Métodos Monte Carlo

La reciente situación de pandemia global ha obligado a varios negocios a adaptarse e incluir cada vez más trabajo remoto. Este es el caso de las carreras de autos, donde por ejemplo Fox ha comenzado a transmitir carreras virtuales.

En este contexto, un piloto misterioso decide participar, pero con una salvedad, quiere que sus acciones sean controladas por código. Para ello decide iniciar desarrollando políticas a mano y luego probarlas.

Para probar una de sus políticas plantea su problema de selección de acciones como un Markov Decision Process (MDP), de modo tal que partiendo de un cierto estado del motor, la pista y velocidad actual, ser capaz de ir subiendo y bajando la velocidad, a fin de maximizar la velocidad sin hacer sufrir al motor, y con esto lograr maximizar la velocidad media hasta la meta.

Dados los siguientes episodios **observados**, ejecutados de acuerdo a una política arbitraria **desconocida**, calcule los valores $V_{\pi_d}(s)$ y $Q_{\pi_d}(s, a)$.

Notas: Los valores de γ y α se asumen 1.

Episodio 1:

Estado (motor/velocidad)	Acción	Recompensa	Estado siguiente
(normal, rápido)	frenar	60	(frío, lento)
(frío, lento)	acelerar	110	(normal, rápido)
(normal, rápido)	acelerar	120	(tibio, rápido)
(tibio, rápido)	frenar	50	(tibio, lento)
(tibio, lento)	frenar	30	(normal, muy lento)
(normal, muy lento)	frenar	0	(normal, detenido)

Episodio 2:

Estado (motor/velocidad)	Acción	Recompensa	Estado siguiente
(normal, muy lento)	acelerar	100	(normal, rápido)
(normal, rápido)	acelerar	125	(tibio, rápido)
(tibio, rápido)	acelerar	130	(caliente, muy rápido)
(caliente, muy rápido)	acelerar	150	(muy caliente, muy rápido)
(muy caliente, muy rápido)	acelerar	-100	(fundido, detenido)

Métodos de diferencias temporales

El piloto, no muy contento con los resultados y el hecho de tener que esperar a jugar todo un episodio para ir actualizando los valores de sus estados, decide incursionar en técnicas que le permitan, haciendo uso de bootstrapping, mejorar cada vez que ejecuta una acción las estimaciones de sus valores $Q^{\pi}(s, a)$.

Para ello se embarca en esta tarea, teniendo inicializados los valores Q de sus estados con los valores del ejercicio anterior o 0 si nunca los visitó.

Asumiendo que está en el estado **(caliente, muy rápido)** y una estrategia epsilon greedy le dice que tiene que acelerar, cuanto vale:

1) $Q^{\pi}((caliente, muyrápido), acelerar)$ de acuerdo con SARSA (si al seleccionar "epsilon-greedy-mente" una acción siguiente esta es acelerar).
2) $Q^{\pi}((caliente, muyrápido), acelerar)$ de acuerdo con QLearning.

Notas:

- Para ambas partes, $S' = (muycaliente, muyrápido)$ y $R = 120$.
- Los valores de γ y α se asumen 1.