1. Introducción al aprendizaje por refuerzo

El aprendizaje por refuerzo es un enfoque computacional al aprendizaje mediante la interacción con el entorno. Este enfoque está a centrado en **aprendizaje dirigido a objetivo**.

El aprendizaje por refuerzo es una aproximación a entender y automatizar el aprendizaje y la toma de decisiones. Se distingue de otros enfoques en el énfasis en el aprendizaje de un agente a través de interacciones con su entorno sin depender de la supervisión externa o tener un modelo completo del entorno.

El aprendizaje por refuerzo utiliza una estructura formal que define las interacciones entre el agente aprendiz y su entorno en términos de: estados, acciones y recompensas. Esta estructura es una forma sencilla de representar características esenciales de un problema de inteligencia artificial. Dichas características incluyen un sentido de causa-efecto, un sentido de incertidumbre y no determinismo y la existencia de metas explícitas.

Los principales elementos del aprendizaje por refuerzo son: la política, función de valor, la función de recompensa y el modelo.

Los conceptos de valor y función de valor son características clave para la mayoría de métodos de aprendizaje por refuerzo. Estas funciones de valor son importantes para una búsqueda eficiente en el espacio de políticas. y permiten diferenciar los métodos de aprendizaje por refuerzo de los métodos evolutivos que buscan directamente en el espacio de políticas.

Actualmente existen dos métodos principales que se usan en RL: los gradientes de política basados en probabilidad y el Q-Learning basado en el valor.

1.1. Inspiración Biológica

Aprender mediante la interacción con el entorno es la idea fundamental detrás de casi todas las teorías de aprendizaje e inteligencia. La idea de que aprendemos interaccionando con nuestro entorno es, probablemente, lo primero que se nos ocurre cuando pensamos en el aprendizaje biológico.

1.2. Características

El aprendizaje por refuerzo es un problema de bucle cerrado porque la influencia de las acciones de los sistemas aprendices será su propia entrada más adelante. Además, no se le dice qué acciones hacer (como en otros métodos de aprendizaje automático), sin embargo se diferencia en que el agente descubre con qué acciones acaba obteniendo una mayor recompensa probándolas. Las acciones que toma el agente involucran, no solamente la recompensa inmediata, sino que también tiene en cuenta las futuras recompensas. Estas tres características son las más importantes del aprendizaje por refuerzo.

1. Ser de bucle cerrado
2. No decirle instrucciones directas como qué acciones hacer.
3. Las consecuencias de las acciones (incluyendo las señales de recompensa) están en juego durante periodos largos de tiempo.

El agente debe tener una meta (o metas) relacionada con el estado del entorno. Se incluyen, entonces, tres aspectos:

1. Percepciones
2. Acciones
3. Meta(s)

1.3. Aprendizaje por Refuerzo vs. Aprendizaje Supervisado

El aprendizaje por refuerzo es diferente del aprendizaje supervisado, el tipo de aprendizaje utilizado en la mayoría de los campos de investigación sobre aprendizaje automático. El aprendizaje supervisado consiste en aprender de un conjunto de entrenamiento etiquetado proporcionado por un supervisor externo.

En problemas interactivos, a menudo no es práctico obtener ejemplos de comportamientos deseables que sean correctos y representativos de todas las situaciones en las que el agente debe actuar.

Consideramos el aprendizaje por refuerzo como un tercer paradigma en el aprendizaje automático junto con el aprendizaje supervisado y el no supervisado.

Uno de los retos que surgen en el aprendizaje por refuerzo, y no en otros paradigmas, es el equilibrio entre la exploración y el aprovechamiento para obtener una mayor recompensa. Para conseguir una mayor recompensa, el agente aprendiz debe preferir acciones que ha intentado en el pasado y las ha encontrado útiles para producir una recompensa. Pero para descubrir dichas acciones, debe intentar acciones que no ha seleccionado antes. El agente deberá aprovechar lo que ya sabe para obtener la recompensa, pero también debe explorar otras acciones para seleccionar mejores acciones en el futuro.

El dilema está en que ni la exploración ni el aprovechamiento pueden ser utilizados exclusivamente sin fallar en la tarea. El agente debe intentar una variedad de acciones y, progresivamente, favorecer aquellas que aparentan ser las mejores.
Un agente aprendiz debe encontrar el equilibrio entre exploración y explotación de los recursos. Por el contrario, los sistemas de aprendizaje supervisado y no supervisado toman la respuestas directamente de los datos de entrenamiento sin explorar otras respuestas.

En el aprendizaje por refuerzo el agente debe aprender por su cuenta las acciones que debe tomar en cada momento en función de la recompensa que obtenga al realizar dicha acción.
En el aprendizaje por refuerzo la respuesta correcta no se da explícitamente. El agente necesita aprender por prueba y error teniendo como única referencia la recompensa obtenida después de cada acción.

El aprendizaje por refuerzo es un proceso de **decisión múltiple**, es decir, forma una cadena de toma de decisiones a través del tiempo requerido para terminar un trabajo específico. Por otro lado, el aprendizaje supervisado es un proceso de decisión única (una instancia, una predicción).

2. Historia

La historia del aprendizaje por refuerzo converge a través de tres hilos de investigación: **el aprendizaje mediante prueba y error**, el **problema de control óptimo** y su solución utilizando programación dinámica y, por último, **los métodos de diferenciación temporal**. Estos tres hilos convergieron a finales de los 80's para producir el campo del aprendizaje por refuerzo.

2.1. Aprendizaje mediante Prueba y Error

La idea del aprendizaje mediante prueba y error comienza en 1850 con el psicólogo británico Conway Lloyd Morgan que utilizó el término para describir sus observaciones en el comportamiento animal.

El primero que utilizó este término como principio de aprendizaje fue Edward Thorndike, que lo llamó la **ley de efecto** porque describe el efecto que situaciones de refuerzo a la hora de realizar ciertas acciones.

El término de **refuerzo** en el contexto del aprendizaje biológico viene definido como el fortalecimiento de unos patrones de comportamiento como resultado de darle un estímulo a un animal (reforzador) en un momento determinado en relación con otro estímulo o respuesta. El refuerzo produce cambios en el comportamiento que persisten después de eliminar el reforzador.

La idea de implementar el aprendizaje mediante prueba y error en los algoritmos aparece durante los primeros pensamientos de inteligencia artificial. En un informe en 1948, Alan Turing describió un diseño de un sistema placer-dolor en el que estaba trabajando junto con la ley de efecto.

En 1952, Claude Shannon muestra un ratón que era capaz de resolver un laberinto utilizando esta técnica.

2.2. Control Óptimo y Programación Dinámica

El término de control óptimo empieza a usarse a finales de los años 50 para describir el problema de diseñar un controlador que minimizase la medida del comportamiento de un sistema dinámico. Uno de los enfoques a este problema fue propuesto por **Richard Bellman**. Este enfoque utilizaba conceptos de los estados de los sistemas dinámicos y las funciones de valor o funciones de retorno óptimo (ecuaciones de Bellman).

Los métodos clásicos para resolver problemas de control óptimo resolviendo las ecuaciones de Bellman se empezó a conocer como **programación dinámica**. Este paradigma se considera uno de los métodos más factibles de resolver problemas de control estocásticos. Sin embargo, sufren de lo que Bellman llamaba *maldición de la dimensionalidad*, es decir, que sus requisitos computacionales crecen exponencialmente con el número de variables de estado, no obstante, sigue siendo más eficiente que otros métodos más generales.

2.3. Métodos de Diferenciación Temporal

Los métodos de aprendizaje de diferenciación temporal se distinguen en que estos son dirigidos mediante la diferencia entre estimaciones temporales sucesivas. Este hilo es más pequeño que los otros dos, pero ha jugado un papel importante en el campo, en parte, porque los métodos de diferencia temporal parecían ser nuevos y únicos para el aprendizaje por refuerzo.

Los orígenes de este tipo de aprendizaje están, en parte, en la psicología del aprendizaje animal. Particularmente en los reforzadores secundarios. Un reforzador secundario es un estímulo combinado con un reforzador primario (comida o dolor) y, como resultado tiene unas propiedades de refuerzo similares.

3. Elementos del Aprendizaje por Refuerzo

3.1. Política

Una política define el comportamiento del agente aprendiz en un momento determinado. Es una función que convierte los estados percibidos en el entorno a acciones que se realizarán cuando el agente esté en dichos estados (función de mapeo).
En psicología se corresponde con lo que se llama **reglas estímulo-respuesta** o **asociaciones**.

Es el núcleo del agente de aprendizaje por refuerzo en el sentido en que utilizando solamente la política, es suficiente para determinar el comportamiento de dicho agente. Generalmente las políticas son estocásticas.

3.2. Señal de Recompensa

Una señal de recompensa define la meta del problema de aprendizaje por refuerzo. En cada paso, el entorno envía al agente un número (recompensa). El único objetivo del agente es maximizar dicha recompensa.

La señal de recompensa define cuáles son los eventos buenos y los perjudiciales para el agente.
La recompensa que se le proporciona al agente en cada momento depende de la acción del agente en ese instante y del estado del entorno del agente. El agente no puede alterar el proceso que otorga la recompensa. La única forma en la que el agente pueda influir en la señal de recompensa es a través de sus acciones, que

pueden tener un efecto directo en la recompensa o indirecto, cambiando el estado del entorno.

La señal de recompensa es la base para alterar la política. Si a una acción seleccionada por la política le corresponde una recompensa baja, entonces la política deberá ser cambiada para seleccionar otra acción cuando vuelva a ocurrir esa situación en el futuro.

En general, las señales de recompensa son funciones estocásticas del estado del entorno y de las acciones realizadas.

3.3. Función de Valor

La función de valor especifica qué es bueno a largo plazo para el agente.
El **valor** de un estado es la cantidad de recompensa total que un agente puede esperar acumular en un futuro, empezando desde el estado actual.

Mientras que la recompensa determina cómo de bueno es el estado actual (inmediato), el valor indica, a largo plazo, como de bueno son los estados después de tener en cuenta los estados más probables que va a seguir el agente y la recompensa de estos estados.

Sin recompensa no habría valores y, el único propósito de estimar los valores es para conseguir una mayor recompensa. La selección de acciones están basadas en un **juicio de valor**. Buscamos acciones que conduzcan a estados de mayor valor, no mayor recompensa, ya que dichas acciones obtendrán la mayor cantidad de recompensa a largo plazo.

Las recompensas las da directamente el entorno, mientras que los valores deben ser estimados y reestimados a través de las observaciones del agente durante su tiempo de vida.

La parte más importante de la mayoría de algoritmos de aprendizaje por refuerzo es el método que estima los valores eficientemente.

3.4. Modelo

El último elemento de los sistemas de aprendizaje por refuerzo es el **modelo** del entorno. Es algo que simula el comportamiento del entorno o, dicho de otra forma, que permite hacer inferencia sobre cómo se comportará el entorno.

Los modelos se usan para planificar, es decir, cualquier manera de decidir el transcurso de una acción considerando futuras situaciones antes de que se experimenten. Los métodos que utilizan modelos y planificación en el aprendizaje por refuerzo se llaman **métodos basados en modelos**, mientras que, por otro lado, existen métodos más simples llamados **métodos libres de modelo** que son aprendices de prueba y error.

4. Limitaciones

La mayoría de métodos de aprendizajes por refuerzo están construidos alrededor de la estimación de una función de valor, pero no es estrictamente necesario hacerlo así para resolver este tipo de problemas. Por ejemplo, los algoritmos genéticos, la programación genética y otros métodos de optimización se utilizan también como enfoque para resolver los problemas de aprendizaje por refuerzo sin tener una función de valor.

Estos tipos de métodos evalúan el tiempo de vida de muchos agentes no aprendices, cada uno utiliza una política diferente para interactuar con su entorno y selecciona aquellos que son capaces de obtener una mayor recompensa.

4.1. Métodos Evolutivos

Los métodos evolutivos se llaman de este modo porque actúan análogamente a como lo hace la evolución biológica. Producen organismos con diferentes habilidades incluso cuando no aprenden durante sus periodos de vida individuales.

Los métodos evolutivos ignoran la mayoría de la estructura útil de los problemas de aprendizaje por refuerzo. Estos métodos no utilizan el hecho de que la política que se busca es una función de estados a acciones, no se dan cuenta en qué estados ha estado el agente durante su tiempo de vida o qué acción ha elegido.

Nuestro enfoque está en los métodos de aprendizaje por refuerzo que involucran el aprendizaje mediante la interacción con su entorno, cosa que los métodos evolutivos no hacen (en la mayoría de los casos).

Cuando decimos que el objetivo de un agente aprendiz es maximizar la señal de recompensa, no estamos asegurando que el agente alcance el máximo valor, sino que, el algoritmo de aprendizaje por refuerzo, intentará aumentar la cantidad de recompensa que recibe. Existen muchos factores que pueden hacer que el agente no alcance el máximo valor (si existiera). En otras palabras, **optimización** no es lo mismo que **optimalidad**.

5. Procesos Finitos de Decisión de Markov

En el aprendizaje por refuerzo, el agente y su entorno interacciones mediante una secuencia de tiempo discreta. Las especificaciones de sus interfaces definen una tarea en particular: las acciones son elegidas por el agente, los estados son la base para realizar dichas elecciones y las recompensas son la base para evaluar las elecciones.

Todo lo que está dentro del agente es conocido y controlable por el mismo agente, mientras que lo que está fuera es incontrolable, pero puede ser conocido o no.

El entorno satisface la propiedad de Markov si su señal de estado resume el pasado sin degradar la capacidad de predecir el futuro. Si se mantiene esta propiedad, entonces el entorno se conoce como un **proceso de decisión de Markov**. Por tanto, un proceso finito de decisión de Markov tiene estados y acciones finitas.

Los procesos de decisión de Markov se definen mediante una tupla (S,A,T,R) donde:

- S: conjunto no vacío de estados. El estado s_i representa el entorno en el paso i.
- A: conjunto no vacío de acciones posibles. En el paso i, el agente ejecuta la acción a_i.
- T: Distribución de probabilidad del conjunto:

  $$T : SxAxS \to \mathcal{P}(S)$$

  donde T(s,a,s') es la probabilidad de que se realice una transición del estado s al estado s', ejecutando la acción a. Se puede reescribir de la forma: $T(s, a, s') = P(s')|(s, a)$
- R: Es el esfuerzo esperado dada una acción desde un estado concreto. R es la función:

  $$R : SxAxS \to \mathbb{R}$$

  de manera que $R(s, a, s') = E(r_{i+1}|s_i = s, a_i = a, s_{i+1} = s')$, donde r_{i+1} es una variable aleatoria que corresponde a ala recompensa recibida en la etapa $i + 1$. En cada etapa el agente recibe un refuerzo dependiendo de la decisión tomada. El objetivo final es maximizar el refuerzo acumulado por el agente a lo largo de la vida del sistema.

Se considera que el DMP en el que se enmarca el problema es **finito** (el conjunto de estados y de acciones es un conjunto finito).

Observación: La distribución de probabilidad definida con la función T solo depende del estado actual, del siguiente y de la acción a ejecutar. Es decir, la probabilidad de alcanzar el próximo estado s_i depende tan solo del estado actual y no de los estados que le preceden. Esta propiedad se conoce como **propiedad de Markov** (Ejemplo del laberinto)

En cada etapa, el agente recibe un refuerzo dependiendo de la decisión tomada. El objetivo final es maximizar el refuerzo acumulado por el agente a lo largo de la vida del sistema. Podemos presentar la recompensa acumulada como:

$$\sum_{k=0}^{\infty} \gamma^k R_k$$

donde $R_k = R(s_k, a_k, s_{k+1})$ es el refuerzo recibido en la etapa k. El parámetro γ es un factor de descuento (entre 0 y 1) en el que, exceptuando cuando $\gamma = 1$, disminuye el esfuerzo recibido por cada etapa que pasa. Este parámetro controla la importancia de los refuerzos a largo plazo, cuanto menor es el valor, menos peso se le otorga a las recompensas de las últimas etapas y más a las primeras y viceversa. Este enfoque de la recompensa se denomina **modelo de recompensa descontada en horizonte infinito** ya que se calcula para un número infinito de pasos.

Hay otros criterios como el **modelo de recompensa en horizonte finito** o el **modelo de recompensa media**. En el primero disponemos de un número finito de pasos en los que se reciben recompensas, por lo que la recompensa acumulada solo se calcula en estos pasos:

$$E = \left\{ \sum_{k=0}^{M} r_{i+k+1} | s_t = s \right\}$$

El modelo de recompensa media trata de maximizar la recompensa media a largo plazo:

$$\lim_{M \to \infty} E = \left\{ \sum_{k=0}^{M} r_{i+k+1} | s_t = s \right\}$$

En el aprendizaje por refuerzo normalmente se utiliza el **modelo de recompensa descontada en horizonte infinito**.
La solución del problema vendrá dada como la asociación de una acción a cada estado de manera que se alcance el objetivo según el modelo de recompensa descontada en horizonte infinito. Esta asociación define el comportamiento del agente y se denomina **política.** La función política se denota:

$$\pi : S \to A$$

La política depende del estado actual (cumple la propiedad de Markov). Existen dos conceptos impotantes en la búsqueda de la optimalidad de la política:

- **Función de valor-estado**: Representa el esfuerzo esperado desde el estado s y siguiendo con la política π, $V_\pi : S \to \mathbb{R}$.

  $$V_\pi(s) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} | s_0 = s, \ \pi \right]$$

  Esta función cuantifica la bondad de esta política para el caso particular de $s_0 = s$.

- **Función de valor-acción**: Representa el esfuerzo esperado empezando desde el estado s, tomando la acción a y siguiendo con la política π, $\mathcal{Q}_\pi : SxA \to \mathbb{R}$

  $$Q_\pi(s,a) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} | s_0 = s, a_0 = a, \ \pi \right]$$

  Esta función explica la bondad de esta política para el caso particular de la pareja estado-acción $(s_0 = s, a_0 = a)$.

De la función valor-estado podemos derivar la **ecuación de Bellman**:

$$\begin{aligned}
V_\pi(s) &= E\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} | s_0 = s, \ \pi \right] \\
&= E\left\{ r_{i+1} + \gamma \sum_{k=0}^{\infty} r_{i+k+2} | s_t = s, \ \pi \right\} \\
&= \sum_a p(s,a) \left[E\left\{ r_{i+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{i+k+2} | s_i = s, a_i = a, \pi \right\} \right] \\
&= \sum_a p(s,a) \sum_{s'} T(s,a,s') \left[E\left\{ r_{r+1} | s_i = s, a_i = a, s_{i+1} = s' \right\} + \gamma E\left\{ \sum_{k=0}^{\infty} \gamma^k r_{i+k+2} | s_{i+1} = s', \pi \right\} \right] \\
&= \sum_a p(s,a) \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma E\left\{ \sum_{k=0}^{\infty} \gamma^k r_{i+k+2} | s_{i+1} = s', \pi \right\} \right] \\
&= \sum_a p(s,a) \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_\pi(s') \right]
\end{aligned}$$

donde $p(s, a)$ representa la probabilidad de tomar la acción a desde el estado s.

La ecuación de Bellman se puede reescribir como un sistema lineal de ecuaciones. Podemos denominas **V** como un vector columna formado por los valores de $V_p i$ en cada estado s, **R** como un vector columna de longitud el numero de estado en el que cada componente es:

$$R_i = \sum_a p(s_i, a) \sum_{s'} T(s_i, a, s') R(s_i, a, s')$$

y **P** es una matriz SxS dada por:

$$P_{i,j} = \sum_a p(s_i, a) T(s_i, a, s_j)$$

Obtenemos el sistema lineal:

$$V = R + \gamma P V$$

cuya solución viene dada por:

$$V = (I - \gamma P)^{-1} R$$

el sistema $V = (I - \gamma P)^{-1} R$ tiene solución **única** con $\gamma < 1$

Podemos relacionar las dos funciones (estado-valor y valor-acción):

$$V_\pi(s) = \sum_a p(s, a) \mathcal{Q}_\pi(s, a)$$

La política óptima (pi^*) obtendrá los valores más altos posibles. En el caso de la función V esto significa que: $V_{pi^*}(s) \geq V_\pi(s)$ para todo estado s. Este valor máximo será:

$$V^*(s) = \max_\pi V_\pi(s) \qquad \forall s \in S$$

Para la función valor-accion (Q) es análogo.

Si relacionamos ambas funciones óptimas:

$$V_\pi(s) = \sum_a p(s, a) \mathcal{Q}_\pi(s, a) \leq \max_a \mathcal{Q}_\pi(s, a)$$

Para el caso óptimo tendremos: $V^*(s) = \max_a \mathcal{Q}^*(s, a) \qquad \forall s \in S$
Para las ecuaciones de Bellman se obtienen las **ecuaciones óptimas de Bellman**:

$$V^*(s) = \max_a \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V^*(s') \right]$$

$$Q^*(s, a) = \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma \max_{a'} Q^*(s', a') \right]$$

En la última ecuación permite encontrar la política óptima si se conocen los valores óptimos de la función $\mathcal{Q}$. La asociación adecuada de la acción a cada estado será la acción que hga máximo el valor de esta función:

$$\pi^*(s) = \arg\max_{a \in A} \mathcal{Q}^*(s, a)$$