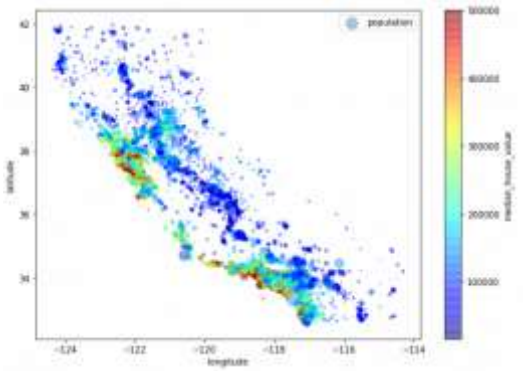
Aprendizaje por refuezo (Reinforcement Learning)

DATA SCIENCE

Previously on… DataScience Bootcamp

- Hemos construido y entrenado modelos que resuelven problemas de clasificación, regresión y agrupamiento o clustering… utilizando métodos supervisados o no-supervisados.

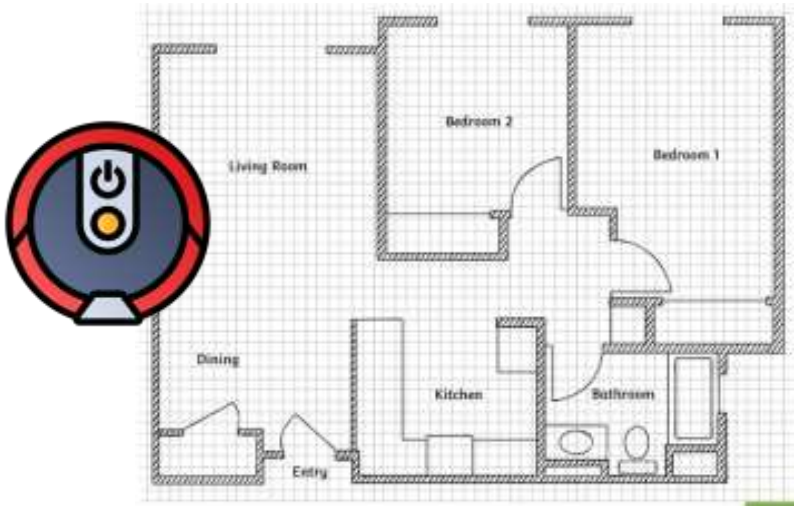


… ahora: Aprendizaje por Refuerzo

El aprendizaje por refuerzo es aquel en el que se busca que un agente (un programa o un robot) "aprenda" como actuar ante determinadas situaciones en base a una recompensa (o castigo, si la recompensa es negativa) que se asigna a la respuesta que proporcione el agente ante dichas situaciones…

“Ejemplos”

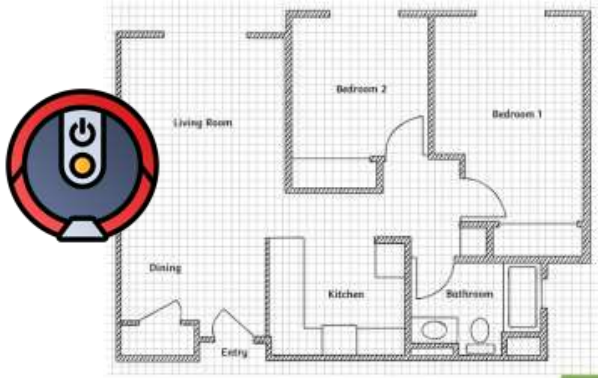
- Enseñar a un lindo perrito a darnos la patita



- Programar un útil robot aspiradora para limpiar nuestra casa

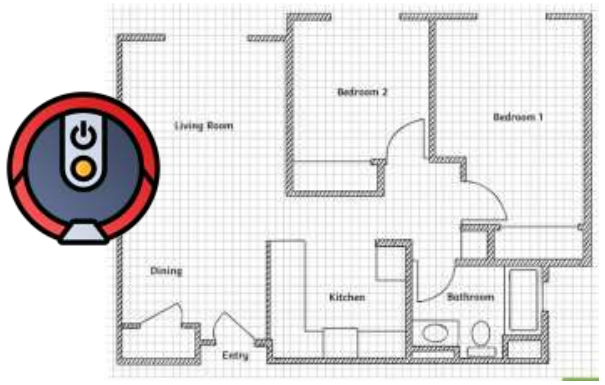
Aplicando aprendizaje por refuerzo

- Le mostramos nuestra mano
- El perro actuará (se quedará quieto, moverá el rabo, se paseará, nos mirará, o nos dará la patita…)
- Si nos da la patita le damos una recompensa (un refuerzo positivo)
- Si no, no hacemos nada o le decimos “NO” (refuerzo negativo)
- Repetimos hasta que el perrito haya aprendido



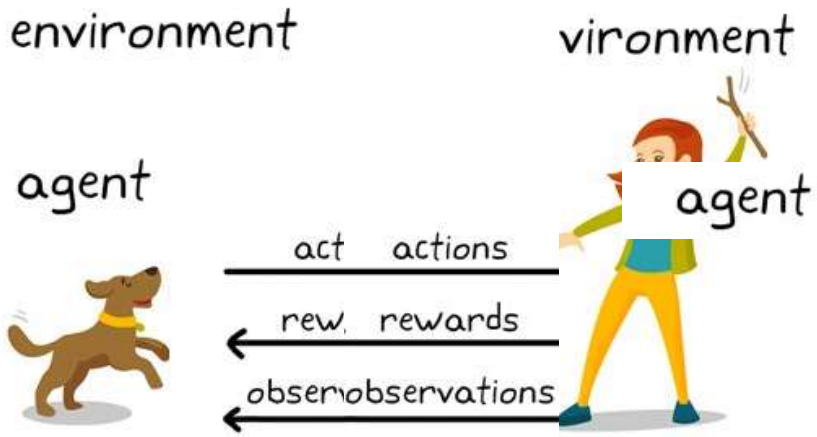
- El robot puede avanzar, girar y aspirar, al tiempo que registra su posición (x,y) y detecta choques
- Comienza avanzando aleatoriamente, y se apunta una “recompensa” cada vez que:
  - Pasa por x,y no visitadas
  - Llena la bolsa un %
- Le añadimos un mecanismo que controle el giro y el avance para busque la mayor recompensa

Elementos del aprendizaje por refuerzo



- **Agente (Agent):** Perrito y Robot
- **Entorno/Ambiente (Environment):** El ser humano y la zona de entrenamiento; la casa y sus condiciones de suciedad.
- **Observaciones (Observations) y Estado (State):** Las posturas y respuestas del dueño/entrenador del perrito; las diferentes combinaciones de posición (x,y), número de veces que se ha pasado por esa posición, y el volumen ocupado de la bolsa de vaciado.
- **Recompensa (Reward):** La galletita, los puntos positivos o negativos ganados por el robot
- **Acción (Action):** comportamientos del perro; avanzar, girar, aspirar, en el caso del robot
- **Política (Policy):** El criterio, reglas, programa o modelo que usa el agente para decidir cuál es la mejor acción dado un estado

Proceso



1. Para empezar se “observa” el **entorno**: Nos da las **observaciones** y el **agente** puede saber el estado en el que está
1. El **agente** decide cómo actuar, en general, en función de ese **estado** y de acuerdo a la **política** que tenga en ese momento
1. Actúa
1. El **entorno** le “devuelve” el **estado** (o las observaciones necesarias para definirlo) resultante y la **recompensa** (positiva o negativa)
1. El **agente** aprende de la experiencia: actualiza la **política** o estrategia**.**
1. Se itera hasta conseguir una estrategia o **política** óptima

Consideraciones y Aplicaciones

- El reinforcement learning se suele considerar como un tercer tipo de aprendizaje (ni supervisado, ni no-supervisado)
- La secuencia de acciones influye en la recompensa final. El “tiempo” sí es importante.
- Ser “glotón” (greedy) no siempre es lo deseado.
- Aplicado a todo tipo de bots: mecánicos, digitales (e.j: ChatGPT)

- … pero también a problemas como los del principio, sobre todo en sistemas de recomendación

- Aunque para la investigación se emplean principalmente videojuegos