Notas del curso

Aprendizaje por refuerzo I
Introducción

Richard S. Sutton y Andrew G. Barto

[En él intentamos dar una idea intuitiva básica de qué es el aprendizaje por refuerzo y en qué se diferencia y relaciona con otros campos, por ejemplo, el aprendizaje supervisado y las redes neuronales, los algoritmos genéticos y la vida artificial, la teoría del control. Intuitivamente, la RL es ensayo y error (variación y selección, búsqueda) más aprendizaje (asociación, memoria). Sostenemos que la RL es el único campo que aborda seriamente las características especiales del problema de aprender de la interacción para alcanzar objetivos a largo plazo].

1 Aprender de la interacción

La idea de que aprendemos interactuando con nuestro entorno es probablemente la primera que se nos ocurre cuando pensamos en la naturaleza del aprendizaje. Cuando un bebé juega, agita los brazos o mira a su alrededor, no tiene un maestro explícito, pero sí una conexión sensoriomotora directa con su entorno. El ejercicio de esta conexión produce una gran cantidad de información sobre causa y efecto, sobre las consecuencias de las acciones y sobre qué hacer para alcanzar objetivos. No cabe duda de que esta interacción contribuye en gran medida a que el bebé desarrolle un sentido de su entorno y de su propio papel en él. La experiencia sigue siendo un poderoso maestro a medida que el bebé se convierte en niño y en adulto, aunque la interacción con la naturaleza cambia significativamente con el tiempo. Tanto si estamos aprendiendo a conducir un coche como a mantener una conversación, todos somos muy conscientes de cómo responde nuestro entorno a lo que hacemos, y tratamos de influir en su comportamiento. Aprender de la interacción es una idea fundamental que subyace en casi todas las teorías del aprendizaje.

Aprendizaje por refuerzo: una aproximación computacional al estudio del aprendizaje

de la interacción. En la última década, el estudio del aprendizaje por refuerzo se ha convertido en un campo extraordinariamente multidisciplinar en el que participan investigadores especializados en inteligencia artificial, psicología, ingeniería de control e investigación operativa,

neurociencia, redes neuronales artificiales y algoritmos genéticos. El aprendizaje por refuerzo tiene raíces especialmente ricas en la psicología del aprendizaje animal, de la que toma su nombre. También se han desarrollado aplicaciones impresionantes del aprendizaje por refuerzo. El creciente interés por el aprendizaje por refuerzo se debe en parte al reto de diseñar sistemas inteligentes que deben funcionar en entornos dinámicos del mundo real. Por ejemplo, para que los robots o "agentes" robóticos sean más autónomos (es decir, menos dependientes de condiciones cuidadosamente controladas) se necesitan métodos de toma de decisiones que sean eficaces en presencia de incertidumbre y que puedan cumplir las limitaciones de tiempo. En estas condiciones, el aprendizaje parece esencial para lograr un comportamiento hábil, y es en estas condiciones en las que el aprendizaje por refuerzo puede tener ventajas significativas sobre otros tipos de aprendizaje.

Aquí desarrollamos el aprendizaje por refuerzo desde el punto de vista de la tecnología artificial.

inteligencia artificial (IA) y la ingeniería. (El aprendizaje por refuerzo también se ha desarrollado en relación con la psicología y la neurociencia). Desde esta perspectiva, el aprendizaje por refuerzo corresponde a una formulación matemática particular del problema del aprendizaje a partir de la interacción. Examinamos detenidamente este problema y, a continuación, exploramos algunos de los muchos algoritmos para resolverlo que se han propuesto en varias disciplinas diferentes. Al presentar estos algoritmos desde una única perspectiva y utilizando una notación unificada, intentamos que sea fácil ver cómo se relacionan entre sí los distintos métodos y cómo pueden combinarse de la forma más provechosa. Sorprendentemente, casi todos estos algoritmos pueden entenderse como combinaciones de unos pocos principios subyacentes.

Quizá el resultado más sorprendente de la visión moderna del refuerzo

El aprendizaje por refuerzo se basa en la estrecha relación que existe entre el aprendizaje y la planificación, entendiendo por planificación la decisión sobre un curso de acción teniendo en cuenta posibles situaciones de fu- tura antes de experimentarlas realmente. Los primeros y más sencillos algoritmos de aprendizaje por refuerzo permiten aprender directamente de la interacción con el entorno sin tener que considerar ninguna situación que no se haya vivido realmente. Utilizando este tipo de aprendizaje por refuerzo, un sistema puede lograr un comportamiento altamente cualificado sin tener ninguna capacidad de predecir cómo podría comportarse su entorno en respuesta a sus acciones (es decir, sin tener ningún tipo de modelo de su entorno). Es casi lo contrario de la planificación. Sin embargo, se han ideado formas más complejas de aprendizaje por refuerzo que están estrechamente relacionadas con los métodos computacionales conocidos como programación dinámica, que sí aprovechan los modelos del entorno, y estos métodos están estrechamente relacionados con los métodos de planificación del espacio de estados de la inteligencia artificial. Hoy está claro que el aprendizaje por refuerzo, en una u otra de sus diversas

formas, puede aplicarse a casi cualquier problema de planificación, a veces con ventajas significativas sobre los métodos de planificación más convencionales.

2 Ejemplos

Una buena manera de introducir el aprendizaje por refuerzo es considerar algunos de los ejemplos y posibles aplicaciones que han guiado su desarrollo:

- *Un maestro del ajedrez hace una jugada. La elección se basa tanto en la planificación -anticipando posibles respuestas y contrarréplicas- como en juicios inmediatos e intuitivos sobre la conveniencia de determinadas posiciones y jugadas.*

- *Un controlador adaptativo ajusta en tiempo real los parámetros de funcionamiento de una refinería de petróleo. El controlador optimiza la relación rendimiento/coste/calidad en función de los costes marginales especificados sin ceñirse estrictamente a los valores de consigna sugeridos originalmente por los ingenieros.*

- *Una cría de gacela se levanta con dificultad minutos después de nacer. Media hora después corre a 50 km/h.*

- *Phil se prepara el desayuno. Cuando se examina de cerca, incluso esta actividad aparentemente mundana se revela como una compleja red de comportamientos condicionales y relaciones meta-submeta entrelazadas: Caminar hasta el armario, abrirlo, seleccionar una caja de cereales y, a continuación, alcanzarla, cogerla y recuperarla. Otras secuencias de comportamiento complejas, afinadas e interactivas son necesarias para obtener un cuenco, una cuchara y una jarra de leche. Cada paso implica una serie de movimientos oculares para obtener información y guiar el alcance y la locomoción. Se realizan continuamente juicios rápidos sobre cómo transportar los objetos o si es mejor llevar algunos de ellos a la mesa del comedor antes de obtener otros. Cada paso está guiado por objetivos, como coger una cuchara o llegar hasta el frigorífico, y está al servicio de otros objetivos, como tener la cuchara para comer una vez preparados los cereales y, en última instancia, obtener alimento.*

- *Un robot móvil decide si debe entrar en una nueva habitación en busca de más basura que recoger o empezar a intentar encontrar el camino de vuelta a su estación de recarga de baterías. Su decisión se basa en la rapidez y facilidad con la que ha encontrado el cargador en el pasado.*

Estos ejemplos comparten características tan básicas que es fácil pasarlos por alto. Todos implican la interacción entre un agente activo que toma decisiones y su entorno, en la que el agente intenta alcanzar un objetivo a pesar de la incertidumbre sobre su entorno. Se permite que las acciones del agente afecten al estado futuro del entorno (por ejemplo, la próxima posición del ajedrez, el nivel de reservas de la refinería, la próxima ubicación del robot),

afectando así a las opciones y oportunidades disponibles para el agente en momentos posteriores. Una elección correcta requiere tener en

tienen en cuenta las consecuencias indirectas y diferidas de las acciones, por lo que pueden requerir previsión o planificación.

Al mismo tiempo, los efectos de las acciones no pueden predecirse por completo, por lo que

el agente debe vigilar con frecuencia su entorno y reaccionar adecuadamente. Por ejemplo, Phil debe vigilar la leche que vierte en su tazón de cereales para evitar que rebose. Todos estos ejemplos implican objetivos explícitos, en el sentido de que el agente puede juzgar el progreso hacia su objetivo basándose en lo que puede percibir directamente. El jugador de ajedrez sabe si gana o no, el controlador de la refinería sabe cuánto petróleo se está produciendo, el robot móvil sabe cuándo se agotan sus baterías y Phil sabe si está disfrutando o no de su desayuno. Además, los objetivos de un agente son sus propios objetivos, no los objetivos de un agente externo o de un diseñador. Si queremos utilizar un sistema de aprendizaje por refuerzo para una aplicación de ingeniería, como mejorar el rendimiento de una refinería de petróleo, tenemos que hacer un sistema de aprendizaje por refuerzo cuyos propios objetivos sean los mismos que los nuestros.

En todos estos ejemplos, el agente puede utilizar su experiencia para mejorar su per-

formancia con el paso del tiempo. El jugador de ajedrez refina la intuición que utiliza para evaluar posiciones, mejorando así su juego; la cría de gacela mejora la eficiencia con la que puede correr; Phil aprende a racionalizar la preparación de su desayuno. El nivel de conocimiento que el agente aporta a la tarea al principio -ya sea por su experiencia previa con tareas relacionadas o por su programación genética- influye en lo que es útil o fácil de aprender, pero la interacción con el entorno es esencial para ajustar el comportamiento y explotar las características específicas de cada tarea.

3 Aprendizaje por refuerzo

El aprendizaje por refuerzo es el aprendizaje de una correspondencia entre situaciones y acciones para maximizar una recompensa escalar o señal de refuerzo. El alumno no necesita que se le diga directamente qué acciones debe emprender, como en la mayoría de las formas de aprendizaje automático, sino que debe descubrir qué acciones producen la mayor recompensa probándolas. En los casos más interesantes y desafiantes, una acción puede afectar no sólo a la recompensa inmediata, sino también a la situación siguiente y, en consecuencia, a todas las recompensas subsiguientes. Estas dos características - la búsqueda por ensayo y error y la recompensa diferida- son las dos características distintivas más importantes del aprendizaje por refuerzo.

Todos los algoritmos de aprendizaje por refuerzo requieren un com-

binación de búsqueda y memoria. La búsqueda es necesaria para encontrar

buenas acciones, y la memoria es necesaria para recordar qué acciones funcionaron bien en qué situaciones en el pasado. La sinergia surge porque la búsqueda proporciona la información que hay que recordar y la memoria facilita y acelera la búsqueda. El aprendizaje por refuerzo consiste en almacenar sistemáticamente en la memoria caché los resultados de la búsqueda para que ésta sea más eficaz en el futuro o incluso se elimine.

Aunque la búsqueda y la memoria son elementos computacionales clave de cualquier algoritmo de aprendizaje por refuerzo, es mejor definir el aprendizaje por refuerzo en términos de una clase particular de problemas de aprendizaje, no de algoritmos particulares. Cualquier algoritmo que sea adecuado para resolver uno de estos problemas se considera un algoritmo de aprendizaje por refuerzo. En el capítulo 2 presentamos una formulación precisa de los problemas de aprendizaje por refuerzo, basada en gran medida en la definición matemática de un proceso de decisión de Markov. Aunque esta formulación nos permite aprovechar una gran riqueza de la teoría matemática existente, nuestra intención principal es proporcionar una representación bastante directa del problema real al que se enfrenta un agente de aprendizaje que interactúa con su entorno para alcanzar un objetivo (o alcanzar múltiples objetivos). Es evidente que un agente de este tipo debe ser capaz de percibir información relativa al estado de su entorno y de emprender acciones que afecten a dicho estado. El agente también debe tener un objetivo u objetivos, definidos en términos de cómo se comporta el entorno a lo largo del tiempo bajo la influencia de sus acciones. Estos tres aspectos -sensación, acción y objetivo- son los componentes básicos del marco teórico que utilizamos a lo largo de este libro.

El aprendizaje por refuerzo es muy diferente del aprendizaje supervisado, el tipo de aprendizaje que se estudia en casi toda la investigación actual sobre aprendizaje automático, reconocimiento estadístico de patrones y redes neuronales artificiales. El aprendizaje supervisado consiste en aprender bajo la tutela de un supervisor, o "maestro", que indica explícitamente al agente de aprendizaje cómo debe responder a las entradas de entrenamiento. Aunque este tipo de aprendizaje puede ser un componente importante de sistemas de aprendizaje más completos, no es adecuado por sí mismo para el tipo de aprendizaje que deben realizar los agentes autónomos. A menudo es muy costoso, o incluso imposible, obtener un conjunto de ejemplos del comportamiento deseado que sea a la vez correcto y representativo de las situaciones en las que el agente tendrá que actuar. En territorio inexplorado - donde cabría esperar que el aprendizaje fuera más beneficioso-, un agente debe ser capaz de aprender de sus propias experiencias en lugar de aprender de un maestro experto. Aunque un agente de aprendizaje por refuerzo también puede aprovechar la información de un profesor experto, la principal fuente de información y retroalimentación es la interacción con su entorno.

Uno de los retos que se plantea en el aprendizaje por refuerzo, y no en otros de aprendizaje, se ha denominado el compromiso entre exploración y explotación. Para obtener muchas recompensas, un agente de aprendizaje por refuerzo debe preferir acciones que haya probado en el pasado y que le hayan resultado eficaces para obtener recompensas. Pero para descubrir qué acciones son éstas, tiene que seleccionar acciones que no ha probado antes. El agente tiene que explotar lo que ya sabe para obtener re ward, mientras que también tiene que explorar para hacer mejores selecciones de acciones en el futuro. El

dilema es que ni la explotación ni la exploración pueden realizarse de forma exclusiva sin fracasar en la tarea. El agente debe probar diversas acciones y favorecer progresivamente las que le parezcan mejores. En una tarea estocástica, cada acción debe probarse muchas veces para estimar de forma fiable su rendimiento esperado.

recompensa. El dilema exploración-explotación ha sido estudiado intensamente por los matemáticos durante muchas décadas. Nos limitaremos a señalar que la cuestión de equilibrar explotación y exploración ni siquiera se plantea en el aprendizaje supervisado, tal y como suele definirse.

Otra característica clave del aprendizaje por refuerzo es que considera explícitamente

el problema global de un agente dirigido por un objetivo que interactúa con un entorno incierto. Esto contrasta con muchos enfoques que tratan subproblemas sin abordar cómo podrían encajar en un panorama más amplio. Por ejemplo, hemos mencionado que gran parte de la investigación sobre aprendizaje automático se ocupa del aprendizaje su- pervisado sin especificar explícitamente cómo sería útil finalmente esa capacidad. Otros investigadores han desarrollado teorías de la planificación con objetivos generales, pero sin tener en cuenta el papel de la planificación en la toma de decisiones en tiempo real, o la cuestión de dónde provendrían los modelos predictivos necesarios para la planificación. Aunque estos enfoques han dado muchos resultados útiles, es evidente que su concentración en subproblemas aislados se ha convertido en una limitación importante.

El aprendizaje por refuerzo adopta la táctica opuesta al comenzar con un completo,

agente interactivo que busca objetivos. Todos los agentes de aprendizaje por refuerzo tienen objetivos explícitos, pueden percibir aspectos de su entorno y elegir acciones para influir en él. Además, normalmente se asume desde el principio que el agente tendrá que operar a pesar de una incertidumbre significativa sobre el entorno al que se enfrenta. Cuando el aprendizaje por refuerzo implica planificación, hay que abordar la interacción entre la planificación y la selección de acciones en tiempo real, así como la cuestión de cómo se adquieren y mejoran los modelos del entorno. Cuando el aprendizaje por refuerzo implica aprendizaje supervisado, lo hace por razones muy específicas que determinan qué capacidades son críticas y cuáles no. Para que la investigación sobre el aprendizaje progrese, sin duda hay que aislar y estudiar subproblemas importantes, pero deben ser subproblemas motivados por funciones claras en agentes completos, interactivos y que buscan objetivos, aunque todavía no se puedan completar todos los detalles del agente completo.

4 Componentes de un agente de aprendizaje por refuerzo

Un agente de aprendizaje por refuerzo suele constar de cuatro componentes básicos: una política, una función de recompensa, una función de valor y un modelo del entorno.

La política es la función de toma de decisiones del agente, que especifica qué

acción emprende en cada una de las situaciones que se le pueden presentar. En psicología, correspondería al conjunto de reglas o asociaciones estímulo-respuesta. Este es el núcleo de un agente de refuerzo, tal y como sugiere la Figura 1, porque por sí solo basta para definir un agente completo que se comporta. Los demás componentes sólo sirven para cambiar y mejorar la política. La política en sí es el determinante último del comportamiento y el rendimiento. En general, puede ser estocástica.

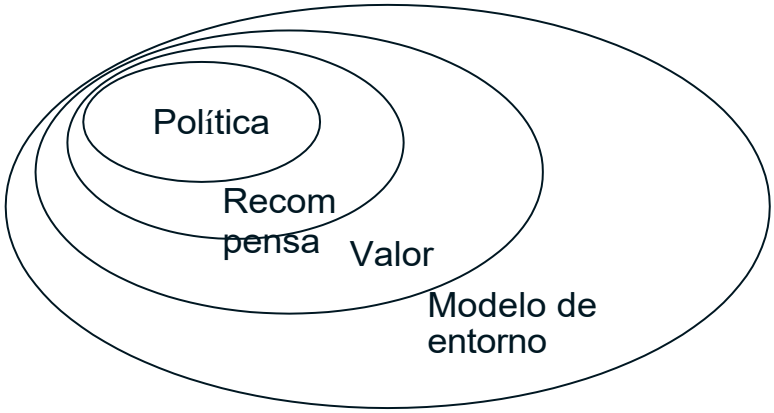

Figura 1: Componentes principales de un agente de aprendizaje por refuerzo.

La función de recompensa define el objetivo del agente de aprendizaje por refuerzo. El objetivo del agente es maximizar la recompensa que recibe a largo plazo. Por tanto, la función de recompensa define cuáles son los sucesos buenos y malos para el agente. Las recompensas son las características inmediatas y definitorias del problema al que se enfrenta el agente. Como tal, la función de recompensa debe ser necesariamente fija. Sin embargo, puede servir de base para modificar la política. Por ejemplo, si una acción seleccionada por la política va seguida de una recompensa baja, entonces la política puede cambiarse para seleccionar alguna otra acción en esa situación en el futuro.

Mientras que la recompensa indica lo que es bueno en un sentido inmediato, el valor especifica lo que es bueno a largo plazo, es decir, porque predice la recompensa. La diferencia entre valor y recompensa es fundamental para el aprendizaje por refuerzo. Por ejemplo, al jugar al ajedrez, dar jaque mate al adversario se asocia a una recompensa alta, pero ganar su reina se asocia a un valor alto. La primera define el verdadero objetivo de la tarea (ganar la partida), mientras que la segunda sólo predice ese verdadero objetivo. Aprender el valor de los estados, o de los pares estado-acción, es el paso crítico en los métodos de aprendizaje por refuerzo que consideramos aquí.

El cuarto y último componente principal de un agente de aprendizaje por refuerzo es un

modelo de su entorno o mundo exterior. Se trata de algo que imita el comportamiento del entorno en cierto sentido. Por ejemplo, dada una situación y una acción, el modelo puede predecir el siguiente estado resultante y la siguiente recompensa. En la Figura 1, el modelo es el componente de mayor tamaño porque es de esperar que ocupe el mayor espacio de almacenamiento. Si hay jSj estados y jAj acciones, un modelo completo ocupará un espacio proporcional al tamaño de jSj × jSj × jAj, porque asigna pares estado-acción a distribuciones de probabilidad sobre los estados, dando la probabilidad de cada posible estado resultante para cada acción realizada en cada estado. Por el contrario, las funciones de recompensa y de valor podrían simplemente asignar

estados a números reales y, por tanto, tener un tamaño jSj, mientras que una política estocástica tiene como máximo un tamaño jSj $\times$ *jAj.*

No todos los agentes de aprendizaje por refuerzo utilizan modelos del entorno. Met-

os métodos que nunca aprenden ni utilizan un modelo se denominan métodos de aprendizaje por refuerzo sin modelo. Los métodos sin modelo son muy sencillos y, quizá sorprendentemente, suelen ser capaces de encontrar el comportamiento óptimo. Los métodos basados en modelos simplemente lo encuentran más rápido (con menos experiencias). El caso más interesante es aquel en el que el agente no tiene un modelo perfecto del entorno a priori, sino que debe utilizar métodos de aprendizaje para alinearlo con la realidad.

5 Resumen

En este capítulo hemos esbozado algunas de las razones por las que cada vez más buscadores prestan atención al aprendizaje por refuerzo. En primer lugar, el aprendizaje por refuerzo se centra en el aprendizaje en línea durante la interacción normal con un entorno dinámico. Esto contrasta con el enfoque de gran parte del aprendizaje automático, tanto simbólico como de redes neuronales artificiales, en sistemas que aprenden offiine a partir de un conjunto preespecificado de ejemplos de entrenamiento proporcionados por un "profesor" explícito y bien informado. Aunque un sistema de aprendizaje por refuerzo también debería ser capaz de aprovechar los conocimientos de los profesores de su entorno, si los hay, su verdadera fuente de información es su propia experiencia. Además, la mayoría de los sistemas de aprendizaje automático no aprenden mientras se utilizan. Es más apropiado llamarlos sistemas aprendidos que sistemas de aprendizaje. Aunque el aprendizaje por refuerzo se utiliza a veces de este modo, la base conceptual del aprendizaje por refuerzo es que un sistema de aprendizaje debe utilizar toda su experiencia, a lo largo de toda su existencia, para mejorar su rendimiento.

El aprendizaje por refuerzo utiliza un marco formal que define la interacción entre el aprendizaje y el refuerzo.

entre el agente y el entorno en términos de situaciones (estados), acciones y recompensas. Este marco pretende ser una forma sencilla de representar las características esenciales del problema de la IA. Estas características incluyen un sentido de causa y efecto, de incertidumbre y no determinismo, y la existencia de objetivos explícitos. Lo más relevante es el formalismo de los procesos de decisión de Markov, que proporciona una forma precisa y relativamente neutral de incluir las características clave. Aunque sólo arañamos su superficie, esta teoría nos permite aprovechar perspectivas y métodos afines desarrollados en el campo del control óptimo estocástico.

Los conceptos de valor y funciones de valor son las características clave de los tipos de

métodos de aprendizaje por refuerzo que consideramos en este libro. Adoptamos la postura de que las funciones de valor son esenciales para una búsqueda eficiente en el espacio de políticas. El uso de funciones de valor distingue a estos

métodos de aprendizaje por refuerzo de los métodos evolutivos conceptualmente más sencillos que buscan directamente en el espacio de políticas, guiados por evaluaciones escalares de políticas enteras. En nuestro enfoque, las funciones de valor permiten a los algoritmos aprovechar los detalles de las interacciones conductuales individuales. Aunque los métodos evolutivos pueden proporcionar resultados útiles para algunos problemas, y los métodos de función de valor pueden utilizarse provechosamente en conjun-

En nuestra opinión, cuando se aplican a problemas de aprendizaje por refuerzo, los métodos evolutivos son intrínsecamente menos eficaces que los métodos de función de valor.

Una vez que se toma la estimación (aprendizaje) de las funciones de valor como un paso computacional clave, la cuestión pasa a ser cuál es la mejor manera de hacerlo. En este libro identificamos tres clases principales de métodos. Los métodos de Monte Carlo estiman el valor de un estado simplemente realizando muchas pruebas a partir de ese estado. A continuación, se promedian las recompensas totales reales recibidas en esos ensayos para obtener una estimación del valor del estado. Los métodos de búsqueda, como la programación dinámica y la búsqueda heurística, pueden considerarse formas sencillas de estimar una función de valor basadas en modelos. Por último, los métodos de diferencia temporal, un desarrollo relativamente nuevo, se basan en el aprendizaje de los valores de los estados utilizando los valores de los estados que les siguen en los ensayos reales.

Este libro se articula en torno al principio de que estas tres clases de me
os métodos de aprendizaje de funciones de valor no son totalmente diferentes: pueden considerarse miembros de una "superfamilia" de métodos. Aunque existen diferencias entre ellos, no es necesario elegir, sino que se pueden mezclar y combinar. Todos tienen en común una operación llamada "copia de seguridad". Algunos realizan copias de seguridad basándose en la experiencia, otros en un modelo, otros a partir de un amplio conjunto de posibles estados siguientes, otros sólo a partir de uno. Las copias de seguridad son de distintos tamaños, formas y fuentes, pero todas comparten características comunes y contribuyen a un cálculo común.

6 Observaciones bibliográficas e históricas

A continuación presentamos un análisis necesariamente abreviado de la historia de las principales ideas del aprendizaje por refuerzo. Aunque el término específico "aprendizaje por refuerzo" nunca ha sido utilizado por los psicólogos, las raíces del aprendizaje por refuerzo se encuentran en las teorías del aprendizaje desarrolladas por los psicólogos experimentales a lo largo de este siglo. Nos llevaría demasiado lejos intentar una visión general de las teorías del refuerzo de la psicología, algo que ya está disponible en muchos libros (por ejemplo, (Mackintosh, 1983)). En su lugar, nos centraremos en las primeras exploraciones más conocidas del poder computacional del aprendizaje por refuerzo, intentando no ocultar el hecho de que las perspectivas computacionales y psicológicas son a veces difíciles de distinguir.

En la década de 1960 se encuentran los términos "refuerzo" y "aprendizaje por refuerzo".

ing" por primera vez en la literatura de ingeniería (tesis de Minsky? (Min- sky,

1961); Waltz & Fu (Waltz & Fu, 1965), Mendel, 1966; Mendel & McClaren (Mendel & McLaren, 1970)). Estos términos se utilizan para referirse a la idea general de aprendizaje a partir de recompensas y castigos: aprendizaje por ensayo y error, en el que las acciones seguidas de buenos o malos resultados se refuerzan o debilitan respectivamente.

Esta noción temprana del aprendizaje por refuerzo es un reflejo exacto de un principio psicológico clásico, la "Ley del Efecto" de Thorndike (1911):

> *"De varias respuestas dadas a la misma situación, las que van acompañadas o seguidas de satisfacción para el animal estarán, en igualdad de condiciones, más firmemente conectadas con la situación, de modo que, cuando se repita, será más probable que se repitan; las que van acompañadas o seguidas de malestar para el animal tendrán, en igualdad de condiciones, sus conexiones con esa situación debilitadas, de modo que, cuando se repita, será menos probable que se produzcan. Cuanto mayor sea la satisfacción o el malestar, mayor será el fortalecimiento o el debilitamiento del vínculo".*

Aunque este principio ha generado una considerable controversia en psicología, así como en otros campos, a lo largo de los años (véase la referencia ???), sigue siendo influyente porque su idea general está respaldada por muchos experimentos y tiene un gran sentido intuitivo. Es una forma elemental y obvia de combinar búsqueda y memoria: búsqueda en forma de probar muchas acciones, y memoria en forma de recordar qué acciones funcionaron mejor. Dennett (-) ofrece una buena explicación del atractivo continuo de la Ley del Efecto, y Cziko (-) ofrece una explicación muy amplia de la utilidad de los métodos, como la Ley del Efecto, que operan utilizando principios de selección, en contraposición a los de instrucción.

Las primeras investigaciones computacionales de la Ley del Efecto que conocemos

de Minsky y de Farley y Clark, ambos publicados en 1954. En su tesis doctoral, Minsky (Minsky, 1954) describe la construcción de una máquina analógica, la SNARC (Stochastic Neural-Analog Reinforcement Calculator), diseñada para aprender por ensayo y error. Farley y Clark ((Farley & Clark, 1954); Clark y Farley, 1955) describen otra máquina de aprendizaje por redes neuronales, pero señalando su capacidad de "generalizar", la discuten más en términos de aprendizaje supervisado que de aprendizaje por refuerzo. Esto inició un patrón de confusión sobre la relación entre estos tipos de aprendizaje. Muchos investigadores parecían creer que estaban estudiando el aprendizaje por refuerzo, cuando en realidad estaban estudiando el aprendizaje supervisado, una confusión que persiste hoy en día. Incluso los libros de texto modernos sobre redes neuronales suelen describir las redes que aprenden a partir de ejemplos de entrenamiento como sistemas de aprendizaje por ensayo y error, ya que utilizan la información de error para actualizar los pesos de las conexiones. Aunque se trata de una confusión comprensible, en realidad pasa por alto el carácter selectivo del aprendizaje a través de la Ley del Efecto, que es lo que el término ensayo-error pretendía describir originalmente.

Está claro que pioneros de las redes neuronales como Rosenblatt (1958,

(Rosenblatt,
1961)) y Widrow y Hoff (Widrow & Hoff, 1960), así como los psicólogos Bush y Mosteller (Bush & Mosteller, 1955), pensaban en el aprendizaje por refuerzo -utilizaban el lenguaje de recompensas y castigos-, pero los sistemas

que estudiaban se convirtieron más claramente en sistemas de aprendizaje supervisado, adecuados para el reconocimiento de patrones y el aprendizaje perceptivo, pero no para la interacción directa con un entorno. En los años sesenta y setenta, el aprendizaje por refuerzo fue quedando gradualmente relegado a un segundo plano y se perdió como tema diferenciado, mientras que el aprendizaje supervisado, sobre todo en forma de reconocimiento de patrones, pasó a ser ampliamente estudiado. En el capítulo 3 se analizan algunos de los detalles de esta transición, incluidas excepciones como la teoría de los autómatas de aprendizaje y los métodos de aproximación estocástica de Kiefer-Wolfowitz. *3.*

Cabe mencionar aquí otras claras excepciones a esta tendencia. En 1963, Andreae

describió una máquina de aprendizaje por refuerzo llamada STeLLA (-), que incluía lo que ahora llamaríamos un modelo de entorno para facilitar el aprendizaje. Andreae se interesaba explícitamente por cómo una máquina podía aprender interactuando con su entorno. Desarrollos posteriores incluyeron un "monólogo interno" para tratar el problema de la observabilidad parcial del estado (-), algo que sigue siendo importante para el aprendizaje por refuerzo. Aunque los trabajos posteriores de Andreae siguieron insistiendo en el aprendizaje por interacción, hicieron más hincapié en el papel de un profesor (-). La investigación pionera de Andreae no es muy conocida, pero sigue aportando enseñanzas a la investigación moderna sobre el aprendizaje por refuerzo.

También en la década de 1960, Donald Michie mantuvo un claro enfoque en el refuerzo

aprendizaje. Describió un sencillo sistema de aprendizaje por refuerzo para aprender a jugar al tres en raya (también conocido como tres en raya) llamado MENACE (Matchbox Educable Noughts and Crosses Engine) (-; -). Consistía en una caja de cerillas para cada posible posición de juego que contenía un número de cuentas de colores, un color para cada movimiento disponible desde esa posición. Sacando una cuenta al azar de la caja de cerillas correspondiente a la posición de juego actual, se podía determinar el movimiento de MENACE. Cuando terminaba una partida, se añadían o retiraban cuentas de las cajas utilizadas durante el juego para reforzar o castigar las decisiones de MENACE. Ahora consideraríamos a MENACE como una colección de autómatas de aprendizaje estocástico simple (Capítulo 3). En 1968, Michie y Chambers (Michie & Chambers, 1968) describieron un aprendiz de refuerzo Tic-Tac-Toe más avanzado llamado GLEE (Game Learning Expectimaxing Engine) que estimaba una función de valor utilizando lo que llamaron Expectimaxing. Ahora lo reconoceríamos como algo estrechamente relacionado con la programación dinámica. Los jugadores de Tic-Tac-Toe de Michie sirvieron de inspiración para nuestro ejemplo de Tic-Tac-Toe de este capítulo, y su discusión sobre cómo descomponer un gran problema en una serie de subproblemas mutuamente

independientes puede conducir a un aprendizaje de refuerzo eficiente influyó en nuestra discusión en la que contrastamos los métodos de función de valor y los métodos evolutivos. (Estos últimos métodos no descomponen los problemas de este modo).

Michie y Chambers (Michie & Chambers, 1968) también describieron una forma más versión avanzada del enfoque MENACE, implementada en un sistema llamado BOXES, que aplicaron al problema de aprender a equilibrar un poste articulado a un carro móvil basándose en una señal de fallo que sólo se producía cuando el poste se caía o el carro llegaba al final de una pista. Este trabajo se inspiró parcialmente en

el sistema de equilibrio de postes de Widrow y Smith (Widrow & Smith, 1964), que aprendió mediante aprendizaje supervisado de un profesor que ya era capaz de realizar la tarea. (Comparar el sistema de equilibrio de polos de Widrow y Smith con el de Michie y Chambers es una buena manera de apreciar la diferencia entre aprendizaje supervisado y aprendizaje por refuerzo). BOXES, que no estimaba una función de valor, inspiró el sistema de equilibrio de polos de Barto, Sutton y An-derson (Barto et al., 1983), que sí estimaba una función de valor. (A estos sistemas les han seguido otros sistemas de aprendizaje por refuerzo con equilibrio de polos demasiado numerosos para mencionarlos).

Aunque Widrow y sus colegas mantuvieron un claro énfasis en la supervisión En 1973, Widrow, Gupta y Maitra (Widrow et al., 1973) modificaron la regla de aprendizaje supervisado de Widrow-Hoff. En 1973, Widrow, Gupta y Maitra (Widrow et al., 1973) modificaron la regla de aprendizaje supervisado de Widrow-Hoff (a menudo llamada regla del mínimo cuadrado medio o LMS) para producir una regla de aprendizaje por refuerzo que pudiera aprender de las señales de éxito y fracaso en lugar de a partir de ejemplos de entrenamiento. Llamaron a esta forma de aprendizaje "aprendizaje se- lectivo bootstrap" y utilizaron la expresión "aprender con un crítico" en lugar de "aprender con un maestro".

Otro investigador que se opuso a la tendencia del aprendizaje supervisado fue Harry Klopf

(Klopf, 1972; Klopf, 1982), que presentó una teoría "hedónica", o "heteroestática", de la función neuronal y la IA. Klopf reconoció que se estaban perdiendo aspectos esenciales del comportamiento adaptativo a medida que los investigadores del aprendizaje se centraban casi exclusivamente en el aprendizaje supervisado. Se trataba de los aspectos hedónicos: el impulso de obtener algún resultado del entorno, de controlarlo hacia fines deseados y alejarlo de fines no deseados. Se trata, por supuesto, de un elemento esencial del aprendizaje por refuerzo. El trabajo de Klopf fue especialmente importante para los autores porque nuestra evaluación de las ideas de Klopf (Barto y Sutton, 1981) nos llevó a apreciar la distinción entre aprendizaje supervisado y aprendizaje por refuerzo, así como a centrarnos finalmente en el aprendizaje por refuerzo.

También hicieron importantes contribuciones al aprendizaje por refuerzo John Holland (Holland, 1975; Holland, 1986). Aunque es más conocido por su desarrollo de los algoritmos genéticos, Holland esbozó una teoría muy general de los sistemas adaptativos que hace hincapié en el aprendizaje interactivo basado en principios de selección. De hecho, su sistema clasificador (Holland, 1986) es un sistema de aprendizaje por refuerzo que actualiza las funciones de valor utilizando lo que él llamó el "algoritmo de la brigada del cubo", que está estrechamente relacionado con los métodos de estimación de funciones de valor que tratamos en este libro. Aunque los algoritmos genéticos son candidatos naturales para implementar lo que llamamos el enfoque evolutivo del

aprendizaje por refuerzo, que contrastamos con los métodos de funciones de valor, Holland no sugirió este enfoque. Los sistemas clasificadores utilizan tanto algoritmos genéticos como funciones de valor.

Aunque la idea de aprender estimando funciones de valor a partir de la experiencia

apareció en la disertación de Minsky (Minsky, 1954), fue introducida de forma más influyente por Samuel (Samuel, 1959) en su programa para aprender a jugar al juego

de damas utilizando lo que ahora llamaríamos un método de diferencia temporal. Consideramos que el trabajo de Samuel es una influencia fundamental en el enfoque del aprendizaje por refuerzo que presentamos en este libro. Sus trabajos (Samuel, 1959; Samuel, 1967) revelan una visión extraordinaria de casi todas las cuestiones que todavía desafían a los investigadores actuales. Estos trabajos son una lectura que merece la pena incluso hoy en día, y tenemos mucho más que decir sobre el jugador de damas de Samuel en capítulos posteriores. Aunque la disertación de Minsky fue una de las primeras incursiones en el aprendizaje por refuerzo, mucho más influyente fue su artículo de 1960 "Steps Toward Artificial Intelligence" (Pasos hacia la inteligencia artificial), que, al igual que los artículos de Samuel, contiene debates convincentes sobre cuestiones que siguen siendo relevantes para el aprendizaje por refuerzo moderno. Cabe destacar la discusión de Minsky sobre lo que él consideraba el principal problema computacional que tendrían que resolver los sistemas complejos de aprendizaje por refuerzo para tener éxito. Lo llamó este el problema de la asignación de créditos:

> *Al aplicar estos métodos a problemas complejos, nos encontramos con una gran dificultad: distribuir el mérito del éxito de una estrategia compleja entre las muchas decisiones que han intervenido.*

Si hay que reforzar las numerosas decisiones que intervienen en la consecución del éxito, hay que evaluar su contribución relativa a ese logro. Minsky analizó el método de estimación de la función de valor utilizado en el juego de damas de Samuel como un enfoque importante de este problema, señalando que está estrechamente relacionado con el fenómeno del refuerzo condicionado que se produce en el aprendizaje animal. Todos los métodos que discutimos en este libro están dirigidos a hacer que la asignación de créditos sea menos problemática para los sistemas de aprendizaje por refuerzo.

De la breve exposición anterior se desprende claramente que las ideas principales del refuerzo

han estado presentes en la IA desde sus inicios. Sin embargo, no ha sido hasta hace relativamente poco que han atraído una atención generalizada. Una de las razones por las que las ideas de refuerzo han tenido históricamente poca influencia en la IA es su asociación con las visiones conductistas del aprendizaje y la inteligencia. En la década de 1960, la investigación en IA siguió los pasos de la psicología, que pasó de enfoques basados en el comportamiento animal a enfoques más cognitivos, dejando poco espacio para las teorías del refuerzo; de hecho, dejando poco espacio para las teorías del aprendizaje de cualquier tipo. Aunque estamos de acuerdo con los críticos que han argumentado que no se puede entender o generar todo el comportamiento inteligente basándose únicamente en los principios de refuerzo, creemos que los sistemas de IA y las teorías cognitivas que se alejan de estos principios de aprendizaje básico también se ven perjudicados. De hecho, el clima se ha caldeado

considerablemente hacia los principios clásicos del aprendizaje, incluido el aprendizaje por refuerzo, porque los investigadores los están utilizando en sistemas que deben tanto a las perspectivas cognitivas como a las teorías anteriores del comportamiento animal.

Un factor relacionado que limitó la influencia de los principios del aprendizaje por refuerzo

en la IA es la reputación de que son demasiado débiles desde el punto de vista computacional para ser de gran utilidad. Sin embargo, en la actualidad existen numerosas pruebas de que el aprendizaje por refuerzo puede ser

muy potente. Algunos de los logros más impresionantes de los sistemas de aprendizaje artificial se han conseguido utilizando el aprendizaje por refuerzo.

Se necesita más información sobre historia y bibliografía: Watkins, Campbell y Craik y Dennett (modelos), Witten , Werbos , nuestro propio material inicial. Búsqueda Heurística, Booker, Hampson. Bellman y Howard. También añadir la cita de Minsky y Selfridge (1963): "en una situación novedosa, prueba métodos como los que han funcionado mejor en situaciones similares"

Referencias

Barto, AG y Sutton, RS (1981). Goal seeking components for adaptive intelligence: An initial assessment. Technical Report AFWAL-TR-81-1070 Air Force Wright Aeronautical Laboratories/Avionics Laboratory Wright-Patterson AFB, OH.

Barto, AG, Sutton, RS, & Anderson, CW (1983). Neuronlike elements that can solve difficult learning control problems. IEEE Transactions on Sys- tems, Man, and Cybernetics, 13, 835-846. Reimpreso en J. A. Anderson y E. Rosenfeld, Neurocomputing: Foundations of Research, MIT Press, Cambridge, MA, 1988.

Bush, RR y Mosteller, F (1955). Stochastic Models for Learning. New York: Wiley.

Farley, BG & Clark, WA (1954). Simulación de sistemas autoorganizados por ordenador digital. IRE Transactions on Information Theory, 4, 76-84.

Holland, JH (1975). Adaptation in Natural and Artificial Systems. Ann Arbor: University of Michigan Press.

Holland, JH (1986). Escaping brittleness: La posibilidad de algoritmos de aprendizaje de propósito general aplicados a sistemas basados en reglas. En: Machine Learning: An Artificial Intelligence Approach, Volume II, (RS Michalski, JG Carbonell, & TM Mitchell, eds) pp. 593-623. San Mateo, CA. San Mateo, CA: Morgan Kaufmann.

Klopf, AH (1972). Brain function and adaptive systems-A heterostatic theory. Technical Report AFCRL-72-0164 Air Force Cambridge Research Laborato- ries Bedford, MA.

Klopf, AH (1982). La neurona hedonista: una teoría de la memoria, el aprendizaje y la inteligencia. Washington, D.C.: Hemishere.

Mackintosh, NJ (1983). Condicionamiento y aprendizaje asociativo. New York: Ox- ford University Press.

Mendel, JM & McLaren, RW (1970). Reinforcement learning control and pattern recognition systems. En: Adaptive, Learning and Pattern Recognition Systems: Theory and Applications, (JM Mendel & KS Fu, eds.) pp. 287-318. New York: J.M. & McLaren, RW (1970). New York: Academic Press.

Michie, D & Chambers, RA (1968). BOXES: Un experimento de control adaptativo. En: Machine Intelligence 2, (E Dale & D Michie, eds) pp. 137-152. Oliver y Boyd.

Minsky, ML (1954). Theory of Neural-Analog Reinforcement Systems and its Application to the Brain-Model Problem. Tesis doctoral Universidad de Princeton.

Minsky, ML (1961). Steps toward artificial intelligence. Proceedings of the Institute of Radio Engineers, 49, 8-30. Reimpreso en E. A. Feigenbaum y otros. Reimpreso en E. A. Feigenbaum y
J. Feldman, editores, Computers and Thought. McGraw-Hill, Nueva York, 406-450, 1963.

Rosenblatt, F (1961). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. 6411 Chillum Place N.W., Washington, D.C.: Spartan Books.

Samuel, AL (1959). Some studies in machine learning using the game of checkers. IBM Journal on Research and Development, 210-229. Reimpreso en E. A. Feigenbaum y J. Feldman, editores, Computers and Thought, McGraw-Hill, Nueva York, 1963.

Samuel, AL (1967). Algunos estudios en aprendizaje de máquinas usando el juego de damas. II-Progresos recientes. Revista IBM de Investigación y Desarrollo, 601-617.

Waltz, MD & Fu, KS (1965). A heuristic approach to reinforcment learning control systems. IEEE Transactions on Automatic Control, 10, 390-398.

Widrow, B, Gupta, NK, & Maitra, S (1973). Punish/reward: Learning with a critic in adaptive threshold systems. IEEE Transactions on Systems, Man, and Cybernetics, 5, 455-465.

Widrow, B & Hoff, ME (1960). Adaptive switching circuits. En: 1960 WESCON Convention Record Part IV pp. 96-104,.

Widrow, B & Smith, FW (1964). Pattern-recognizing control systems. En: Computer and Information Sciences (COINS) Proceedings , Washington, D.C.: Spartan.