Universidad Católica del Norte

Departamento de Ingeniería de Sistemas y Computación

Proyecto: Fundamentos de Aprendizaje por Refuerzo

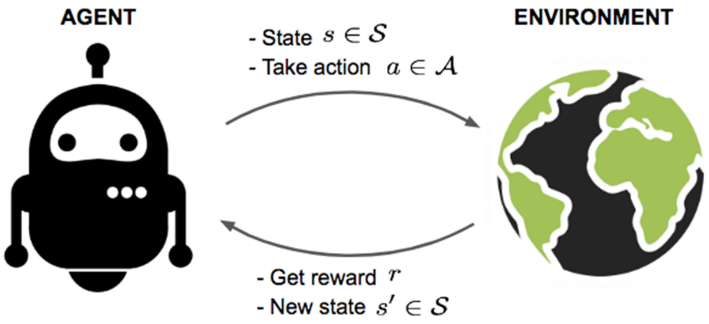
---

1. Introducción

En esta Sección se hace una breve descripción de los conceptos y técnicas involucradas en el desarrollo del proyecto para la asignatura, luego en la Sección II se describe en detalle el proyecto a realizar y en la Sección III se describe el método de evaluación, entregas y características de estas entregas durante el proyecto.

1.1. *Aprendizaje por Refuerzo*

El aprendizaje por refuerzo (*Reinforcement Learning*, RL) es una rama del aprendizaje automático en la que un agente autónomo aprende a tomar decisiones secuenciales mediante la interacción con un entorno, con el objetivo de maximizar una señal de recompensa acumulada a lo largo del tiempo. A diferencia del aprendizaje supervisado, donde se dispone de ejemplos con respuestas correctas, en el aprendizaje por refuerzo el agente no recibe indicaciones explícitas sobre qué acción es la mejor, sino que debe descubrirla a través de la experiencia, probando distintas acciones y observando sus consecuencias. Este proceso implica la exploración de nuevas acciones y la explotación de aquellas que han demostrado ser efectivas, equilibrando ambas para mejorar su comportamiento. Los elementos clave de este enfoque son el agente, el entorno, el conjunto de estados posibles, las acciones disponibles, la función de recompensa, la política de decisión y la función de valor, que en conjunto definen la dinámica de aprendizaje y adaptación del agente.



1.2. *Proceso de decisión de Markov*

El proceso decisión de Markov (*Markov Decision Process*, MDP) (ver Fig. 1) se refiere a un proceso de toma de decisiones estocástico que utiliza un marco matemático para modelar la toma de decisiones de un sistema dinámico [1]. Se utiliza en escenarios donde los resultados son aleatorios o controlados por quien toma las decisiones, quien las toma secuencialmente a lo largo del tiempo. El MDP evalúa qué acciones debe tomar el agente teniendo en cuenta el estado actual y el entorno del sistema. El MDP postula que el futuro es independiente del pasado, dado el presente. Esto significa que, se puede predecir el estado siguiente a partir del estado actual, sin necesidad de conocer el estado anterior.

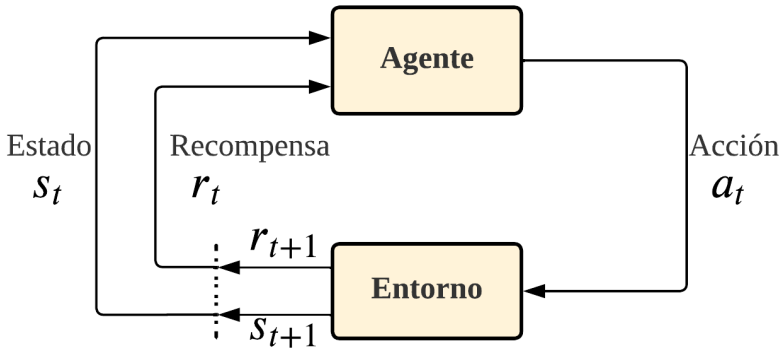


Figura 1: Diagrama del MDP.

El principio de RL contiene términos claves para entender este campo. A continuación, se mencionan los principales términos:

- **Agente:** es la entidad que toma decisiones e interactúa con el entorno para aprender cómo lograr un objetivo específico, el cual es maximizar las recompensas acumuladas.
- **Entorno:** representa el contexto en el que opera el agente y en el que se desarrollan las interacciones.
- **Acciones:** se refiere a una decisión o movimiento que realiza un agente para interactuar con su entorno.
- **Estados:** representa la condición del entorno en un momento específico, proporcionando al agente la información necesaria para tomar decisiones.
- **Recompensa:** valor numérico que indica la retroalimentación que recibe el agente después de realizar una acción en un estado específico del entorno.
- **Política:** estrategia que guía en el comportamiento del agente en función de los estados. Actúa como un mapeo entre la acción y el estado actual.

La idea principal del MDP es que el agente busca aprender una política óptima, es decir, una estrategia que maximice la recompensa acumulada a largo plazo. El MDP está definido por una tupla $(\mathscr{S}, \mathscr{A}, \mathscr{T}, \mathscr{R}, \gamma)$, donde:

- $\mathscr{S}$: es un conjunto de estados
- $\mathscr{A}$: es un conjunto de acciones
- $\mathscr{T}(s'|s,a)$: función de transición
- $\mathscr{R}(s,a)$: función de recompensa
- γ: factor de descuento

1.3. *Función de recompensas*

La función de recompensas en el aprendizaje por refuerzo es un mecanismo que asigna un valor numérico (llamado recompensa) a cada acción que realiza el agente en un determinado estado del entorno. Su propósito es indicar qué tan buena o mala fue la acción tomada con respecto al objetivo final del agente. Si la recompensa es alta (por ejemplo, positiva), significa que la acción contribuyó favorablemente a alcanzar la meta; si es baja o negativa, significa que la acción fue desfavorable o indeseada. La función de recompensas define el comportamiento deseado del agente y guía su proceso de aprendizaje, ya que el agente intentará maximizar la suma total de recompensas recibidas a lo largo del tiempo. Esta función no le dice explícitamente al agente qué hacer, sino que le da señales sobre las consecuencias de sus decisiones, permitiéndole ajustar su política de actuación para mejorar su desempeño.

2. Descripción del proyecto

El presente proyecto comprende el diseño e implementación de un agente basado en RL para la resolución de tareas en un entorno simulado. En este trabajo se deberá desarrollar un agente que aprenda a resolver una tarea específica dentro de un entorno controlado, utilizando técnicas de RL. El agente deberá ser capaz de:

- Observar estados del entorno.
- Seleccionar acciones posibles según una política.
- Recibir recompensas en función de su comportamiento.
- Ajustar su política de actuación para mejorar su desempeño a través de la experiencia.

2.1. Resultados

Para este trabajo se toma en cuenta las siguientes tareas y consideraciones:

1. Realice una reflexión sobre la selección del algoritmo de RL más adecuado para la resolución de su problema específico y argumentar su selección.
2. Cada grupo deberá entrenar su agente respetando la función de recompensa original de su entorno definida en la librería Gymnasium. En el caso de que el grupo esté utilizando un entorno externo a la librería Gymnasium, deben referenciar en el documento escrito la fuente del entorno obtenido.
3. Diseñar una función de recompensa totalmente de su propiedad, con el objetivo de facilitar o acelerar el proceso de aprendizaje y mejorar el comportamiento del agente. Entrenar nuevamente el agente bajo esta nueva función de recompensa.
4. Obtener las curvas de recompensa total acumulada y promedio por episodio, diferencias en velocidad de convergencia, y comparación de comportamiento cualitativo del agente para los dos entrenamientos (recompensa original vs recompensa diseñada).
5. Obtener las siguiente métricas cuantitativas:
    - **Recompensa promedio por episodio** (para ambas funciones de recompensa)
    - **Desviación estándar de la recompensa** (para medir la estabilidad del aprendizaje)
    - **Número de episodios hasta convergencia** (definido por un criterio claro, por ejemplo, alcanzar un umbral de recompensa promedio).
    - **Tiempo de entrenamiento total** (en pasos o segundos).

3. Evaluación

La evaluación de la actividad considera 40 % el trabajo funcional, 40 % el documento escrito, y 20 % la presentación oral del mismo.

4. Fechas

- **Fecha de la presentación:** Lunes 15 de Diciembre de 2025
- **Fecha de la entrega del escrito:** Miércoles 17 de Diciembre de 2025
- **Consultas:** christian.camacho@ce.ucn.cl

Referencias

[1] Barto Andrew and Sutton Richard S. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.