Matemáticas Aplicadas y Ciencias de la Computación
Inteligencia Artificial: Representación y Solución de Problemas

Universidad del Rosario

Q-learning en el Mundo del Taxi

Diryon Yonith Mora Romero, Laura Valentina Gonzalez Rodriguez

Resumen

En este artículo, exploramos el desafío del aprendizaje por refuerzo en el entorno del Taxi, donde nuestro agente busca optimizar su ruta en una ciudad dinámica. Centrado en la filosofía del Q-learning clásico, el agente se sumerge en la exploración del entorno, actualizando sus estrategias para encontrar de manera eficiente a un pasajero y llevarlo a su destino en un mundo preestablecido. Explorando las complejidades del Q-learning, buscamos ajustar cuidadosamente los parámetros para perfeccionar la capacidad del agente de cumplir su tarea de manera óptima.

1. Introducción

En el intrincado entramado de la inteligencia artificial y el fascinante campo del aprendizaje por refuerzo, emerge un microcosmos único: el escenario del Taxi. En este desafío clásico, nuestro agente, asumiendo el papel de taxi, se sumerge en la tarea de navegar por las caóticas calles de una ciudad virtual, sorteando obstáculos y descifrando las complejidades del tráfico urbano. Sin embargo, este no es un simple trayecto; es una coreografía compleja entre decisiones y recompensas, donde el agente persigue la armonía perfecta entre la eficiencia temporal y las satisfacciones de los pasajeros.

En este intrépido viaje, la filosofía del Q-learning clásico se erige como el faro conductor que guía al taxi a través de un laberinto de posibilidades. El Q-learning, venerado en el ámbito del aprendizaje por refuerzo, se convierte en el mapa cognitivo del agente, evaluando el valor de cada acción en cada estado. En esta danza algorítmica, el agente se sumerge en la exploración del entorno, actualizando sus Q-valores con cada recompensa obtenida y cada estimación del valor del siguiente estado. Así, en cada esquina de la ciudad, el taxi aprende, evoluciona y traza su propio camino hacia una política óptima, buscando la máxima recompensa a largo plazo.

Este viaje de descubrimiento nos lleva al corazón del problema del taxi de gym. Nuestro agente, el taxi, enfrenta el desafío de aprender cómo llegar de manera óptima hasta el pasajero y llevarlo a su destino en un mundo preestablecido. Aunque los muros de este mundo están claramente definidos, la ubicación del pasajero y su destino varían entre cuatro posiciones en cada ejecución, agregando una capa adicional de complejidad, veasé la Figura 1 para una mejor visualización del entorno. En este entorno totalmente observable, determinista, episódico, estático, discreto y conocido, el algoritmo de Q-learning se convierte en la herramienta esencial. Ajustando meticulosamente sus parámetros, nuestro agente se sumerge en el proceso de aprendizaje, perfeccionando su capacidad para cumplir su tarea de manera eficaz y eficiente.

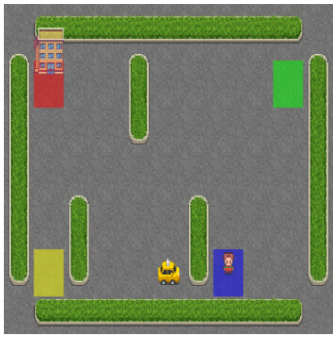

Figura 1: Entorno del Taxi

2. Métodos

Para abordar el problema del taxi en el entorno Gym, se implementó el algoritmo de aprendizaje por refuerzo conocido como Q-learning. En este enfoque, el taxi actúa como un agente que toma decisiones secuenciales para maximizar su recompensa a lo largo del tiempo.

Representación del Problema

El entorno del taxi en Gym se modela como un espacio con estados, acciones y recompensas. Los estados incluyen la ubicación del taxi, la posición del pasajero, la ubicación del destino y la presencia del pasajero en el taxi. Las acciones posibles son movimientos en las cuatro direcciones, recoger y dejar al pasajero. Las recompensas se asignan según la ejecución de acciones, se aplican las siguientes:

- Se otorga una recompensa de -1 por cada paso realizado, a menos que se desencadene otra recompensa.
- Se otorga una recompensa de +20 por entregar exitosamente al pasajero.
- Se penaliza con -10 al ejecutar las acciones de "recoger" y "dejar" al pasajero de manera ilegal.

Algoritmo de Q-learning

Se utiliza una tabla Q para almacenar los valores de utilidad de cada par estado-acción. Inicialmente, la tabla se inicia con ceros. La actualización de los valores Q se realiza mediante la ecuación de Bellman:

$$Q(s,a) = (1-\alpha)\cdot Q(s,a) + \alpha \cdot \left(R + \gamma \cdot \max Q(s',a')\right)$$

Donde:

- $Q(s,a)$ es el valor actualizado para el par estado-acción.
- α es la tasa de aprendizaje.
- R es la recompensa obtenida por la acción a en el estado s.
- γ es el factor de descuento.
- $\max Q(s',a')$ es el valor máximo de la tabla Q para el próximo estado s'.

Se emplea un enfoque ϵ-greedy para equilibrar la exploración y explotación. Con probabilidad ϵ, el agente elige una acción aleatoria; de lo contrario, selecciona la acción con el máximo valor en la tabla Q. Además, se ajustan hiperparámetros como la tasa de aprendizaje (α), el factor de descuento (γ) y la probabilidad de exploración (ϵ) para optimizar el rendimiento del algoritmo.[1]

Implementación

Se instanció un agente Q-learning basado en un enfoque de agentes en inteligencia artificial. Los parámetros clave, como la tasa de aprendizaje (α) y el factor de descuento (γ), se configuraron para guiar el proceso de aprendizaje. El agente sigue un bucle de entrenamiento a través de múltiples episodios. En cada episodio, el agente elige acciones, interactúa con el entorno y actualiza la tabla Q según la ecuación de Bellman.

En cada paso, el agente selecciona una acción basada en su política actual. Debido a la estrategia epsilon-greedy, con probabilidad ϵ, el agente elige explorar nuevas acciones al azar; de lo contrario, explota las acciones conocidas. Este enfoque ayuda al agente a descubrir nuevas estrategias mientras utiliza las ya aprendidas.

Después de ejecutar una acción en el entorno, el agente observa el nuevo estado resultante y la recompensa asociada. Se utiliza la fórmula de actualización de Q-values para

ajustar las estimaciones del agente. La actualización considera la recompensa obtenida, el Q-value actual y el Q-value máximo esperado para el próximo estado. La tasa de aprendizaje (α) controla la magnitud de esta actualización, determinando cuánto se ajustan los valores Q en cada paso.

Tras actualizar los Q-values, el agente ajusta su política para reflejar las nuevas estimaciones. Este proceso contribuye a una toma de decisiones más informada, ya que el agente adapta sus acciones según las recompensas y experiencias pasadas.

3. Resultados y Discusiones

Configuración de Parámetros

Para el entrenamiento del agente Q-learning, después de realizar multiple experimentación y ajuste, se seleccionaron los siguientes parámetros:

- Tasa de Descuento (gamma): 0.8
- Exploración (epsilon): 0.2
- Tasa de Aprendizaje (alpha): 0.6

Resultados del Entrenamiento

Durante el entrenamiento, el agente Q-learning mostró los siguientes resultados:

- Suma Promedio de Recompensas: -540.9465
- Porcentaje de Éxito: 99.875 %

La Figura 2, que representa el histograma de la suma de recompensas en función de su frecuencia, destaca principalmente valores menores a 0, con la mayoría de los datos concentrados entre 0 y -1000. Esto sugiere que el agente enfrentó desafíos significativos durante el entrenamiento, pero logró superar obstáculos para alcanzar el objetivo en la mayoría de los episodios.

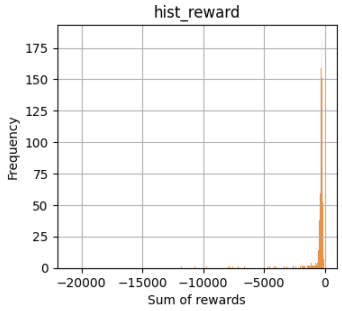

Figura 2: Recompensas en el entrenamiento

Resultados de las Pruebas

Al evaluar el agente entrenado, se obtuvieron los siguientes resultados:

- Suma Promedio de Recompensas: 8.07
- Porcentaje de Éxito: 100. %

Estos resultados indican que el agente Q-learning logró generalizar con éxito sus aprendizajes del entrenamiento para realizar tareas similares en un nuevo entorno. La Figura 3, que representa el histograma de recompensas con valores positivos, respalda esta conclusión al mostrar que el agente recibió predominantemente recompensas positivas durante las pruebas.

Figura 3: Recompensas en la prueba

La Figura 4, que presenta las recompensas por episodio (episode vs total_reward), revela la oscilaciones en las recompensas, que varían entre 2 y 14. Este comportamiento ser por la dinámica del entorno del taxi y las decisiones

secuenciales tomadas por el agente para maximizar sus recompensas.

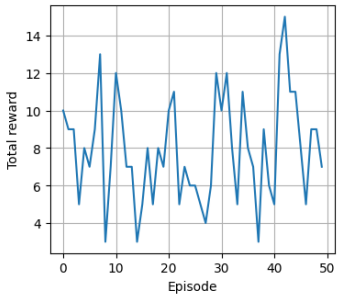

Figura 4: Recompensas por episodio

Comparación sin Entrenamiento

Se evidencia claramente la diferencia de rendimiento entre un agente entrenado y uno no entrenado a través de los gifs proporcionados en el repositorio de GitHub[2]. Estos gifs ilustran cómo el agente entrenado toma decisiones más informadas y eficientes en comparación con el agente no entrenado.

4. Discusión

La elección cuidadosa de parámetros, como la tasa de descuento, la probabilidad de exploración y la tasa de aprendizaje, contribuyó al éxito del agente Q-learning en el entorno del taxi. Aunque enfrentó desafíos durante el entrenamiento, logró aprender una política efectiva que le permitió alcanzar sus objetivos en la mayoría de las situaciones de prueba.

El análisis de las recompensas y el rendimiento en episodios proporciona una comprensión más profunda de cómo el agente Q-learning se adapta y responde a las complejidades del entorno. Estos resultados respaldan la efectividad del enfoque de aprendizaje por refuerzo, específicamente Q-learning, en la resolución del problema del taxi en Gym.

Se destacan áreas para futuras mejoras, como la exploración de técnicas avanzadas de ajuste de hiperparámetros y la consideración de enfoques de aprendizaje profundo para problemas más complejos.

5. Conclusiones

El agente Q-learning demostró ser capaz de aprender una política efectiva para el problema del taxi, alcanzando un elevado porcentaje de éxito del 100 % en las pruebas. Los ajustes cuidadosos de los parámetros, como la tasa de descuento y la probabilidad de exploración, fueron cruciales para el éxito del aprendizaje por refuerzo.

Además, el agente no solo logró un alto rendimiento durante el entrenamiento, sino que también demostró su capacidad para generalizar sus aprendizajes a nuevas situaciones durante las pruebas. Esto destaca la adaptabilidad del enfoque Q-learning y su capacidad para tomar decisiones informadas en entornos desconocidos.

No obstante, nuestros experimentos sugieren que, en este contexto, los hiperparámetros del Q-learning no desempeñan un papel tan crucial como el número de episodios y rondas. A pesar de configuraciones diferentes, como 2000 episodios con 10 rondas y pruebas con 2000 episodios y 1000 rondas, así como hiperparámetros completamente distintos, se observaron resultados comparables.

A pesar del éxito alcanzado, hay oportunidades para mejorar el rendimiento y la eficiencia del agente. Se identifican posibles áreas de mejora, como la exploración de técnicas avanzadas de ajuste de hiperparámetros y la consideración de enfoques de aprendizaje profundo para problemas más complejos.

Referencias

[1] S. M. Kerner, "What is q-learning?" May. 2023. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/Q-learning

[2] "Repositorio github." [Online]. Available: https://github.com/Lau-Gonz/AI-taxi-qlearning.git