**Aprendizaje por Refuerzo**

La historia del aprendizaje de refuerzo tiene dos hilos principales, ambos largos y ricos, que fueron perseguidos independientemente antes de entrelazarse en el aprendizaje de refuerzo moderno. Uno de los hilos se refiere al aprendizaje por ensayo y error que comenzó en la psicología del aprendizaje animal. Este hilo recorre algunos de los primeros trabajos sobre inteligencia artificial y condujo al resurgimiento del aprendizaje de refuerzo a principios de los años 80. El otro hilo se refiere al problema del control óptimo y su solución usando funciones de valor y programación dinámica. En su mayor parte, este hilo no implicaba aprendizaje. Aunque los dos hilos han sido en gran medida independientes, las excepciones giran en torno a un tercer hilo, menos distintivo, relativo a los métodos de diferenciación temporal, como los utilizados en el ejemplo del tres en raya de este capítulo. Los tres hilos se unieron a finales de la década de 1980 para producir el moderno campo de aprendizaje de refuerzo como lo presentamos en este libro.

El hilo que se centra en el aprendizaje por ensayo y error es con el que estamos más familiarizados y sobre el que más tenemos que decir en esta breve historia. Antes de hacerlo, sin embargo, discutimos brevemente el hilo de control óptimo.

El término "control óptimo" se utilizó a finales de los años 50 para describir el problema de diseñar un controlador para minimizar una medida del comportamiento de un sistema dinámico a lo largo del tiempo. Una de las aproximaciones a este problema fue desarrollada a mediados de los 50 por Richard Bellman y otros a través de la extensión de una teoría del siglo XIX de Hamilton y Jacobi. Este enfoque utiliza los conceptos de estado de un sistema dinámico y de una función de valor, o "función de retorno óptima", para definir una ecuación funcional, ahora a menudo llamada la ecuación de Bellman. La clase de métodos para resolver problemas de control óptimo mediante la resolución de esta ecuación llegó a conocerse como programación dinámica (Bellman, 1957a). Bellman (1957b) también introdujo la versión estocástica discreta del problema de control óptimo conocido como procesos de decisión Markovianos (MDPs), y Ronald Howard (1960) ideó el método de iteración de políticas para los MDPs. Todos estos son elementos esenciales que subyacen a la teoría y los algoritmos del aprendizaje de refuerzo moderno.

La programación dinámica se considera la única forma viable de resolver los problemas de control óptimo estocástico general. Sufre lo que el Hombre de las campanas llamó "la maldición de la dimensionalidad", lo que significa que sus requisitos de cálculo crecen exponencialmente con el número de variables de estado, pero sigue siendo mucho más eficiente y más ampliamente aplicable que cualquier otro método general. La programación dinámica se ha desarrollado ampliamente desde finales del decenio de 1950, incluidas las extensiones a los MDP parcialmente observables (estudiados por Lovejoy, 1991), muchas aplicaciones (estudiados por White, 1985, 1988, 1993), los métodos de aproximación ap- (estudiados por Rust, 1996) y los métodos asíncronos (Bertsekas, 1982, 1983). Se dispone de muchos tratamientos modernos excelentes de programación dinámica (por ejemplo, Bertsekas, 1995; Puterman, 1994; Ross, 1983; y Whittle, 1982, 1983). Bryson (1996) ofrece una historia autorizada de control óptimo.El campo ha recorrido un largo camino desde entonces, evolucionando y madurando en varias direcciones. El aprendizaje de refuerzo se ha convertido gradualmente en una de las áreas de investigación más activas en el aprendizaje de las máquinas, la inteligencia artificial y la investigación del trabajo en red neural. El campo ha desarrollado fuertes fundamentos matemáticos y aplicaciones impresionantes. El estudio computacional del aprendizaje de refuerzo es ahora un gran campo, con cientos de investigadores activos en todo el mundo en disciplinas diversas como la psicología, la teoría de control, la inteligencia artificial y la neurociencia. Particularmente importantes han sido las contribuciones que han establecido y desarrollado las relaciones con la teoría del control óptimo y la programación dinámica. El problema general de aprender de la interacción, está todavía lejos de ser resuelto, pero nuestra comprensión de ello ha mejorado significativamente. Ahora podemos situar las ideas componentes, como el aprendizaje de la diferencia temporal, la programación dinámica y la aproximación de funciones, dentro de una perspectiva coherente con respecto al problema general.

Desde el departamento de I+D de Vates, mostramos nuestro interés al escribir este artículo proporcionando un relato claro y simple de las ideas clave y algoritmos de aprendizaje de refuerzo. Queríamos que nuestro tratamiento fuera accesible a los lectores de todas las disciplinas relacionadas, pero no podíamos cubrir todas estas perspectivas en detalle. En su mayor parte, nuestro tratamiento toma el punto de vista de la inteligencia artificial y la ingeniería.

**El problema de aprendizaje de refuerzo**

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 sus brazos o mira a su alrededor, no tiene un maestro explícito, pero tiene una conexión sensorial-motora directa con su entorno. El ejercicio de esta conexión produce una gran cantidad de información sobre la causa y el efecto, sobre las consecuencias de las acciones y sobre qué hacer para alcanzar los objetivos. A lo largo de nuestras vidas, tales interacciones son sin duda una fuente importante de conocimiento sobre nuestro entorno y sobre nosotros mismos. Ya sea que estemos aprendiendo a conducir un auto o a mantener una conversación, somos muy conscientes de cómo nuestro medio ambiente responde a lo que hacemos, y buscamos influenciar lo que sucede a través de nuestro comportamiento. Aprender de la interacción es una idea fundamental que subyace a casi todas las teorías de aprendizaje e inteligencia.

En este libro exploramos un enfoque computacional para aprender de la interacción. En lugar de teorizar directamente sobre cómo aprenden las personas o los animales, exploramos situaciones de aprendizaje idealizadas y evaluamos la efectividad de varios métodos de aprendizaje. Es decir, adoptamos la perspectiva de un investigador o ingeniero de inteligencia artificial. Exploramos diseños de máquinas que son eficaces en la resolución de problemas de aprendizaje de interés científico o económico, evaluando los diseños a través de análisis matemáticos o experimentos computacionales. El enfoque que exploramos, llamado aprendizaje de refuerzo, está mucho más enfocado en el aprendizaje dirigido por objetivos a partir de la interacción que otros enfoques del aprendizaje de las máquinas.

**Aprendizaje por refuerzo**

El aprendizaje por refuerzo es como muchos temas, como el aprendizaje de la máquina, la planificación y el alpinismo, en el sentido de que es simultáneo, o sea, un problema, una clase de métodos de solución que funcionan bien en la clase de problemas, y el campo que estudia estos problemas y su solución se encuentra. Los problemas de aprendizaje de refuerzo implican aprender qué hacer -cómo mapear las situaciones a las acciones- para maximizar una señal numérica de recompensa. De manera esencial son problemas de circuito cerrado porque las acciones del sistema de aprendizaje influyen en sus entradas posteriores. Además, al modelo no se le dice qué acciones debe tomar, como en muchas formas de aprendizaje automático, sino que debe descubrir qué acciones producen la mayor recompensa probándolas. En los casos más interesantes y difíciles, las acciones pueden afectar no sólo a la recompensa inmediata sino también a la siguiente situación y, a través de ella, a todas las recompensas posteriores. Estas tres características -ser de circuito cerrado de manera esencial, no tener instrucciones directas sobre qué acciones tomar y dónde se desarrollan las consecuencias de las acciones, incluidas las señales de recompensa, durante períodos de tiempo prolongados- son los tres rasgos distintivos más importantes de los problemas de aprendizaje de refuerzo.

Una especificación completa de los problemas de aprendizaje de refuerzo en términos de control óptimo de los procesos de decisión de Markov, pero la idea básica es simplemente capturar los aspectos más importantes del problema real al que se enfrenta un agente de aprendizaje que interactúa con su entorno para lograr un objetivo. Es evidente que tal agente debe ser capaz de percibir el estado del medio ambiente en cierta medida y debe ser capaz de tomar medidas que afecten al estado. El agente también debe tener una o varias metas relacionadas con el estado del medio ambiente. La formulación tiene por objeto incluir sólo estos tres aspectos -sensación, acción y meta- en sus formas más sencillas posibles sin trivializar ninguno de ellos.

Cualquier método que se adapte bien a la solución de este tipo de problemas lo consideramos un método de aprendizaje de refuerzo. El aprendizaje de refuerzo es diferente del aprendizaje supervisado, el tipo de aprendizaje estudiado en la mayoría de las investigaciones actuales en el campo del aprendizaje de las máquinas. El aprendizaje supervisado es el aprendizaje de un conjunto de ejemplos etiquetados proporcionados por un supervisor externo conocedor. Cada ejemplo es una descripción de una situación junto con una especificación -la etiqueta- de la medida correcta que el sistema debe adoptar ante esa situación, que a menudo consiste en identificar una categoría a la que pertenece la situación. El objeto de este tipo de aprendizaje es que el sistema extrapole o generalice sus respuestas para que actúe correctamente en situaciones no presentes en el conjunto de la capacitación. Este es un tipo de aprendizaje importante, pero por sí solo no es adecuado para aprender de la interacción. En los problemas interactivos suele ser poco práctico obtener ejemplos de comportamiento deseado que sean a la vez correctos y representativos de todas las situaciones en las que el agente tiene que actuar.

El aprendizaje de refuerzo también es diferente de lo que los buscadores de aprendizaje automático llaman **aprendizaje no supervisado**, que normalmente consiste en encontrar estructuras ocultas en colecciones de datos no etiquetados. Los términos aprendizaje supervisado y aprendizaje no supervisado parecen clasificar exhaustivamente los paradigmas del aprendizaje automático, pero no es así. Aunque uno podría estar tentado de pensar en el aprendizaje de refuerzo como una especie de aprendizaje no supervisado porque no se basa en ejemplos de comportamiento correcto, el aprendizaje de refuerzo trata de maximizar una señal de recompensa en lugar de tratar de encontrar una estructura oculta. Descubrir la estructura en la experiencia de un agente puede ser ciertamente útil en el aprendizaje de refuerzo, pero por sí mismo no aborda el problema del agente de aprendizaje de refuerzo de maximizar una señal de recompensa. Por lo tanto, consideramos que el aprendizaje de refuerzo es un tercer paradigma de aprendizaje de máquina, junto con el aprendizaje supervisado, el aprendizaje no supervisado, y quizás también otros paradigmas.

Uno de los desafíos que surgen en el aprendizaje por refuerzo, y no en otros tipos de aprendizaje, es el equilibrio entre la exploración y la explotación. Para obtener mucha recompensa, un agente de aprendizaje de refuerzo debe preferir las acciones que ha intentado en el pasado y que ha encontrado eficaces para producir recompensa. Pero para descubrir tales acciones, tiene que probar acciones que no ha seleccionado antes. El agente tiene que explotar lo que ya sabe para obtener la recompensa, pero también tiene que explorar para hacer mejores selecciones de acciones en el futuro. El dilema es que ni la exploración ni la explotación se pueden llevar a cabo exclusivamente sin fracasar en la tarea. El agente debe probar una variedad de acciones y favorecer progresivamente las que parezcan mejores. En una tarea estocástica, cada acción debe ser probada muchas veces para obtener una estimación fiable de su recompensa esperada. El dilema exploración-exploración ha sido estudiado intensamente por los matemáticos durante muchas décadas. Por ahora, simplemente observamos que toda la cuestión del equilibrio entre la exploración y la explotación ni siquiera se plantea en el aprendizaje supervisado y no supervisado, al menos en sus formas puristas.

Otra característica clave del aprendizaje por refuerzo es que considera explícitamente todo el problema de un agente dirigido por un objetivo que interactúa con un entorno incierto. Esto contrasta con muchos enfoques que consideran los sub problemas sin abordar cómo podrían encajar en un cuadro más amplio. Por ejemplo, hemos mencionado que gran parte de la investigación sobre el aprendizaje automático se ocupa del aprendizaje supervisado sin especificar explícitamente cómo sería útil finalmente esa capacidad. Otros investigadores han elaborado teorías de planificación con objetivos generales, pero sin considerar el papel de la planificación en la adopción de decisiones en tiempo real, ni la cuestión de la procedencia de los modelos de predicción necesarios para la planificación. Aunque estos enfoques han dado muchos resultados útiles, su enfoque en sub problemas aislados es una limitación significativa.

El aprendizaje por refuerzo toma la dirección opuesta, comenzando con un agente completo, interactivo y de búsqueda de objetivos. Todos los agentes de aprendizaje de refuerzo tienen objetivos explícitos, pueden percibir aspectos de sus entornos y pueden elegir acciones para influir en ellos. Además, se suele suponer desde el principio que el agente tiene que operar a pesar de la gran incertidumbre sobre el entorno al que se enfrenta. Cuando el aprendizaje de refuerzo implica planificación, tiene 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 de entorno. Cuando el aprendizaje de refuerzo implica un aprendizaje supervisado, lo hace por razones específicas que determinan qué capacidades son críticas y cuáles no. Para que la investigación sobre el aprendizaje avance, es necesario aislar y estudiar los subproblemas importantes, pero deben ser subproblemas que desempeñen funciones claras en agentes completos, interactivos y con fines de búsqueda de objetivos, aunque no se puedan completar todos los detalles del agente completo.

Uno de los aspectos más excitantes del aprendizaje moderno de refuerzo es su sustantiva y fructífera interacción con otros discípulos de la ingeniería y la ciencia. El aprendizaje de refuerzo es parte de una tendencia de décadas dentro de la inteligencia artificial y el aprendizaje de máquinas hacia una mayor integración con la estadística, la optimización y otras materias matemáticas. Por ejemplo, la capacidad de algunos métodos de aprendizaje de refuerzo para aprender con aproximaciones parametrizadas aborda la clásica "maldición de la dimensionalidad" en la investigación de operaciones y la teoría de control. De manera más distintiva, el aprendizaje de refuerzo también ha interactuado fuertemente con la psicología y la neurociencia, con beneficios sustanciales en ambos sentidos. De todas las formas de aprendizaje de las máquinas, el aprendizaje de refuerzo es el más cercano al tipo de aprendizaje que hacen los humanos y otros animales, y muchos de los algoritmos centrales del aprendizaje de refuerzo fueron originalmente inspirados por sistemas de aprendizaje biológico. Y el aprendizaje de refuerzo también ha devuelto, tanto a través de un modelo psicológico de aprendizaje animal que coincide mejor con algunos de los datos empíricos, como a través de un modelo influyente de partes del sistema de recompensa del cerebro. El cuerpo de este libro desarrolla las ideas de aprendizaje de refuerzo que pertenecen a la ingeniería y la inteligencia artificial, con conexiones a la psicología y la neurociencia resumidas en los capítulos ?? y ??.

Finalmente, el aprendizaje de refuerzo es también parte de una tendencia más amplia de la inteligencia artificial hacia principios generales simples. Desde finales de los años 60, muchos investigadores de la inteligencia artificial presumieron que no hay principios generales por descubrir, que la inteligencia se debe en cambio a la posesión de vastos números de trucos, procedimientos y heurística con propósitos especiales. A veces se dijo que si pudiéramos obtener suficientes hechos relevantes en una máquina, digamos un millón o mil millones, entonces se volvería inteligente. Los métodos basados en el método general

principios, como la búsqueda o el aprendizaje, se caracterizaban por ser "métodos débiles", mientras que los que se basaban en conocimientos específicos se denominaban "métodos fuertes". Este punto de vista sigue siendo común hoy en día, pero mucho menos dominante. Desde nuestro punto de vista, era simplemente prematuro: se había puesto muy poco esfuerzo en la búsqueda de principios generales para concluir que no había ninguno. La IA moderna incluye ahora mucha investigación que busca principios generales de aprendizaje, búsqueda y toma de decisiones, así como tratar de incorporar grandes cantidades de conocimiento de dominio. No está claro hasta qué punto el péndulo oscilará, pero la investigación sobre el aprendizaje de refuerzo es ciertamente parte del giro hacia principios generales más simples y menos generales de la inteligencia artificial.

**Ejemplos**

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

* Un maestro de ajedrez hace una jugada. La elección se basa tanto en la planificación -anticipando posibles respuestas y contrarrespuestas- como en juicios inmediatos e intuitivos sobre la conveniencia de determinadas posiciones y jugadas.
* Un controlador adaptativo ajusta los parámetros de la operación de una refinería de petróleo en tiempo real. El controlador optimiza el equilibrio entre rendimiento, costo y calidad sobre la base de los costos marginales especificados, sin ceñirse estrictamente a los puntos de ajuste sugeridos originalmente por los ingenieros.
* Un becerro de gacela lucha hasta los pies minutos después de nacer. Media hora más tarde está corriendo a 20 millas por hora.
* Un robot móvil decide si debe entrar en una nueva habitación en busca de más basura para recoger o empezar a tratar de encontrar el camino de vuelta a su estación de recarga de baterías. Toma su decisión basándose en el nivel de carga actual de su batería y en la rapidez y facilidad con la que ha podido encontrar el cargador en el pasado.
* Phil prepara su desayuno. Examinada de cerca, incluso esta actividad aparentemente mun- dane revela una compleja red de comportamiento condicional y de relaciones entre objetivos y subobjetivos: caminar hacia el armario, abrirlo, seleccionar una caja de cereales, luego alcanzar, agarrar y recuperar la caja. Otras secuencias complejas, afinadas e interactivas de comportamiento son re-preguntadas para obtener un tazón, una cuchara y una jarra de leche. Cada paso implica una serie de los movimientos oculares para obtener información y para guiar el movimiento de alcance y de localización. Continuamente se juzga con rapidez cómo llevar 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 agarrar una cuchara o llegar al refrigerador, y está al servicio de otros objetivos, como tener la cuchara con la que comer una vez que el cereal está preparado y obtener a tiempo los alimentos. Tanto si es consciente de ello como si no, Phil accede a la información sobre el estado de su cuerpo que determina sus necesidades nutricionales, su nivel de hambre y sus preferencias alimentarias.

**Elementos del aprendizaje por refuerzo**

Más allá del ambiente, se pueden identificar cuatro subelementos principales de un sistema de aprendizaje de refuerzo: una política, una señal de recompensa , una función de valor y, opcionalmente, un modelo de entorno.

Una política define la forma en que el agente de aprendizaje se comporta en un momento dado. En términos generales, una política es un mapeo desde los estados percibidos del entorno hasta las acciones que se deben tomar cuando se está en esos estados. Corresponde a lo que en psicología se llamaría un conjunto de reglas o asociaciones de estímulo-respuesta (siempre que los estímulos incluyan los que pueden provenir del interior del animal). En algunos casos la política puede ser una simple función o tabla de búsqueda, mientras que en otros puede implicar una computación extensiva, como un proceso de búsqueda. La política es el núcleo de un agente de aprendizaje de refuerzo en el sentido de que por sí sola es suficiente para determinar el comportamiento. En general, las políticas pueden ser estocásticas.

Una señal de recompensa define el objetivo en un problema de aprendizaje de refuerzo. En cada paso de tiempo, el entorno envía al agente de aprendizaje de refuerzo un único número, una recompensa. El único objetivo del agente es maximizar la recompensa total que recibe a largo plazo. La señal de recompensa define así cuáles son los eventos buenos y malos para el agente. En un sistema biológico, podríamos pensar que las recompensas son análogas a las experiencias de placer o dolor. Son las características inmediatas y definitorias del problema al que se enfrenta el agente. La recompensa que se envía al agente en cualquier momento depende de la acción actual del agente y del estado actual de su entorno. El agente no puede alterar el proceso que hace esto. La única forma en que el agente puede influir en la señal de recompensa es a través de sus acciones, que pueden tener un efecto directo en la recompensa, o un efecto indirecto a través del cambio de estado del entorno. En nuestro ejemplo anterior de Phil desayunando, el agente de aprendizaje de refuerzo que dirige su comportamiento puede recibir diferentes señales de recompensa cuando desayuna, dependiendo de cuánta hambre tenga, de su estado de ánimo y de otras características de su cuerpo, que son parte del entorno de su agente de aprendizaje de refuerzo interno. La señal de recompensa es la base principal para alterar la política. Si una acción seleccionada por la política es seguida de una baja recompensa, entonces la política puede ser cambiada para seleccionar alguna otra acción en esa situación en el futuro. En general, las señales de recompensa pueden ser funciones estocásticas del estado del medio ambiente y las acciones tomadas.

Mientras que la señal de recompensa indica lo que es bueno en un sentido inmediato, una función de valor especifica lo que es bueno a largo plazo. En términos generales, el valor de un estado es la cantidad total de recompensa que un agente puede esperar acumular en el futuro, a partir de ese estado. Mientras que las recompensas determinan la conveniencia inmediata e intrínseca de los estados ambientales, los valores indican la conveniencia a largo plazo de los estados, después de tener en cuenta los estados que probablemente seguirán, y las recompensas disponibles en esos estados. Por ejemplo, un estado siempre puede producir una baja recompensa inmediata pero aún así tener un alto valor porque es seguido regularmente por otros estados que producen altas recompensas. O lo contrario podría ser cierto. Para hacer una analogía humana, las recompensas son algo así como el placer (si es alto) y el dolor (si es bajo), mientras que los valores corresponden a un juicio más refinado y previsor de lo complacidos o disgustados que estamos de que nuestro entorno se encuentre en un estado determinado. Expresado de esta manera, esperamos que quede claro que las funciones de los valores formalizan una idea básica y familiar.

Las recompensas son en cierto sentido primarias, mientras que los valores, como predicciones de recompensas, son secundarios. Sin recompensas no podría haber valores, y el único propósito de estimar los valores es lograr más recompensas. No obstante, son los valores los que más nos preocupan a la hora de tomar y evaluar decisiones. Las decisiones de acción se toman sobre la base de juicios de valor. Buscamos acciones que produzcan estados de mayor valor, no la mayor recompensa, porque estas acciones nos proporcionan la mayor cantidad de recompensas a largo plazo. En la toma de decisiones y la planificación, la cantidad derivada llamada valor es la que más nos preocupa. Lamentablemente, es mucho más difícil determinar los valores que las recompensas. Las recompensas son básicamente dadas directamente por el medio ambiente, pero los valores deben ser estimados y reestimados a partir de las secuencias de observaciones que un agente hace a lo largo de toda su vida. De hecho, el componente más importante de casi todos los algoritmos de aprendizaje de refuerzo que consideramos es un método para estimar valores de manera eficiente. El papel central de la estimación de valores es posiblemente lo más importante que hemos aprendido sobre el aprendizaje de refuerzo en las últimas décadas.

El cuarto y último elemento de algunos sistemas de aprendizaje de refuerzo es un modelo del entorno. Esto es algo que imita el comportamiento del entorno, o más generalmente, que permite hacer inferencias sobre cómo se comportará el entorno. Por ejemplo, dado un estado y una acción, el modelo podría predecir el próximo estado y la próxima recompensa resultantes. Los modelos se utilizan para la planificación, con lo que nos referimos a cualquier forma de decidir el curso de acción considerando posibles situaciones futuras antes de que sean realmente experimentadas. Los métodos para resolver los problemas de aprendizaje de refuerzo que utilizan modelos y planificación se denominan métodos basados en modelos, en contraposición a los métodos más sencillos y gratuitos que son explícitamente de ensayo y error para los estudiantes, considerados casi como ello contrario de la planificación.

**1.4 Limitaciones y alcance**

La mayoría de los métodos de aprendizaje de refuerzo que consideramos están estructurados en torno a la estimación de funciones de valor, pero no es estrictamente necesario hacerlo para resolver los problemas de aprendizaje de refuerzo. Por ejemplo, métodos como los algoritmos genéticos, la programación genética, el recocido simulado y otros métodos de optimización se han utilizado para abordar los problemas de aprendizaje de refuerzo sin recurrir nunca a las funciones de valor. Estos métodos evalúan el comportamiento de "toda la vida" de muchos agentes no docentes, cada uno de los cuales utiliza una política diferente para interactuar con su entorno, y seleccionan a los que son capaces de obtener la mayor recompensa. Llamamos a estos métodos evolutivos porque su funcionamiento es análogo a la forma en que la evolución biológica produce organismos con habilidad de ser humano, incluso cuando no aprenden durante sus vidas individuales. Si el espacio de las políticas es suficientemente pequeño, o puede estructurarse de manera que las buenas políticas sean comunes o fáciles de encontrar -o si se dispone de mucho tiempo para la búsqueda-, entonces los métodos evolutivos pueden ser eficaces. Además, los métodos evolutivos tienen ventajas en los problemas en los que el agente de aprendizaje no puede percibir con precisión el estado de su entorno.

Nos centramos en los métodos de aprendizaje de refuerzo que implican el aprendizaje mientras se interactúa con el entorno, lo que no hacen los métodos evolutivos (no evolucionan los algoritmos de aprendizaje, como en algunos de los enfoques que se han estudiado). Creemos que los métodos capaces de aprovechar los detalles de las interacciones de comportamiento individual pueden ser mucho más eficientes que los métodos evolutivos en muchos casos. Los métodos evolutivos ignoran gran parte de la estructura útil del problema del aprendizaje de refuerzo: no utilizan el hecho de que la política que buscan es una función de los estados a las acciones; no se dan cuenta de los estados por los que pasa un individuo durante su vida, ni de las acciones que selecciona. En algunos casos esta información puede ser engañosa (por ejemplo, cuando los estados se perciben erróneamente), pero más a menudo debería permitir una búsqueda más eficiente. Aunque la evolución y el aprendizaje comparten muchas áreas y naturalmente trabajan juntos, no consideramos que los métodos evolutivos por sí mismos sean especialmente adecuados para reforzar los problemas de aprendizaje. Para simplificar, en este libro cuando usamos el término "método de aprendizaje de refuerzo" no incluimos los métodos evolutivos.

Sin embargo, incluimos algunos métodos que, como los métodos evolutivos, no apelan a funciones de valor. Estos métodos buscan en espacios de políticas definidas por una colección de parámetros numéricos. Estiman las direcciones en las que los parámetros deben ajustarse para mejorar lo más rápidamente posible el rendimiento de una política. Sin embargo, a diferencia de los métodos evolutivos, producen estos es- timos mientras el agente está interactuando con su entorno y así puede aprovechar los detalles de las interacciones de comportamiento individuales. Métodos como este, llamados métodos de gradiente de política, han demostrado ser útiles en muchos problemas, y algunos de los métodos más simples de aprendizaje de refuerzo entran en esta categoría. De hecho, algunos de estos métodos aprovechan las estimaciones de las funciones de valor para “improbar” sus estimaciones de gradientes. En general, la distinción entre los métodos de gradiente de política y otros métodos que incluimos como métodos de aprendizaje de refuerzo no está claramente definida.

La conexión del aprendizaje de refuerzo con los métodos de optimización merece algún comentario adicional porque es fuente de un malentendido común. Cuando decimos que el objetivo de un agente de aprendizaje de refuerzo es maximizar una señal de recompensa numérica, por supuesto no estamos insistiendo en que el agente tiene que lograr realmente el objetivo de la máxima recompensa. Tratar de maximizar una cantidad no significa que esa cantidad se maximice alguna vez. El punto es que un agente de aprendizaje de re-información siempre está tratando de aumentar la cantidad de recompensa que recibe. Muchos factores pueden impedirle alcanzar el máximo, incluso si existe uno. En otras palabras, la optimización no es lo mismo que la optimización.

El documento que presentamos va a tomar los aspectos más básicos de los métodos mediante el cual el aprendizaje por refuerzo logra alcanzar sus objetivos.

**Métodos de solución tabulares**

**Multi-arm Bandits o Bandido de múltiples brazos**

En la teoría de la probabilidad, el problema de los bandidos con múltiples brazos (a veces denominado problema de los bandidos con brazos K-[1] o N[2]) es un problema en el que se debe asignar un conjunto fijo y limitado de recursos entre las opciones en competencia (alternativas) de manera que se maximice su ganancia esperada, cuando las propiedades de cada opción sólo se conocen parcialmente en el momento de la asignación, y pueden comprenderse mejor a medida que pasa el tiempo o mediante la asignación de recursos a la opción. Este es un problema clásico de aprendizaje de refuerzo que ejemplifica el dilema de compromiso entre exploración y explotación. Su nombre proviene de imaginar a un jugador en una fila de máquinas tragaperras (a veces conocidas como "bandidos de un solo brazo"), que tiene que decidir a qué máquinas jugar, cuántas veces jugar a cada máquina y en qué orden jugarlas, y si continuar con la máquina actual o probar una máquina diferente. El problema de los bandidos de varios brazos también entra en la amplia categoría de la programación estocástica.

En el problema, cada máquina proporciona una recompensa aleatoria de una distribución de probabilidad específica para esa máquina. El objetivo del jugador es maximizar la suma de las recompensas obtenidas a través de una secuencia de tirones de palanca. La compensación crucial a la que se enfrenta el jugador en cada prueba es entre la "explotación" de la máquina que tiene el mayor rendimiento esperado y la "exploración" para obtener más información sobre los rendimientos esperados de las otras máquinas. La compensación entre la exploración y la explotación también se enfrenta en el aprendizaje de la máquina. En la práctica, los bandidos multiarmados se han utilizado para modelar problemas como la gestión de proyectos de investigación en una gran organización como una fundación científica o una empresa farmacéutica. En las primeras versiones del problema, el jugador comienza sin tener un conocimiento inicial de las máquinas.

Herbert Robbins en 1952, al darse cuenta de la importancia del problema, construyó estrategias de selección de población convergentes en "algunos aspectos del diseño secuencial de los experimentos". Un teorema, el índice Gittins, publicado por primera vez por John C. Gittins, da una política óptima para maximizar la recompensa descontada esperada.

**Procesos finitos de decisión de Markov**

En matemáticas, un proceso de decisión de Markov (MDP) es un proceso de control estocástico de tiempo discreto. Proporciona un marco matemático para modelar la toma de decisiones en situaciones en las que los resultados son en parte aleatorios y en parte bajo el control de una persona que toma decisiones. Los MDP son útiles para estudiar los problemas de optimización resueltos mediante la programación dinámica y el aprendizaje de refuerzo. Los MDPs fueron conocidos al menos desde los años 50; un cuerpo central de investigación sobre los procesos de decisión de Markov resultó del libro de 1960 de Ronald Howard, Dynamic Programming and Markov Processes. Se utilizan en muchas disciplinas, incluyendo la robótica, el control automático, la economía y la manufactura. El nombre de MDPs viene del matemático ruso Andrey Markov ya que son una extensión de las cadenas de Markov.

En cada paso del tiempo, el proceso está en algún estado, y la persona que toma la decisión puede elegir cualquier acción que esté disponible en el estado ***s***. El proceso responde en el siguiente paso de tiempo moviéndose aleatoriamente a un nuevo estado ***s'***, y dando a la persona que toma la decisión una recompensa correspondiente .

La probabilidad de que el proceso se mueva a su nuevo estado ***s'*** está influenciada por la acción elegida. Específicamente, viene dada por la función de transición de estado . Por lo tanto, el siguiente estado ***s'*** depende del estado actual ***s*** y de la acción de la persona que toma la decisión ***a***. Pero dados los estados y el estilo a, es condicionalmente independiente de todos los estados y acciones previas; en otras palabras, las transiciones de estado de un MDP satisfacen la propiedad de Markov.

Los procesos de decisión de Markov son una extensión de las cadenas de Markov; la diferencia es la adición de acciones (que permiten elegir) y recompensas (que dan motivación). Por el contrario, si sólo existe una acción para cada estado (por ejemplo, "esperar") y todas las recompensas son las mismas (por ejemplo, "cero"), un proceso de decisión de Markov se reduce a una cadena de Markov.

**Programación dinámica**

La programación dinámica es principalmente una optimización sobre la recursividad simple. Dondequiera que veamos una solución recursiva que tiene llamadas repetidas para las mismas entradas, podemos optimizarla usando la Programación Dinámica. La idea es simplemente almacenar los resultados de los subproblemas, para que no tengamos que volver a calcularlos cuando sea necesario más tarde. Esta simple optimización reduce las complejidades del tiempo de lo exponencial a lo polinómico. Por ejemplo, si escribimos una simple recursividad para los Números de Fibonacci, obtenemos una complejidad temporal exponencial y si la optimizamos almacenando soluciones de subproblemas, la complejidad temporal se reduce a lineal.

**Métodos de Monte Carlo**

El método de Montecarlo es un método no determinista o estadístico numérico, usado para aproximar expresiones matemáticas complejas y costosas de evaluar con exactitud. El método se llamó así en referencia al Casino de Montecarlo (Mónaco) por ser “la capital del juego de azar”, al ser la ruleta un generador simple de números aleatorios. El nombre y el desarrollo sistemático de los métodos de Montecarlo datan aproximadamente de 1944 y se mejoraron enormemente con el desarrollo de la computadora.

El uso de los métodos de Montecarlo como herramienta de investigación proviene del trabajo realizado en el desarrollo de la bomba atómica durante la Segunda Guerra Mundial en el Laboratorio Nacional de Los Álamos en EE. UU. Este trabajo conllevaba la simulación de problemas probabilísticos de hidrodinámica concernientes a la difusión de neutrones en el material de fisión. Esta difusión posee un comportamiento eminentemente aleatorio. En la actualidad es parte fundamental de los algoritmos de raytracing para la generación de imágenes 3D.

En la primera etapa de estas investigaciones, John von Neumann y Stanislaw Ulam refinaron esta ruleta y los métodos "de división" de tareas. Sin embargo, el desarrollo sistemático de estas ideas tuvo que esperar al trabajo de Harris y Herman Kahn en 1948. Aproximadamente en el mismo año, Enrico Fermi, Nicholas Metropolis y Ulam obtuvieron estimadores para los valores característicos de la ecuación de Schrödinger para la captura de neutrones a nivel nuclear usando este método.

El método de Montecarlo proporciona soluciones aproximadas a una gran variedad de problemas matemáticos posibilitando la realización de experimentos con muestreos de números pseudoaleatorios en una computadora. El método es aplicable a cualquier tipo de problema, ya sea estocástico o determinista. A diferencia de los métodos numéricos que se basan en evaluaciones en N puntos en un espacio M-dimensional para producir una solución aproximada, el método de Montecarlo tiene un error absoluto de la estimación que decrece como en virtud del teorema del límite central.

La invención del método de Montecarlo se asigna a Stanislaw Ulam y a John von Neumann. Ulam ha explicado cómo se le ocurrió la idea mientras jugaba un solitario durante una enfermedad en 1946. Advirtió que resulta mucho más simple tener una idea del resultado general del solitario haciendo pruebas múltiples con las cartas y contando las proporciones de los resultados que computar todas las posibilidades de combinación formalmente. Se le ocurrió que esta misma observación debía aplicarse a su trabajo de Los Álamos sobre difusión de neutrones, para la cual resulta prácticamente imposible solucionar las ecuaciones íntegro-diferenciales que gobiernan la dispersión, la absorción y la fisión. “La idea consistía en probar con experimentos mentales las miles de posibilidades, y en cada etapa, determinar por casualidad, por un número aleatorio distribuido según las probabilidades, qué sucedería y totalizar todas las posibilidades y tener una idea de la conducta del proceso físico”.

Podían utilizarse máquinas de computación, que comenzaban a estar disponibles, para efectuar las pruebas numéricas y en efecto reemplazar el aparato experimental del físico. Durante una de las visitas de von Neumann a Los Álamos en 1946, Ulam le mencionó el método. Después de cierto escepticismo inicial, von Neumann se entusiasmó con la idea y pronto comenzó a desarrollar sus posibilidades en un procedimiento sistemático. Ulam expresó que Montecarlo “comenzó a tener forma concreta y empezó a desarrollarse con todas sus fallas de teoría rudimentaria después de que se lo propuse a Johnny”.

A principios de 1947 Von Neumann envió una carta a Richtmyer a Los Álamos en la que expuso de modo influyente tal vez el primer informe por escrito del método de Montecarlo. Su carta fue encuadernada junto con la respuesta de Richtmyer como un informe de Los Álamos y distribuida entre los miembros del laboratorio. Von Neumann sugería aplicar el método para rastrear la generación isótropa de neutrones desde una composición variable de material activo a lo largo del radio de una esfera. Sostenía que el problema era adecuado para el ENIAC y estimaba que llevaría 5 horas calcular la acción de 100 neutrones a través de un curso de 100 colisiones cada uno.

Ulam estaba particularmente interesado en el método Montecarlo para evaluar integrales múltiples. Una de las primeras aplicaciones de este método a un problema determinista fue llevada a cabo en 1948 por Enrico Fermi, Ulam y von Neumann cuando consideraron los valores singulares de la ecuación de Schrödinger.

**Aprendizaje por diferencia temporal**

Diferencia temporal (TD) de aprendizaje se refiere a una clase de modelo libre de aprendizaje por refuerzo métodos que aprenden por bootstrapping de la estimación actual de la función de valor. Estos métodos de muestreo del medio ambiente, como los métodos de Monte Carlo, y realizar actualizaciones en base a las estimaciones actuales, como la programación dinámica métodos.

Mientras que los métodos de Monte Carlo sólo se ajustan sus estimaciones una vez conocido el resultado final, los métodos TD ajustar para que coincida con las predicciones más tarde, más precisa, las predicciones sobre el futuro antes de que se conozca el resultado final. Esta es una forma de bootstrapping, como se ilustra con el siguiente ejemplo:

"Suponga que desea predecir el tiempo para el sábado, y usted tiene algún modelo que predice el tiempo del sábado, dado el clima de cada día de la semana. En el caso estándar, se tenga que esperar hasta el sábado y luego ajustar todos sus modelos. Sin embargo, cuando lo es, por ejemplo, el viernes, usted debe tener una idea bastante buena de lo que el tiempo sería el sábado - y por lo tanto ser capaz de cambiar, por ejemplo, el modelo del sábado antes del sábado llega ".

Los métodos de diferencias temporales están relacionadas con el modelo de diferencia temporal de aprendizaje animal que hemos visto en la introducción de este texto.

**Trazos de elegibilidad**

Los trazos de elegibilidad son uno de los mecanismos básicos del aprendizaje de refuerzo. Por ejemplo, en el popular algoritmo TD(λ), la λ se refiere al uso de un rastro de elegibilidad. Casi cualquier método de diferencia temporal (TD), como el Q-learning o el Sarsa, puede combinarse con trazas de elegibilidad para obtener un método más general que pueda aprender de forma más eficiente.

Hay dos formas de ver los rastros de elegibilidad. La visión más teórica, que enfatizamos aquí, es que son un puente entre los métodos de TD y los de Monte Carlo. Cuando los métodos de TD se aumentan con trazas de elegibilidad, producen una familia de métodos que abarcan un espectro que tiene los métodos de Monte Carlo en un extremo y los métodos de TD de un paso en el otro. En medio se encuentran los métodos intermedios que a menudo son mejores que cualquiera de los dos métodos extremos. En este sentido, las trazas de elegibilidad unifican los métodos de TD y de Monte Carlo de una manera valiosa y reveladora.

La otra forma de ver las trazas de elegibilidad es más mecánica. Desde esta perspectiva, un rastro de elegibilidad es un registro temporal de la ocurrencia de un evento, como la visita de un estado o la realización de una acción. El rastro marca los parámetros de memoria asociados al evento como elegibles para someterse a cambios de aprendizaje. Cuando se produce un error de TD, sólo se asigna crédito o culpa por el error a los estados o acciones elegibles. Por lo tanto, los rastros de elegibilidad ayudan a cerrar la brecha entre los eventos y la información de capacitación. Al igual que los métodos de TD en sí, los rastros de elegibilidad son un mecanismo básico para la asignación temporal de créditos.

Por razones que se pondrán de manifiesto en breve, la visión más teórica de los rastros de elegibilidad se denomina vista hacia adelante, y la visión más mecanicista se denomina vista hacia atrás. La vista hacia adelante es más útil para comprender lo que se calcula por los métodos que utilizan trazas de elegibilidad, mientras que la vista hacia atrás es más apropiada para desarrollar la intuición acerca de los propios algoritmos. En este capítulo presentamos ambas vistas y luego establecemos los sentidos en los que son equivalentes, es decir, en los que describen los mismos algoritmos desde dos puntos de vista. Como de costumbre, consideramos primero el problema de la predicción y luego el problema del control. Es decir, consideramos primero cómo se utilizan las trazas de elegibilidad para ayudar a predecir los rendimientos en función del estado para una póliza fija (es decir, en la estimación). Sólo después de explorar los dos puntos de vista de los rastros de elegibilidad dentro de este escenario de predicción, extendemos las ideas a los valores de acción y a los métodos de control.

**Planificación y aprendizaje con métodos tabulares**

Por modelo del medio ambiente entendemos cualquier cosa que un agente pueda usar para predecir cómo responderá el medio ambiente a sus acciones. Dado un estado y una acción, un modelo produce una predicción del siguiente estado y la siguiente re-protección. Si el modelo es estocástico, entonces hay varios posibles estados y recompensas siguientes, cada uno con alguna probabilidad de ocurrir. Algunos modelos producen una descripción de todas las posibilidades y sus probabilidades; a éstas las llamamos modelos de distri- bución. Otros modelos producen sólo una de las posibilidades, muestreadas de acuerdo con las probabilidades; a estos los llamamos modelos de muestra. Por ejemplo, el con- sider modelando la suma de una docena de dados. Un modelo de distribución produciría todas las sumas posibles y sus probabilidades de ocurrir, mientras que un modelo de muestra produciría una suma individual dibujada de acuerdo con esta distribución de probabilidad. El tipo de modelo asumido en la programación dinámica de las estimaciones de las probabilidades de transición de estado y las recompensas esperadas, p(sJ s, a) y r(s, a, sJ)- es un modelo de distribución.

Los modelos pueden utilizarse para imitar o simular la experiencia. Dado un estado inicial y una acción, un modelo de muestra produce una posible transición, y un modelo de distribución genera todas las transiciones posibles ponderadas por sus probabilidades de ocurrencia. Dado un estado de partida y una política, un modelo de muestra podría producir un episodio completo, y un modelo de distribución podría generar todos los posibles episodios y sus probabilidades. En cualquiera de los casos, decimos que el modelo se utiliza para simular el entorno y producir una experiencia simulada.

La palabra planificación se utiliza de varias maneras diferentes en distintos campos. Utilizamos el término para referirnos a cualquier proceso computacional que tome un modelo como entrada y produzca o mejore una política para interactuar con el entorno modelado:

Modelo Planing Politicas

Dentro de la inteligencia artificial, hay dos enfoques distintos de la planificación según nuestra definición. En la planificación del espacio estatal, que incluye el enfoque que tomamos en este libro, la planificación se ve principalmente como una búsqueda a través del espacio estatal de una política o camino óptimo para un objetivo. Las acciones causan transiciones de un estado a otro, y las funciones de valor se calculan sobre los estados. En lo que llamamos planificación del espacio de los planes, la planificación se ve en cambio como una búsqueda a través del espacio de los planes. Los operadores transforman un plan en otro, y las funciones de valor, si las hay, se definen sobre el espacio de los planes. La planificación del plan-espacio incluye métodos evolutivos y planificación de orden parcial, un tipo de planificación popular en la inteligencia artificial en la que el orden de los pasos no está completamente terminado en todas las etapas de la planificación. Los métodos de planificación del espacio son difíciles de aplicar eficazmente a los problemas de control óptimo estocástico que son el centro de atención del aprendizaje de reinicio, y no los consideramos más allá.

**Métodos de solución aproximados**

Usando el aprendizaje de refuerzo (RL), los agentes (controladores) pueden aprender a interactuar de forma óptima con entornos complejos (sistemas). RL puede abordar problemas de una variedad de dominios, incluyendo inteligencia artificial (AI), robótica, control automático, investigación de operaciones, economía, medicina, etc. Por ejemplo, en la IA, RL proporciona una forma de construir agentes de aprendizaje que optimizan su comportamiento en entornos inicialmente desconocidos. En el control automático, RL puede en principio resolver problemas de control óptimo no lineal y estocástico, sin necesidad de un modelo. Un agente de RL interactúa con su entorno midiendo los estados y aplicando acciones de acuerdo con su política.

Una señal de recompensa indica el rendimiento inmediato, y el objetivo es encontrar una política óptima que maximice la función de valor, es decir, la recompensa acumulada a largo plazo en función de los estados y posiblemente de las acciones. La LR clásica requiere representaciones exactas de las funciones de valor y las políticas, y por lo tanto se limita a problemas pequeños y discretos. Sin embargo, la mayoría de los problemas realistas tienen variables con un número grande o infinito de valores posibles (por ejemplo, variables continuas); en esos problemas, las funciones de valor y las políticas deben ser aproximadas. Gracias a los recientes avances en los algoritmos basados en la aproximación, RL ha obtenido éxitos impresionantes en aplicaciones como el control de robots, el vuelo autónomo de helicópteros, la interconexión de un cerebro animal con un brazo robótico, el tratamiento simulado de las infecciones por VIH, etc.

Desafortunadamente, para un recién llegado al campo de la RL aproximada será difícil de comprender, debido a lo intrincado y extenso mundo de algoritmos y resultados, junto con la escasez de papers y modelos libres que se encuentran a disposición. En esta visión general, intentamos generar lecturas adicionales para profundizar en los diferentes temas que poseen un importante componente aproximado de RL que alivia esta situación. Partiendo de su programación dinámica exacta (DP) y de las raíces de RL, describimos tres clases principales de técnicas de RL aproximada: iteración de valores aproximados, iteración de políticas y búsqueda de políticas. Introducimos algoritmos fuera de línea, así como algoritmos en línea, que recogen sus propios datos mediante la interacción con el sistema. Prestamos atención a los métodos de gradiente de políticas y de actores críticos, así como a la iteración de políticas basada en la simulación. Para concluir, se comparan los diversos tipos de enfoques, se esbozan varias formas importantes de mejorar los métodos aproximados de RL y se proporcionan indicaciones para una lectura más profunda.

**Bibliografía**

Para un conocimiento general adicional del aprendizaje por refuerzo, remitimos al lector a los libros de Szepesv'ari (2010), Bertsekas y Tsitsiklis (1996) y Kaelbling (1993a). Los libros que adoptan una perspectiva de investigación de control u operación son los de Sistemas Inteligentes, Powell (2011), y Lewis y Liu (2012). Dos números especiales de la revista Machine Learning se centran en el aprendizaje de refuerzo: Sutton (1992) y Kaelbling (1996). En ellos se presentan encuestas útiles de Barto (1995b); Kaelbling, Littman y Moore (1996); y Keerthi y Ravindran (1997). El volumen editado por Weiring y van Otterlo (2012) ofrece un excelente panorama de los acontecimientos recientes.