Diego Alonzo 20172
Brian Carrillo 21108

Montecarlo

Implementación

Este laboratorio utiliza el entorno Blackjack-v1 de Gymnasium, donde cada estado se representa como una tupla (suma del jugador, carta visible del dealer, indicador de As usable). Las acciones posibles son Stick (0) y Hit (1), y la recompensa se obtiene al final del episodio con valores típicos de +1, 0 o −1. La generación de trayectorias se realiza mediante una función de simulación que itera hasta el estado terminal, acumulando estados, acciones y recompensas. Sobre estas trayectorias se aplica aprendizaje Monte Carlo de primera visita en un esquema on-policy con exploración epsilon-greedy. Específicamente, se estima Q(s, a) por promedio incremental de los retornos observados, recorriendo el episodio en reversa y actualizando solo la primera visita de cada par (s, a). Finalmente, la política de comportamiento durante el aprendizaje selecciona la acción con mayor valor esperado con probabilidad $1 - \varepsilon$ y explora con probabilidad ε.

Evolución de la política

En las primeras etapas la política es más ruidosa por la limitada cobertura del espacio de estados y la exploración activa; el patrón típico muestra una preferencia amplia por Hit en gran parte del dominio, especialmente para sumas del jugador más bajas. Conforme se acumulan episodios, emergen fronteras más nítidas entre regiones de Stick y Hit. Sin As usable, la política es más conservadora en general, pero arriesga en los extremos: pide con sumas bajas porque quedarse casi no ofrece chance de ganar y también en algunas manos altas (como 16–17) frente a cartas fuertes del dealer, donde quedarse suele significar perder. Con As usable, el "colchón" que ofrece contar el As como 11 sin pasarse permite jugar de forma más agresiva, pidiendo incluso con totales que normalmente serían de quedarse, y retrasando el umbral de stick hasta manos muy fuertes (≥ 18).

Resultados

Tras el aprendizaje, la política greedy derivada de Q se evaluó con simulación de múltiples episodios. En una corrida de referencia en el notebook, el rendimiento promedio reportado se situó alrededor de −0. Este valor es coherente con la naturaleza del Blackjack con reglas estándar y baraja infinita, donde la esperanza a largo plazo del jugador óptimo suele ser ligeramente negativa. Las superficies 3D de $V^*(s)$ reflejan este comportamiento: el valor esperado crece con la suma del jugador, especialmente entre 18 y 21, y decrece cuando el dealer muestra cartas fuertes. Asimismo, la superficie con As usable domina a la de "Sin As usable" en amplias regiones del estado, confirmando el beneficio de disponer de un As que puede contarse como 11 sin provocar un bust. Por diseño, los mapas de política se centran en el rango de sumas 12–21 porque por debajo de 12 el Hit es casi siempre la acción dominante, y la estructura de decisiones es menos informativa.

Diego Alonzo 20172
Brian Carrillo 21108

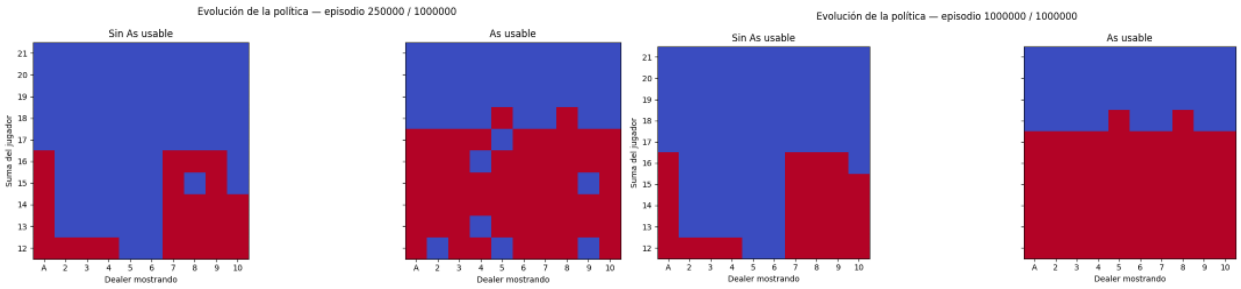

Figura 1. Política óptima en episodio 250,000 y tras todos los episodios.

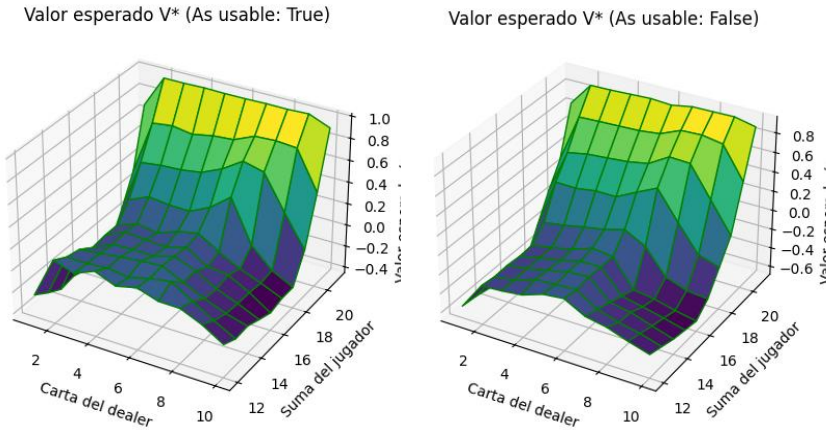

Figura 3. Gráfica 3D del valor esperado según As usable y no usable.

Conclusiones

La implementación de Monte Carlo on-policy con exploración epsilon-greedy, junto con promedios incrementales de primera visita, produce políticas plausibles y estables que convergen hacia fronteras de decisión acordes con la teoría. La separación entre los casos con y sin As usable es clara y se traduce en decisiones más agresivas cuando el jugador dispone del As usable, mientras que frente a cartas fuertes del dealer la política se vuelve más conservadora, especialmente si no hay As usable. El rendimiento ligeramente negativo obtenido es consistente con expectativas teóricas del juego y valida tanto el pipeline de aprendizaje como las visualizaciones. Aunque el método es simple y no requiere modelo, su eficiencia muestral es limitada y exige un número elevado de episodios para alisar Q y estabilizar la política, sobre todo en estados menos visitados. Para mejorar la calidad y la estabilidad de los resultados, resulta útil incrementar el número de episodios, aplicar un decaimiento suave de ε desde 0.1 hacia 0.01 para reducir exploración al final sin empobrecer la cobertura inicial, y fijar semillas y opciones del entorno que alineen las condiciones con las de referencias estándar.