

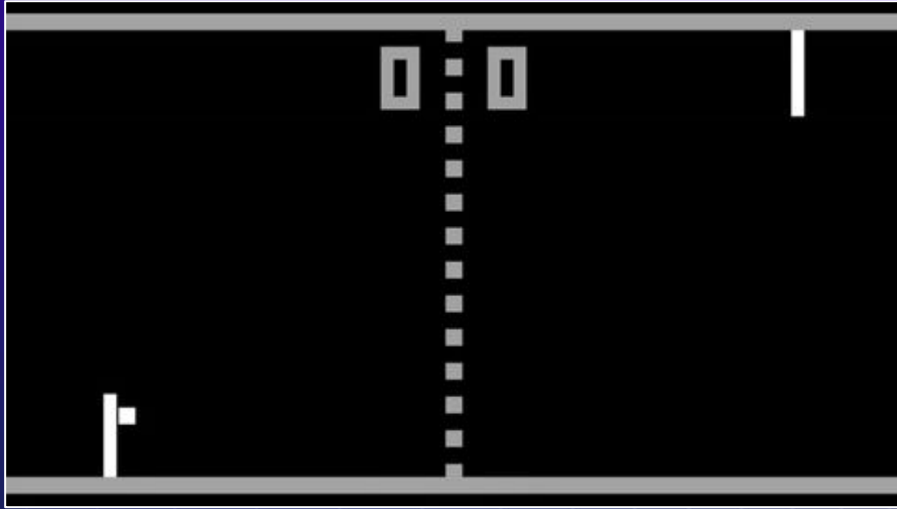


Aprendizaje por refuerzo Q-Learning

Rigonatto, Eduardo

Aprendizaje por refuerzo

El aprendizaje de refuerzo es otro enfoque del aprendizaje automático, donde después de cada acción, el agente obtiene retroalimentación



Ejemplo intuitivo: Pong

.....

Ejemplo intuitivo: Pong



**Controlar la
paleta**

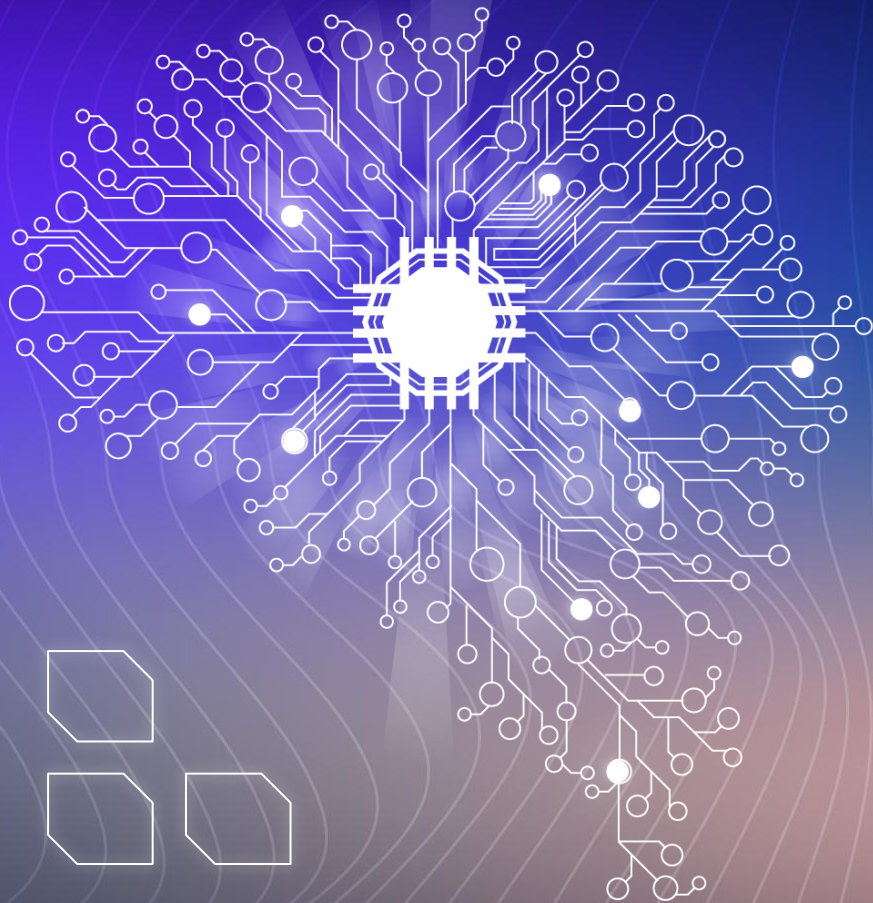


**Golpear la
bola**



**Marcar
puntos**

**Cómo logramos
esto con una
máquina?**



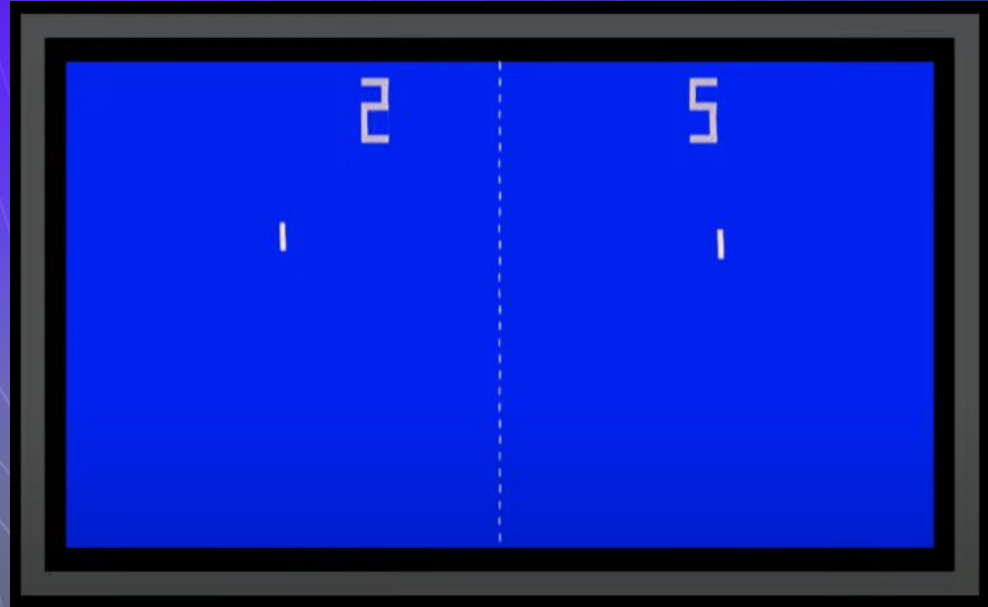


El Agente





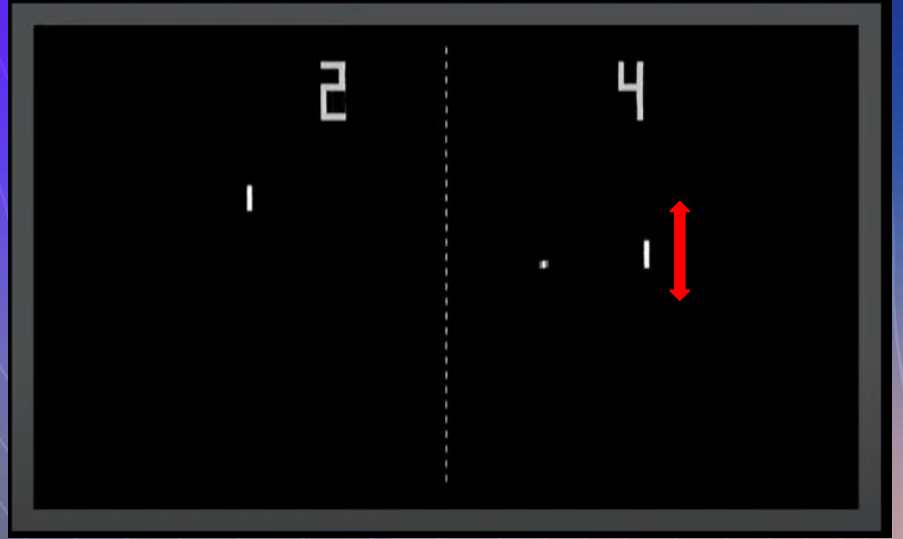
El Entorno



El Estado



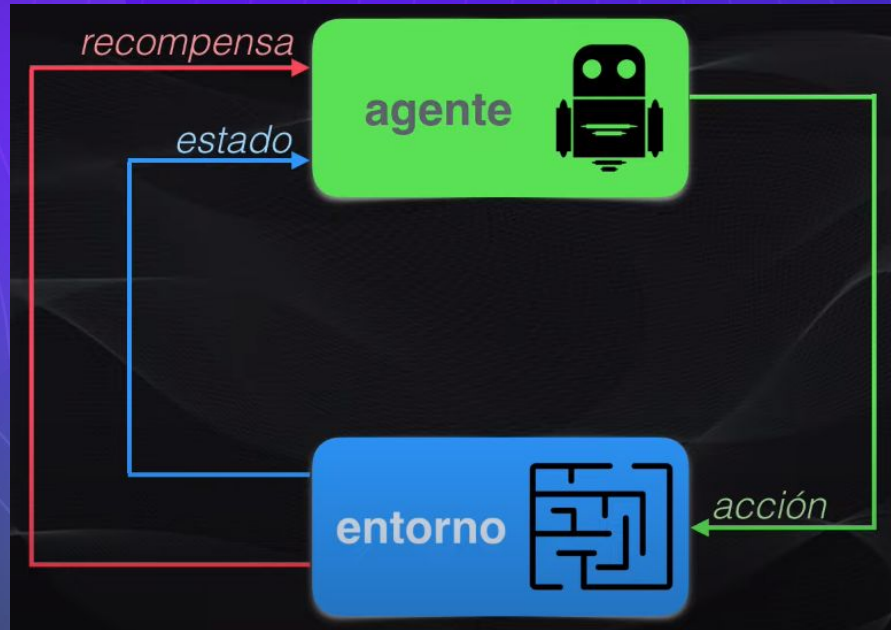
Accion





Recompensa







Q-Learning

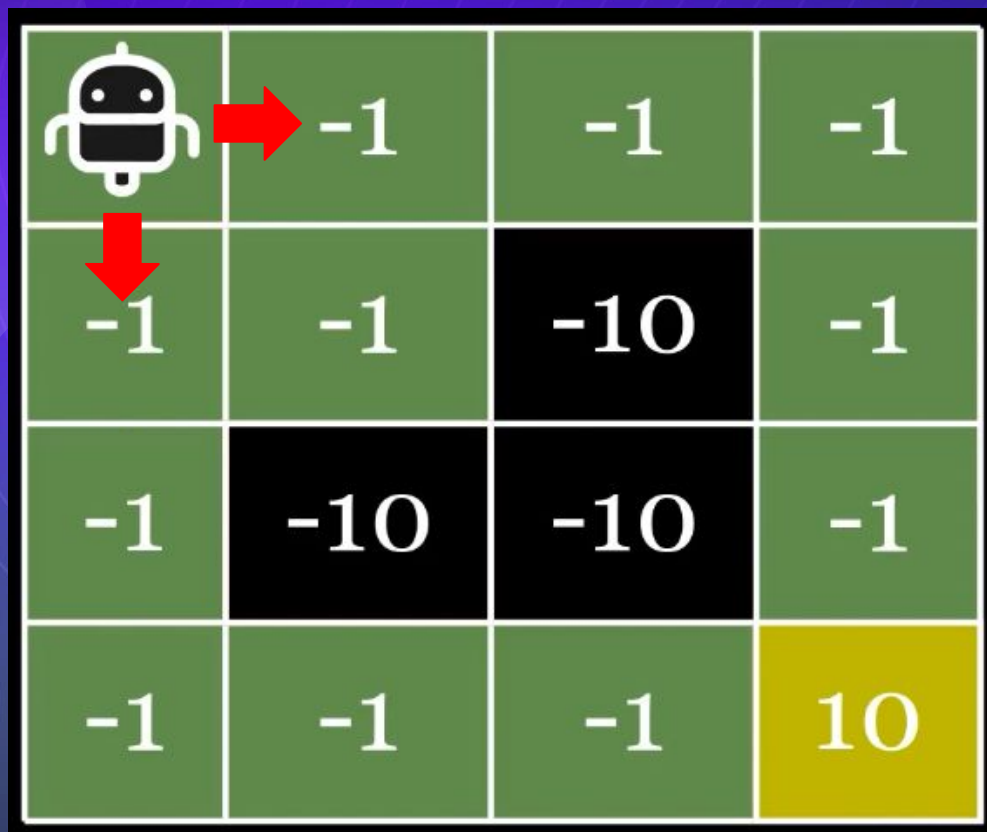


**Aprendizaje reforzado
libre de modelos**



Politica: La funcion Q

	-1	-1	-1
-1	-1	-10	-1
-1	-10	-10	-1
-1	-1	-1	10



	-1	-1	-1
-1	-1	-10	-1
-1	-10	-10	-1
-1	-1	-1	10

arriba=0.89 abajo=0.32 izquierda=0.13 derecha=0.75	arriba=0.60 abajo=0.34 izquierda=0.16 derecha=0.64	arriba=0.06 abajo=0.31 izquierda=0.43 derecha=0.99	arriba=0.69 abajo=0.79 izquierda=0.18 derecha=0.93
arriba=0.20 abajo=0.25 izquierda=0.97 derecha=0.58	arriba=0.94 abajo=0.14 izquierda=0.81 derecha=0.31	arriba=0.81 abajo=0.57 izquierda=0.48 derecha=0.96	arriba=0.68 abajo=0.95 izquierda=0.76 derecha=0.14
arriba=0.96 abajo=0.20 izquierda=0.82 derecha=0.43	arriba=0.11 abajo=0.27 izquierda=0.32 derecha=0.78	arriba=0.1 abajo=0.21 izquierda=0.36 derecha=0.88	arriba=0.02 abajo=0.19 izquierda=0.31 derecha=0.66
arriba=0.43 abajo=0.74 izquierda=0.39 derecha=0.91	arriba=0.48 abajo=0.42 izquierda=0.22 derecha=0.42	arriba=0.97 abajo=0.36 izquierda=0.67 derecha=0.45	arriba=0.78 abajo=0.75 izquierda=0.22 derecha=0.41



La funcion Q


$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max(Q'(s', a')) - Q(s, a)]$$


$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max(Q'(s', a')) - Q(s, a)]$$

Q(s,a)


Nuevo valor de recompensa para ese estado en el que nos encontremos y la acción que ejecutaremos

arriba=0.89
abajo=0.32
izquierda=0.13
derecha=0.75

$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max(Q'(s', a')) - Q(s, a)]$$


Q(s,a)

Nuevo valor de recompensa para ese estado en el que nos encontremos y la acción que ejecutaremos


$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max_{a'} Q'(s', a')) - Q(s, a)]$$



Q(s,a)

Valor del estado y acción actual



arriba=0.89 abajo=0.32 izquierda=0.13 derecha=0.75	arriba=0.60 abajo=0.34 izquierda=0.16 derecha=0.64	arriba=0.06 abajo=0.31 izquierda=0.43 derecha=0.99	arriba=0.69 abajo=0.79 izquierda=0.18 derecha=0.93
arriba=0.20 abajo=0.25 izquierda=0.97 derecha=0.58	arriba=0.94 abajo=0.14 izquierda=0.81 derecha=0.31	arriba=0.81 abajo=0.57 izquierda=0.48 derecha=0.96	arriba=0.68 abajo=0.95 izquierda=0.76 derecha=0.14
arriba=0.96 abajo=0.20 izquierda=0.82 derecha=0.43	arriba=0.11 abajo=0.27 izquierda=0.32 derecha=0.78	arriba=0.1 abajo=0.21 izquierda=0.36 derecha=0.88	arriba=0.02 abajo=0.19 izquierda=0.31 derecha=0.66
arriba=0.43 abajo=0.74 izquierda=0.39 derecha=0.91	arriba=0.48 abajo=0.42 izquierda=0.22 derecha=0.42	arriba=0.97 abajo=0.36 izquierda=0.67 derecha=0.45	arriba=0.78 abajo=0.75 izquierda=0.22 derecha=0.41



arriba=0.89
abajo=0.32
izquierda=0.13
derecha=0.75


$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max_{a'} Q'(s', a')) - Q(s, a)]$$

Tasa de aprendizaje

Determina el tamaño del paso en cada iteración




$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max_{a'} Q'(s', a')) - Q(s, a)]$$

Recompensa

Es la recompensa inmediata después de tomar la acción a en el estado s



	-1	-1	-1
-1	-1	-10	-1
-1	-10	-10	-1
-1	-1	-1	10


$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma (\max_{a'} Q'(s', a')) - Q(s, a)]$$

Factor de descuento

Determina la importancia de las recompensas futuras




$$Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max(Q'(s', a')))] - Q(s, a]$$

$\max(Q'(s', a'))$

Máximo valor para el nuevo estado y cualquier accion



arriba=0.89 abajo=0.32 izquierda=0.13 derecha=0.75	arriba=0.60 abajo=0.34 izquierda=0.16 derecha=0.64	arriba=0.06 abajo=0.31 izquierda=0.43 derecha=0.99	arriba=0.69 abajo=0.79 izquierda=0.18 derecha=0.93
arriba=0.20 abajo=0.25 izquierda=0.97 derecha=0.58	arriba=0.94 abajo=0.14 izquierda=0.81 derecha=0.31	arriba=0.81 abajo=0.57 izquierda=0.48 derecha=0.96	arriba=0.68 abajo=0.95 izquierda=0.76 derecha=0.14
arriba=0.96 abajo=0.20 izquierda=0.82 derecha=0.43	arriba=0.11 abajo=0.27 izquierda=0.32 derecha=0.78	arriba=0.1 abajo=0.21 izquierda=0.36 derecha=0.88	arriba=0.02 abajo=0.19 izquierda=0.31 derecha=0.66
arriba=0.43 abajo=0.74 izquierda=0.39 derecha=0.91	arriba=0.48 abajo=0.42 izquierda=0.22 derecha=0.42	arriba=0.07 abajo=0.11 izquierda=0.67 derecha=6.98	arriba=0.78 abajo=0.75 izquierda=0.22 derecha=0.41

arriba=0.89 abajo=0.32 izquierda=0.13 derecha=0.75	arriba=0.60 abajo=0.34 izquierda=0.16 derecha=0.64	arriba=0.06 abajo=0.31 izquierda=0.43 derecha=0.99	arriba=0.69 abajo=0.79 izquierda=0.18 derecha=0.93
arriba=0.20 abajo=0.25 izquierda=0.97 derecha=0.58	arriba=0.94 abajo=0.14 izquierda=0.81 derecha=0.31	arriba=0.81 abajo=0.57 izquierda=0.48 derecha=0.96	arriba=0.68 abajo=0.95 izquierda=0.76 derecha=0.14
arriba=0.96 abajo=0.20 izquierda=0.82 derecha=0.43	arriba=0.11 abajo=0.27 izquierda=0.32 derecha=0.78	arriba=0.1 abajo=0.21 izquierda=0.36 derecha=0.88	arriba=0.02 abajo=0.19 izquierda=0.31 derecha=0.66
arriba=0.43 abajo=0.74 izquierda=0.39 derecha=0.91	arriba=0.48 abajo=0.42 izquierda=0.22 derecha=0.42	arriba=0.07 abajo=0.11 izquierda=0.67 derecha=6.98	arriba=0.78 abajo=0.75 izquierda=0.22 derecha=0.41

$$Nuevo Q(s, a) = Q(s, a) + \alpha [R(s, a) + \gamma(\max(Q'(s', a')) - Q(s, a)]$$

$$Nuevo Q(s, a) = 0.42 + 0.1 [-1 + 0.95(6.98) - 0.42] = 0.9411$$

Ejemplo práctico: Nim

Nim



01

Se colocan varios montones de elementos sobre una mesa

02

Los jugadores se turnan para tomar elementos de uno de los montones.

03

Un jugador puede tomar cualquier cantidad de elementos, pero solo de un solo montón.

04

El jugador que toma el último elemento de los montones pierde.



.....

Gracias por su atencion!

.....

