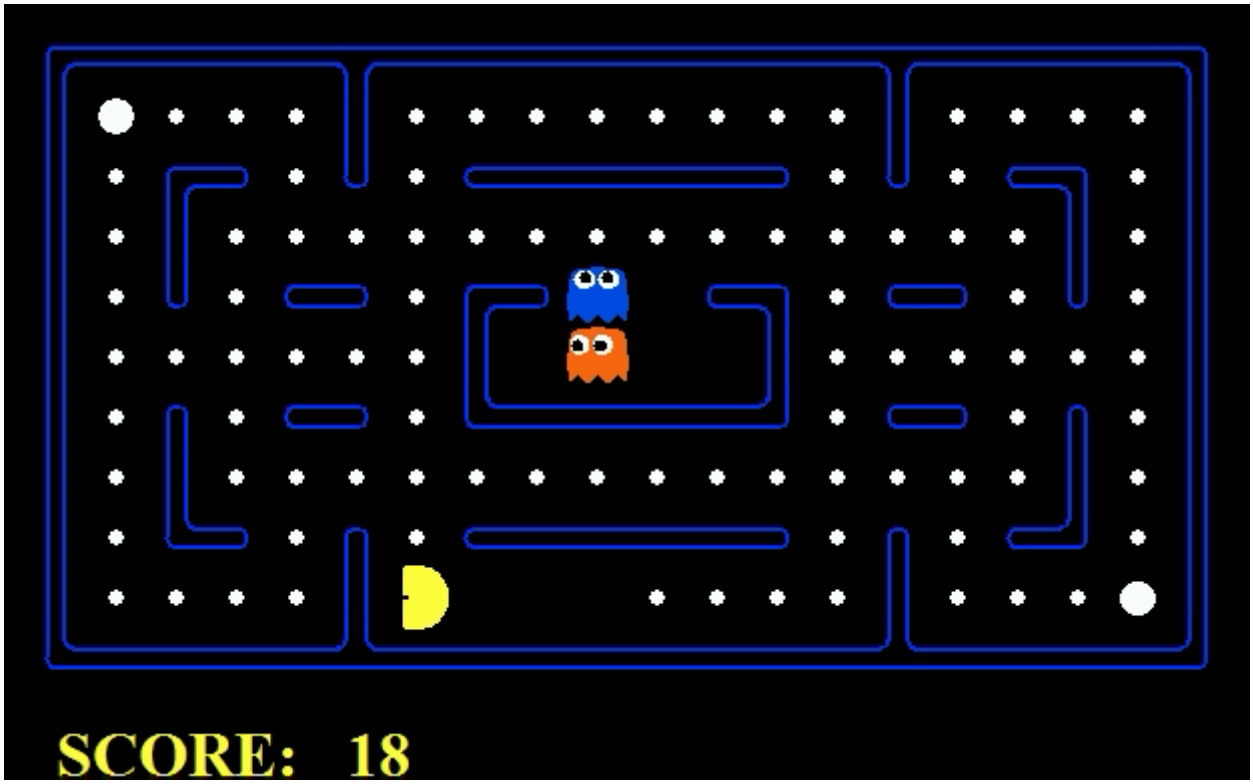
Aprendizaje por Refuerzo: Introducción al mundo del RL

Miguel Silva
Apr 27, 2019 · 7 min read

Hola a todos, esta será una serie de posts donde se explicará la teoría detrás de el aprendizaje por refuerzo lo más simple y cautivadoramente que pueda para juntos introducirnos en este mundo, al final dejaré links donde se pueda investigar más profundamente.

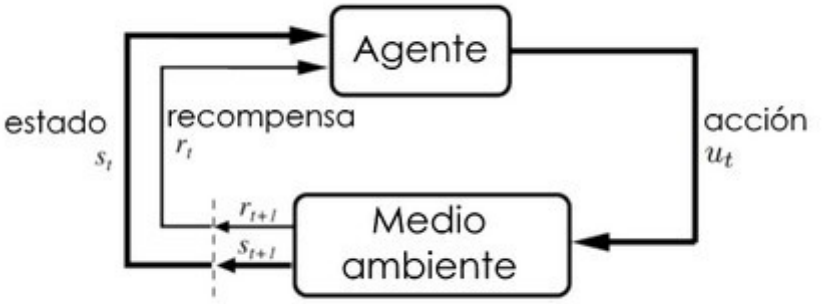
Si deseas ver toda la serie de post completas accede aquí.
Me centraré en hacerlo en español debido a que no he encontrado un recurso parecido sobre este tema en este idioma. Cualquier comentario y/o modificación que se proponga será bien recibido.



Pacman siendo jugado por un agente de aprendizaje por refuerzo.

¿Que es aprendizaje por refuerzo?

El aprendizaje por refuerzo es una área de la inteligencia artificial que esta centrada en descubrir que acciones se debe tomar para maximizar la señal de recompensa, en otras palabras se centra en como mapear situaciones a acciones que se centren en encontrar dicha recompensa. Al agente no se le dice que acciones tomar, si no al contrario el debe experimentar para encontrar que acciones lo llevan a una mayor recompensa, los casos mas desafiantes son los que no llevan a una recompensa inmediata si no en la siguientes situaciones.



[Figure source: Sutton & Barto, 1998]

Diferencia con aprendizaje supervisado

El aprendizaje supervisado esta basado en un set de ejemplos que han sido etiquetados previamente, en otras palabras te muestra exactamente que acción debes tomar para cada circunstancia, la que mayormente consiste en identificar o categorizar la situación; por lo cual el objetivo de este tipo de aprendizaje es generalizar o extrapolar para situaciones que no estén presentes en los datos de entrenamiento. Esto sin embargo no es beneficioso para los problemas interactivos, debido a que al saber que acciones debe desempeñar el agente esta técnica termina siendo inefectiva en territorio no explorado, es ahí donde el agente debe aprender por su propia experiencia.

Diferencia con aprendizaje no supervisado

El aprendizaje no supervisado se centra en encontrar estructuras dentro de una colección de datos, cuando el aprendizaje por refuerzo se enfoca en aumentar la señal de recompensa.

Debido a lo que hemos visto el aprendizaje por refuerzo se considera un tercer paradigma de aprendizaje automático.
Uno de los desafíos mas grandes que existen en el aprendizaje por refuerzo es el intercambio entre **explotación y exploración**, para que el agente pueda encontrar mayor recompensa debe explotar lo que le ha funcionado en el pasado, pero para

descubrir estas acciones debe explorar nuevas acciones. Tomar alguna de las dos acciones exclusivamente lleva al error, por lo cual deben ser balanceadas progresivamente; una acción se debe probar muchas veces para saber cuanto recompensa esperar con mayor confianza, este dilema ha sido estudiado extensivamente sin encontrar una solución definitiva.

Otro aspecto clave del aprendizaje por refuerzo es el enfoque, el cual se centra en el problema de un agente orientado hacia un objetivo en un medio ambiente desconocido. Todos los agentes en aprendizaje por refuerzo tienen objetivos explícitos, pueden sentir los cambios en el ambiente y pueden tomar acciones que influencien al mismo, al tratarse de un enfoque que engloba todo el problema orientado hacia un objetivo no necesariamente significa un organismo completo o un robot.

Ejemplos

Open AI Dexterity

Open AI desarrollo un sistema llamado Dactyl el cual esta completamente entrenado en simulaciones de computadora y transfiere el conocimiento aprendido hacia el mundo real, este sistema aprende desde cero utilizando aprendizaje por refuerzo, el

objetivo de este sistema es demostrar que el entrenamiento por refuerzo en simulaciones puede lograr un gran impacto en la vida real, incluso con lo "ruidoso" que significa esto, por ejemplo con sensores que no responden a tiempo y con datos parciales, igual logra manipular con destreza el objeto y lograr el objetivo propuesto (mostrar una cara específica del cubo).

Demostración de dactyl.

AlphaStar

Los juegos se han usado por décadas como una manera de probar el desempeño de sistemas de inteligencia artificial, como la capacidad de los mismos ha aumentado se ha buscado con el tiempo juegos mucho mas complejos y desafiantes que contengan elementos básico de inteligencia que permitan resolver problemas científicos o de la vida real. Como ya ha ocurrido con otros juegos antes (Atari, Mario, Quake, Dota 2), en una serie de partidas de prueba AlphaStar venció al jugador profesional Grzegorz "MaNa" Komincz por 5–0, esto se dio en condiciones de partidas oficiales entre jugadores profesionales.

AlphaStar fue entrenado inicialmente utilizando aprendizaje supervisado sobre partidas anónimas liberadas por Blizzard, esto permitió que aprendiera micro y macro estrategias de jugadores reales, luego el sistema comenzó a jugar contra si mismo para mejorar sus estrategias por medio de aprendizaje por refuerzo.

AlphaStar vs MaNa

Navegación de Robot

Cada día se encuentran nuevas aplicaciones del aprendizaje por refuerzo, este tipo de aprendizaje resulta siendo útil para mapear entradas de sensores, agarrar objetos y controlar movimientos de robots, es por esto que se ha generado este sistema de prueba el cual puede servir para gente que posee algún impedimento para moverse libremente, estos robots pueden recorrer grandes distancias, traer las compras, medicinas y en general cualquier tipo de paquetes, es el futuro de los repartos y se esta haciendo con aprendizaje por refuerzo.

Robot desplazándose en ambiente con obstáculos.

Elementos del Aprendizaje por refuerzo

A parte del agente y el medio ambiente se puede identificar cuatro sub-elementos de un sistema de aprendizaje por refuerzo: una **política/policy, señal de recompensa/reward signal, función de valor/value function**, y opcionalmente, un **modelo** del medio ambiente.

Una **política** define el modo que el agente se comporta en un momento definido, una política vendría a mapear que acciones se deberían llevar a cabo dado los estados que se han percibido del medio ambiente. Una política puede representar una función simple o una tabla de búsqueda, en la mayoría de casos puede involucrar un calculo extensivo, la política representa el núcleo del agente y es suficiente para determinar el comportamiento de este.

Ejemplo de la política implementada en el paper DeepMimic. (link en las referencias)

La **señal de recompensa** es lo que define el objetivo del problema que se quiere solucionar, cada paso de tiempo, el ambiente retorna un número llamado *recompensa*, el objetivo del agente es maximizar esta recompensa a largo plazo, en este caso la señal de recompensa decide cuales son buenos o malos eventos para el agente. La señal de recompensa termina siendo uno de los recursos mas importantes para la política debido a que si la acción conlleva a una recompensa baja, la política deberá variar para futuros eventos que involucren a esa acción en la misma situación.

Es aquí donde se genera la diferencia con la **función de valor**, ya que la señal de recompensa señala que será lo mejor para corto plazo, pero para largo plazo la encargada es la función de valor.
El *valor* representa a la cantidad de recompensa que puede esperar el agente en el futuro, iniciando desde el estado actual. Es decir la recompensa es lo que se debe buscar, sin recompensa no puede haber valor, y el valor solo sirve para encontrar una mayor recompensa, las acciones se guían en los valores para encontrar una mayor recompensa, por eso se buscan acciones que puedan traer mayor valor, no mayor recompensa, por que dichas acciones son las que obtienen mayor recompensa a largo plazo.

Por ultimo el **modelo** del medio ambiente, este modelo imita el comportamiento del medio ambiente y generalmente ayuda a inferir como el medio ambiente se va a comportar, por ejemplo, dado un estado y una acción el modelo puede predecir el siguiente estado y la siguiente recompensa. Normalmente estos modelos se utilizan para planificar que acciones se llevaran a cabo, los métodos que utilizan modelos para resolver problemas de aprendizaje por refuerzo se llaman **métodos basados en modelos** o **model-based**, por otro lado tambien existe el método llamado **model-free o sin modelo** el cual se basa en prueba y error, lo que es visto como lo contrario a la planificación.

Aparte de estos métodos tambien existen los basados en el **valor/value based, basados en política/policy based, Actor crítico/actor critic** que se encuentra basado en la política y el valor.

Problemas dentro del aprendizaje por refuerzo

Los principales problemas en tomas de decisiones secuenciales se pueden dividir en dos ramas:

Aprendizaje por Refuerzo

- El medio ambiente es desconocido.
- Los agentes interactúan con el medio ambiente.
- Los agentes mejoran su política.

Planificación

- Un modelo del medio ambiente es conocido.

- El agente realiza cálculos con el modelo. (sin ninguna interacción externa)
- El agente mejora su política.
- Búsqueda, deliberación, razonamiento, etc.

Muchas gracias por leer este artículo, en los siguientes post entraremos a conocer mas de este mundo, si deseas ver el siguiente ingresa a Aprendizaje por Refuerzo: Procesos de Decisión de Markov — Parte 1

Referencias

- UCL Course on RL Lecture 1: Introduction to Reinforcement Learning
- Richard Sutton's Intro to Reinforcement Learning
- Long-Range Robotic Navigation via Automated Reinforcement Learning
- Learning Dexterity
- AlphaStar: Mastering the Real-Time Strategy Game StarCraft II
- DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills

Artificial Intelligence Reinforcement Learning Inteligencia Artificial Aprendizaje Por Refuerzo Machine Learning