma de Baja California, marzo 2010

Plan de la presentación

Conceptos Básicos.

Métodos tabulares.

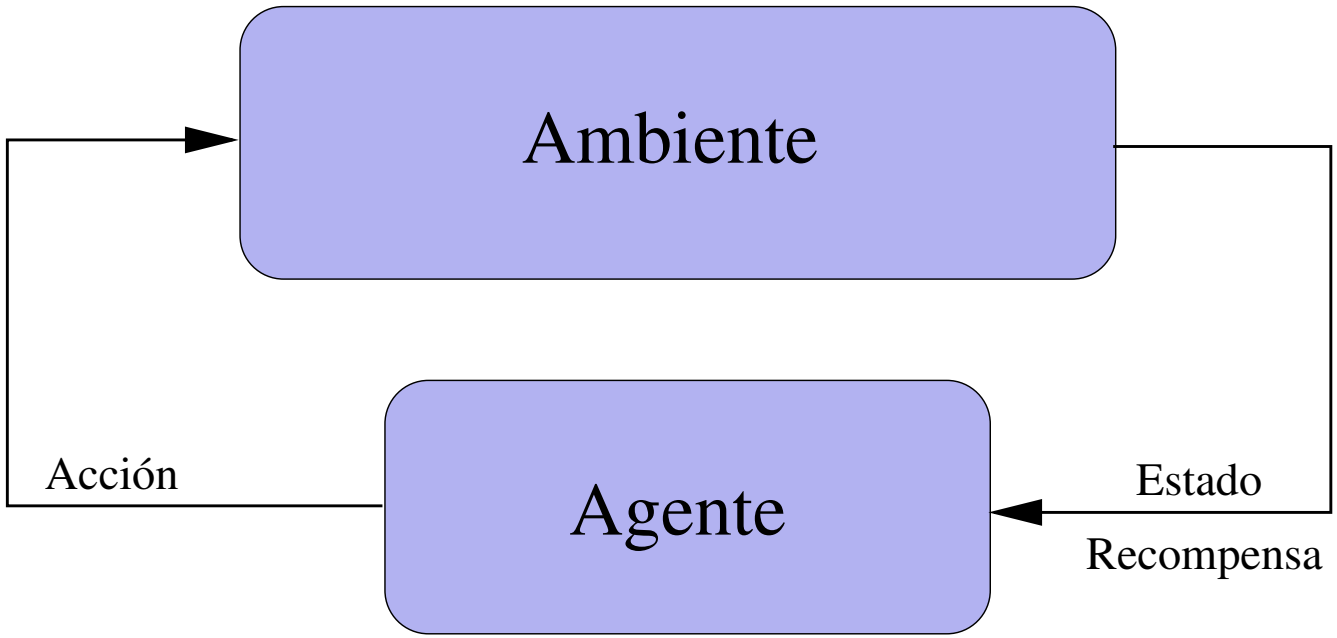
Estados continuos.

Conclusiones.

¿Aprendizaje?

- El aprendizaje supervisado utiliza un conjunto de datos de aprendizaje previamente clasificado. Aprendizaje con maestro.
- El aprendizaje no supervisado utiliza un conjunto de datos sin clasificar. Descubrimiento de conocimiento en bases de datos (KDD).
- El aprendizaje reforzado utiliza la interacción con el medio para establecer una política de comportamiento. Aprendizaje con crítico.

Esquema general del aprendizaje reforzado.



Elementos principales.

- Conjunto de estados, $s_t \in S$, con al menos un estado inicial y posiblemente estados finales.
- Conjunto de acciones en cada estados, $a_t \in A(s_t)$.
- Valor de recompensa, $r_t \in \mathbb{R}$.
- Una política, π, con $\pi(s_t, a_t) \in [0, 1]$

$$\pi = \left[\begin{array}{c|ccc} & s_1 & \cdots & s_n \\ \hline a_1 & 0.1 & \cdots & 0.9 \\ \vdots & \vdots & \ddots & \vdots \\ a_m & 0.9 & \cdots & 0.7 \end{array}\right]$$

Objetivo.

- Encontrar una política subóptima π^* de operación.
- El criterio de optimización es la maximización del regreso, definido como:

$$R_t = \sum_{k=t}^{T} r_k \qquad \text{(episódico)},$$

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \qquad \text{(continuo)}.$$

- Utilizando la exploración/explotación de información.

Funciones de valor.

- Evaluación de un estado,

$$V^{\pi}(s) = E_{\pi}\Big\{ R_t | s_t = s \Big\},$$
$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}(s') \right].$$

- Evaluación de una acción en un estado,

$$Q^{\pi}(s,a) = E_{\pi}\Big\{ R_t | s_t = s, a_t = a \Big\}.$$

- Permite encontrar políticas óptimas,

$$V^{*}(s) = \max_{\pi} V^{\pi}(s), \qquad Q^{*}(s,a) = \max_{\pi} Q^{\pi}(s,a).$$

Método de diferencias temporales.

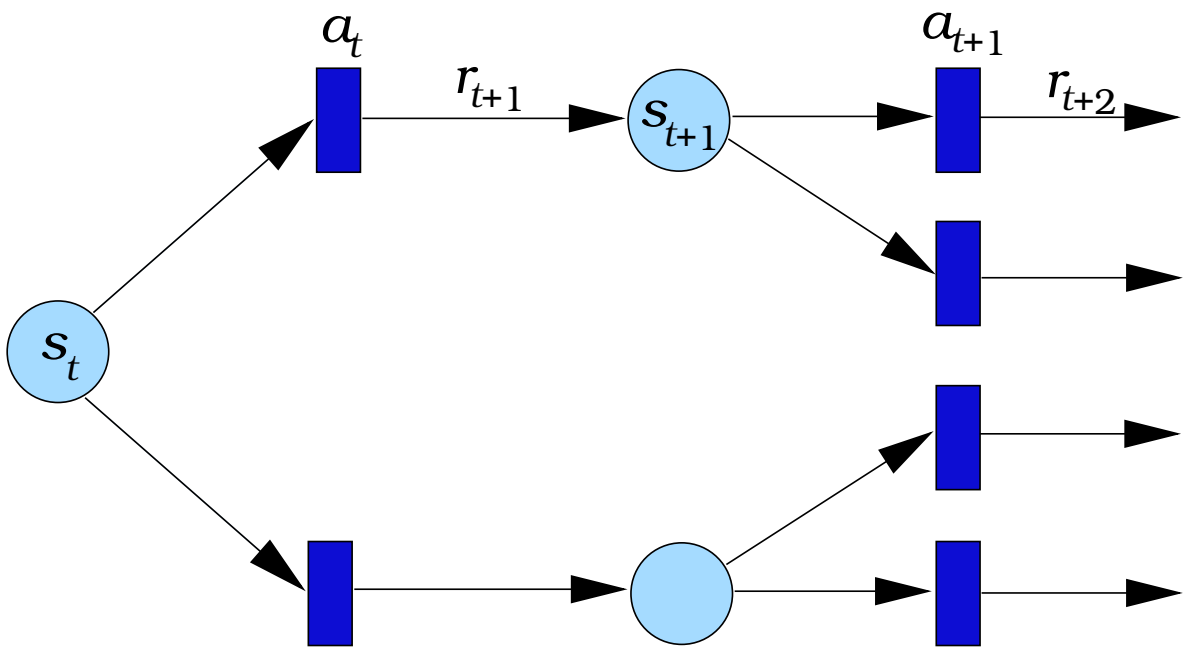
- Utilizan la experiencia para encontrar la función de valor.
- Calcula una política pseudo–óptima.
- Se basan en la actualización por el nuevo estado:

$$V_t(s_t) \leftarrow V_t(s_t) + \alpha\Big[V_{t+1}(s_t) - V_t(s_t)\Big]$$

$$V_t(s_t) \leftarrow V_t(s_t) + \alpha\Big[r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)\Big],$$

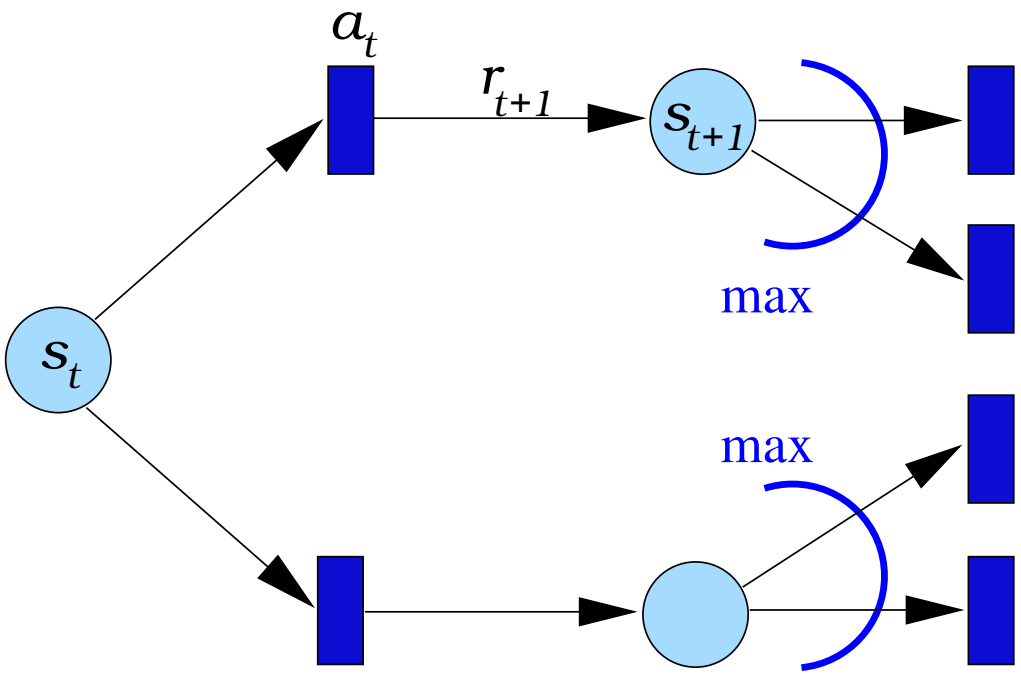
donde $\alpha \in [0, 1]$ es el factor de aprendizaje.

Método SARSA.



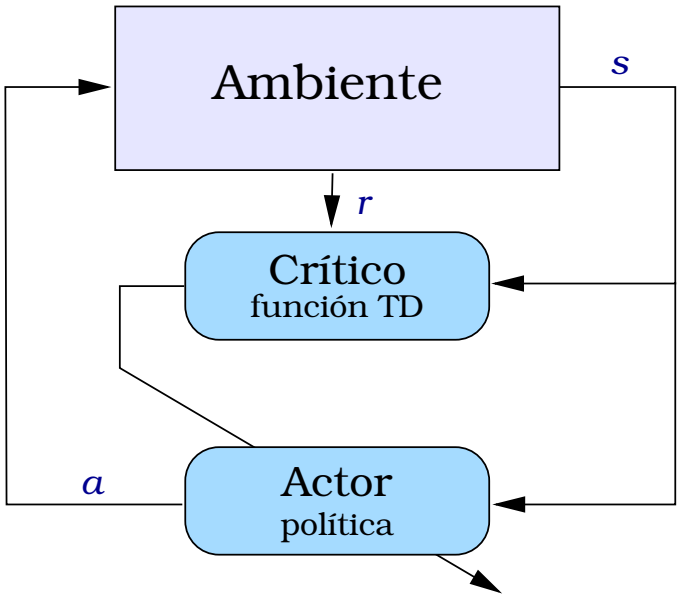
$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right]$$

Método QLearning.



$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

Método *Actor/Critic*.



$$p(s_t, a_t) = p(s_t, a_t) + \beta \left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right]$$

$$\pi_t(s, a) = \frac{e^{p(s,a)}}{\sum_b e^{p(s,b)}}$$

Exploración / Explotación del conocimiento.

- Utilizar el conocimiento adquirido por el agente.
- Explorar racionalmente estados desconocidos.
- Los métodos clásicos son:
  - Avaro
  - ϵ–Avaro
  - *Softmax* o Distribución de Boltzmann,

$$P(a_t|s_t) = \frac{e^{p(a_t,s_t)/T}}{\sum_b e^{p(b,s_t)/T}},$$

donde T es la *temperatura*.

Para más información


R.S. Sutton y A.G. Barto.
Reinforcement Learning. An Introduction.
MIT Press, 2002 (4^a Impresión).


L. Kaelbling, M. Littman y A. Moore.
Reinforcement Learning: A Survey.
Journal of Artificial Inteligence Research, 4:237–285, 1996.

Gracias por su atención