Azul Noguera, Rocio Gonzalez, Valentina Pancaldi y Paula Jordan

Informe de Trabajo Práctico 1

Materia: AI & Neurociencias
Docentes: Erik Ernst y Agustin Gravano

Decisiones tomadas

En el desarrollo de nuestro agente para el juego 'Diez Mil', comenzamos definiendo las políticas y los estados de manera simplificada, centrándonos en dos acciones fundamentales: plantarse o tirar todos los dados que no sumaban puntos. Esta simplificación permitió que el agente aprendiera de manera efectiva en los primeros intentos. Sin embargo, pronto nos dimos cuenta de que esta aproximación no siempre conducía a un comportamiento óptimo, por lo que decidimos ajustar gradualmente las políticas.

Observamos que, para tomar decisiones sobre si tirar o plantarse, los estados relevantes en las políticas aprendidas deberían considerar la cantidad de dados no utilizados y el puntaje acumulado hasta el momento, en lugar de los números específicos de los dados restantes. Inicialmente, intentamos distinguir entre:

- *Dados restantes: (2, 2, 4) con un puntaje actual de 250*
- *Dados restantes: (2, 3, 6) con un puntaje actual de 250*

como estados diferentes. Sin embargo, constatamos que la probabilidad de obtener puntos en el próximo turno era la misma en ambos casos, ya que esta probabilidad no depende de los números específicos de los dados restantes, sino del número total de dados que se pueden volver a lanzar. Por lo tanto, decidimos redefinir los estados relevantes como una tupla de (cantidad de dados no utilizados, puntaje actual). En este nuevo esquema, los dos estados anteriores se representarían como:

- *Cantidad de dados no utilizados: 3 y puntaje actual = 250*

Además, consideramos importante incluir el puntaje actual, ya que, en situaciones donde quedan 3 dados no utilizados, el puntaje acumulado influye en la decisión. Por ejemplo, si el puntaje actual es 200, es más conveniente tirar los dados, dado que el riesgo es menor y existe la posibilidad de sumar puntos adicionales. En cambio, si el puntaje actual es 3100, es preferible plantarse para evitar el riesgo de perder una gran cantidad de puntos acumulados.

Otra observación relevante fue el comportamiento cuando se disponen de 6 dados para lanzar. En este caso, la estrategia se vuelve más compleja. La posibilidad de obtener puntos

adicionales es alta, pero también lo es el riesgo de perder puntos si el resultado de los dados no es favorable. Por lo tanto, en estas situaciones, el agente debe equilibrar cuidadosamente el riesgo y la recompensa potencial, ajustando su estrategia en función del puntaje actual y de la cantidad de dados disponibles.

Combinación de hiperparámetros

Al entrenar el agente, probamos varias combinaciones de hiper parámetros, empezamos con configuraciones comunes, pero fuimos ajustándolos a lo largo del proceso para mejorar la exploración y el aprovechamiento del conocimiento adquirido.

- **Episodios:** Comenzamos con un número bajo para realizar pruebas rápidas, pero fuimos aumentando para poder ver el desempeño del agente a medida que se encontraba en instancias más avanzadas del juego.
- **Alpha**: Ajustamos esta tasa entre 0.1 y 0.5, buscando un equilibrio entre la velocidad de aprendizaje y la estabilidad del agente.
- **Gamma**: Este parámetro fue probado con valores cercanos a 0.9, con el fin de balancear la recompensa inmediata frente a la acumulada a largo plazo.
- **Epsilon**: Experimentamos con valores decrecientes a lo largo del entrenamiento para fomentar la exploración inicial y luego enfocarnos en la explotación de las mejores estrategias. Terminamos eligiendo un valor de 0.2 que fue el que nos arrojó los mejores resultados.

Conclusiones

El desempeño del agente desarrollado ha demostrado mejoras significativas en comparación con los enfoques anteriores, como el agente Plantón y el agente Aleatorio. Aunque los resultados finales no alcanzaron las expectativas iniciales en relación con el esfuerzo invertido, la experiencia obtenida a través del proceso de prueba y error ha sido de suma importancia.

A través de la experimentación y el ajuste de nuestras políticas, descubrimos que las estrategias simplificadas iniciales no reflejaban toda la complejidad del juego. Al darnos cuenta de que la probabilidad de obtener puntos dependía del número total de dados disponibles y no de los números específicos, redefinimos nuestros estados y políticas, mejorando el rendimiento del agente. Aprendimos a considerar el puntaje actual y los dados no utilizados para afinar las decisiones del agente, minimizando riesgos y maximizando la eficiencia. Aunque los resultados finales no cumplieron con las expectativas, la experiencia nos proporcionó una comprensión más profunda del juego y una base sólida para futuros desarrollos en agentes de aprendizaje por refuerzo.