Aprendizaje por refuerzo

Verónica E. Arriola-Rios

Inteligencia Artificial

21 de agosto de 2021

Antecedentes: Psicología

¿Conductismo?

Figura: Asociación de estímulos

Reforzadores positivos y negativos

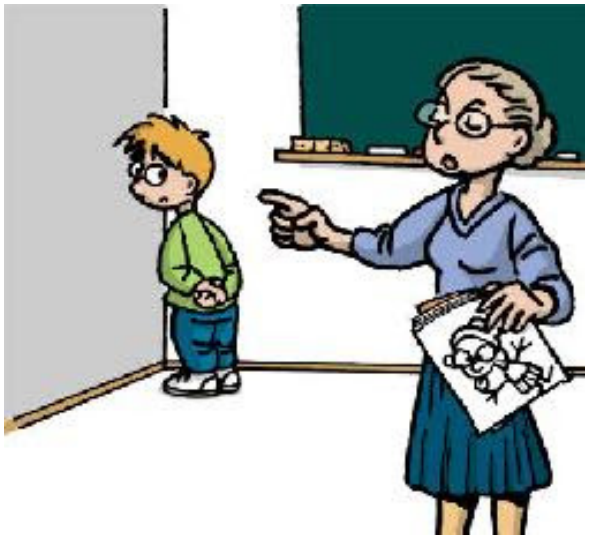


Figura: Izquierda: reforzador negativo. Derecha: reforzador positivo.

Aprendizaje por refuerzo

- Cada vez que un agente realiza una acción a en su ambiente, un entrenador le otorga un *premio* o un *castigo* $r(s, a)$ para indicar lo deseable que es el estado resultante.

$$s_0 \xrightarrow[r_0]{a_0} s_1 \xrightarrow[r_1]{a_1} s_2 \xrightarrow[r_2]{a_2} \cdots$$

- El objetivo del agente es aprender a elegir *secuencias de acciones* que produzcan la recompensa más alta, a partir de esta recompensa indirecta y postergada.
- Se busca aprender una política de control

$$\pi : S \to A \tag{1}$$

que indique la acción $a \in A$ que se debe realizar, en el estado actual $s \in S$.
Se puede hacer una analogía con aprender a reconocer por instinto la acción más conveniente dado un estado.

- Entonces se tienen:
  - Una función de transición:

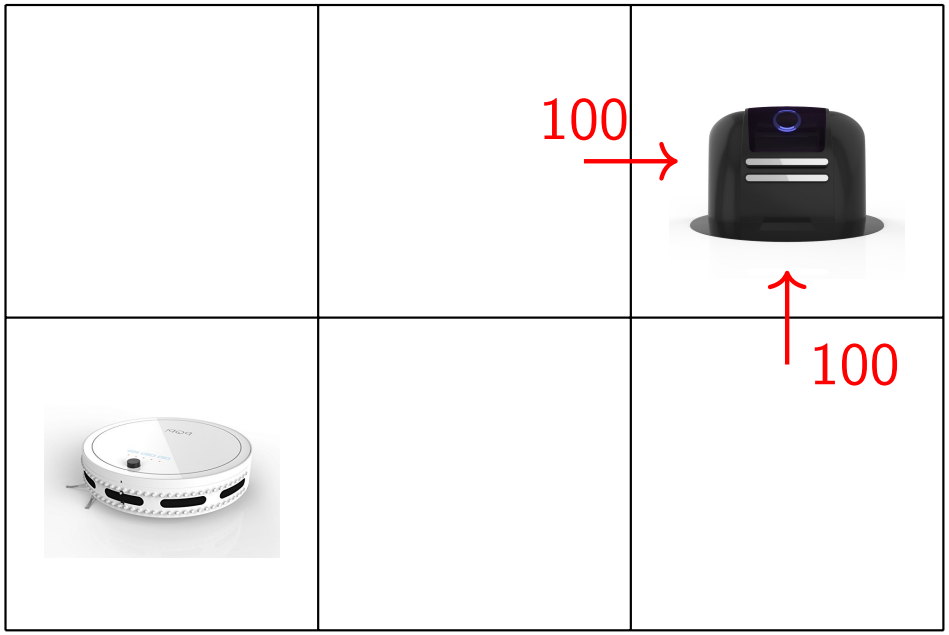
$$\gamma : S \times A \rightarrow S = \gamma(s, a) \rightarrow s' \tag{2}$$

  - Una función de recompensa que otorga el ambiente (o entrenador).

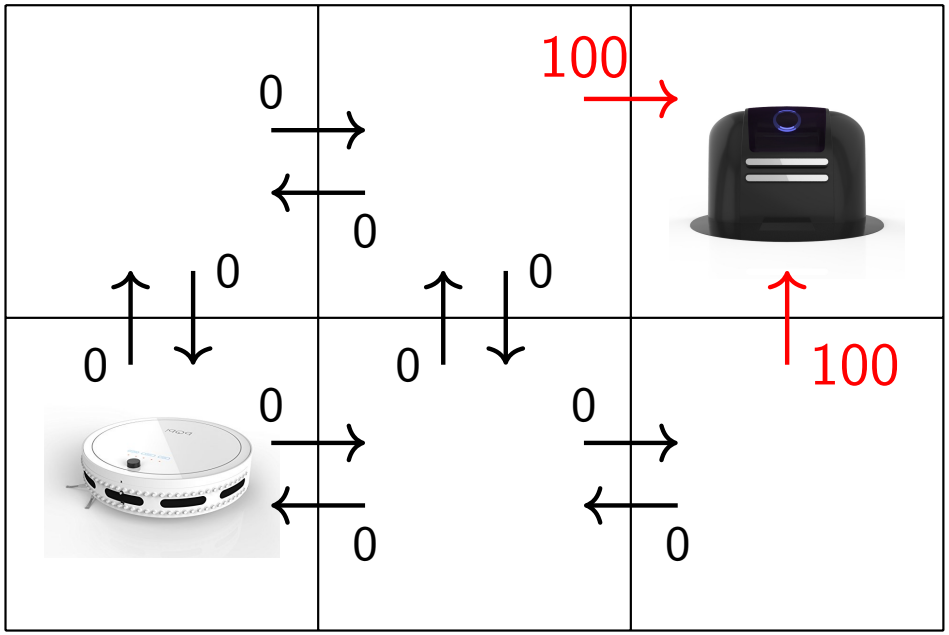
$$r : S \times A \rightarrow \mathbb{R} = r(s, a) \tag{3}$$

Figura: En el juego Assault, se ganan puntos por acertar disparos a los enemigos y se pierde vida cuando los enemigos aciertan o tocan a la nave.

Ejemplo



- El robot recibirá una recompensa de 100 cuando ejecute aquella acción que le permita llegar a su cargador.



- Pero no recibe ninguna recompensa para ninguna otra acción.
- ¿Qué secuencia de acciones debe realizar para alcanzar la recompensa?

Recompensa acumulada descontada

- Se desea que la recompensa sea alcanzada en el menor tiempo posible.
- Sea $V^{\pi}(s_t)$ el *valor acumulado* alcanzado por la política π desde algún estado s_t:

$$V^{\pi}(s_t) \equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \quad (4)$$

$$\equiv \sum_{i=0}^{\infty} \gamma^i r_{t+i} \quad (5)$$

con $0 \leqslant \gamma < 1$.

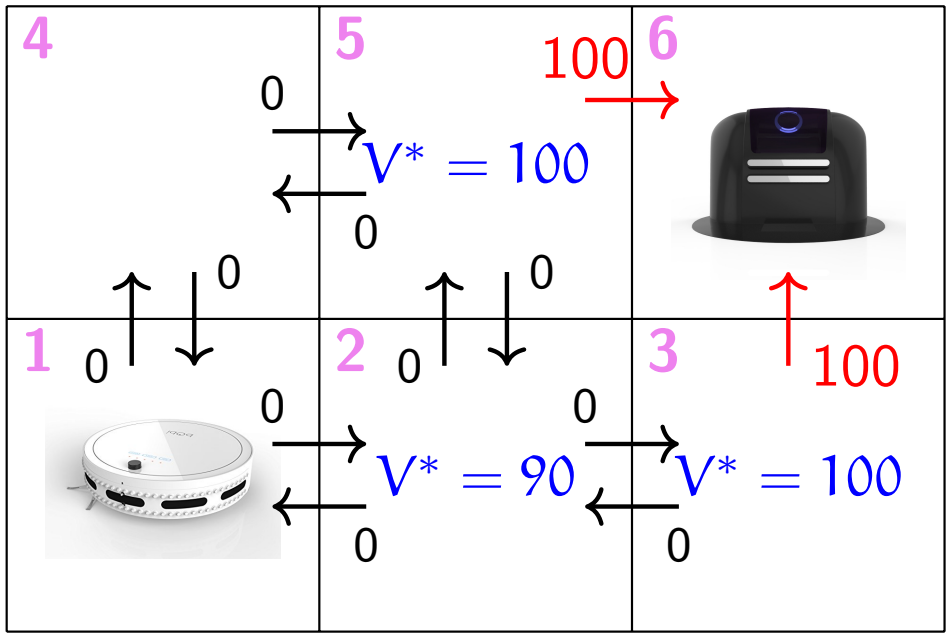
¿Qué quiere aprender el agente?

El agente busca aprender cuál es la política π que maximiza $V^{\pi}(s)$ para todos los estados s.

$$\pi^{*} \equiv \text{argmax}_{\pi} V^{\pi}(s), (\forall s) \tag{6}$$

Recompensa acumulada en el caso óptimo

π	
s	a
1	$\rightarrow$
2	$\rightarrow$
3	$\uparrow$
4	$\rightarrow$
5	$\rightarrow$
6	Fin



- $\gamma = 0.9$

Aprendizaje Q

El algoritmo se basa en la definición de la función Q como:

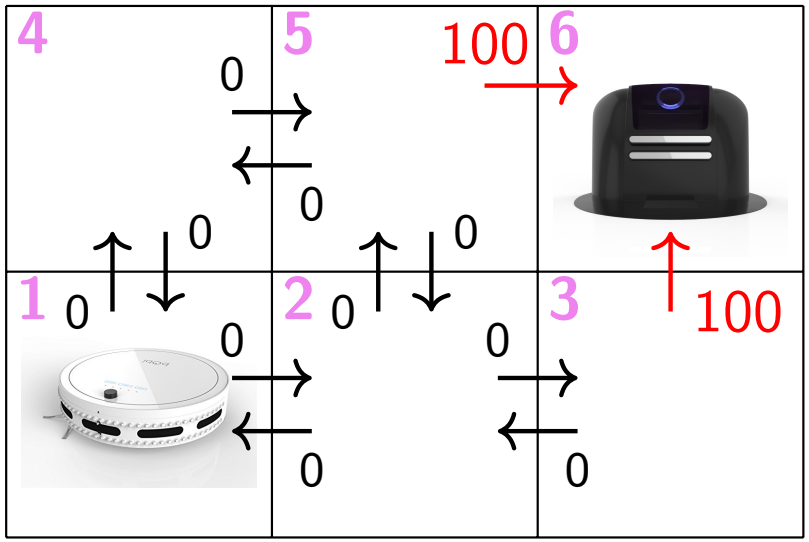
$$Q(s,a) = r(s,a) + \gamma V^*(\gamma(s,a)) \quad (7)$$

$$= r(s,a) + \gamma \max_{a'} Q(\gamma(s,a), a') \quad (8)$$

- El robot realiza un recorrido, buscando su cargador.
- En cada paso actualiza la tabla de la aproximación a la función Q, $\hat{Q}$ utilizando:

$$\hat{Q}(s_t, a_t) \leftarrow r_t + \gamma \max_{a_{t+1}} \hat{Q}(s_{t+1}, a_{t+1}) \quad (9)$$

$\hat{Q}(a_t, s_t)$		
s	a	$\hat{Q}$
1	$\rightarrow$	
1	$\uparrow$	
2	$\leftarrow$	
2	$\rightarrow$	
2	$\uparrow$	
3	$\leftarrow$	
3	$\uparrow$	
4	$\rightarrow$	
4	$\downarrow$	
5	$\leftarrow$	
5	$\rightarrow$	
5	$\downarrow$	
6	Fin	



$\hat{Q}(a_t, s_t)$		
s	a	$\hat{Q}$
1	$\rightarrow$	0
1	$\uparrow$	0
2	$\leftarrow$	0
2	$\rightarrow$	0
2	$\uparrow$	0
3	$\leftarrow$	0
3	$\uparrow$	0
4	$\rightarrow$	0
4	$\downarrow$	0
5	$\leftarrow$	0
5	$\rightarrow$	0
5	$\downarrow$	0
6	Fin	**0**

$\Rightarrow$

$\hat{Q}(a_t, s_t)$		
s	a	$\hat{Q}$
1	$\rightarrow$	
1	$\uparrow$	
2	$\leftarrow$	
2	$\rightarrow$	
2	$\uparrow$	
3	$\leftarrow$	
3	$\uparrow$	
4	$\rightarrow$	
4	$\downarrow$	
5	$\leftarrow$	
5	$\rightarrow$	
5	$\downarrow$	
6	Fin	

Referencias I

 Mitchell, Tom M. (1997). *Machine Learning*. McGrawHill.