UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Guatemala 16 de de 2021
Paola De León - 20361

Reinforcement Learning

Laboratorio 1

Task 1

1. **¿Qué pasa si algunas acciones tienen probabilidades de cero?**

El que hayan acciones con probabilidad de cero significa que estas nunca serán seleccionadas pues no tendrán relevancia a nivel de las reglas de la política, ni forman parte de la acción óptima. Implicando que son acciones que nunca se explotarán y por lo tanto afectarán el aprendizaje del agente pues sus acciones estarán sesgadas y este incluso puede llegar a converger en un óptimo local, donde cree que ya encontró la mejor estrategia, pero en realidad no ha explorado todas las opciones posibles.

2. **¿Qué pasa si la póliza es determinística?**
   a. $\pi_1(a) = 1$ **para algún** a**.**

Si una política es determinística, es decir, selecciona siempre una acción fija para un estado dado $\pi_1(a) = 1$ para alguna acción a, el agente no tendrá la oportunidad de explorar otras posibles acciones. Esto puede llevar a un aprendizaje subóptimo, ya que el agente nunca descubrirá si otras acciones pueden resultar en mejores recompensas a largo plazo. La falta de exploración significa que el agente podría quedar atrapado en un óptimo local, donde la solución encontrada no es la mejor posible, pero no tiene manera de descubrir alternativas debido a su naturaleza determinística.

Además, una política determinística depende en gran medida de la inicialización de los valores de acción. Si los valores iniciales están mal definidos, el agente podría seleccionar consistentemente acciones subóptimas. Para mitigar estos problemas, se suelen utilizar estrategias de exploración como Epsilon-greedy o Boltzmann exploration, que permiten al agente probar acciones diferentes y ajustar su política de manera más efectiva.

3. **Investigue y defina a qué se le conoce como cada uno de los siguientes términos, asegúrese de definir qué consiste cada una de estas variaciones y cómo difieren de los k-armed bandits**
    a. Contextual bandits
        i. Los contextual bandits son una extensión de los k-armed bandits donde el agente tiene acceso a información adicional (contexto) antes de tomar una acción. Este contexto puede ser cualquier información relevante del entorno que puede influir en la recompensa asociada con cada acción. En lugar de seleccionar una acción basada únicamente en las recompensas históricas, el agente utiliza el contexto para tomar decisiones más informadas.
        ii. La diferencia clave con los k-armed bandits es la inclusión de este contexto, que permite una adaptación más fina a diferentes situaciones, mejorando potencialmente la eficiencia de aprendizaje.
    b. Dueling bandits
        i. Los dueling bandits son una variante donde, en lugar de obtener una recompensa para una única acción, el agente selecciona pares de acciones y recibe información comparativa sobre cuál de las dos es mejor. Este enfoque se utiliza cuando las recompensas absolutas son difíciles de definir o medir, pero es más factible determinar cuál de dos acciones es superior.
        ii. La diferencia con los k-armed bandits radica en la estructura de feedback: en lugar de una recompensa cuantitativa para una acción específica, se obtiene una preferencia relativa entre dos acciones.
    c. Combination bandits
        i. Los combinatorial bandits se aplican en situaciones donde el agente debe elegir combinaciones de acciones en lugar de una sola acción. En este modelo, las recompensas se obtienen a partir de combinaciones específicas de acciones seleccionadas simultáneamente. Este enfoque es relevante en contextos donde las interacciones entre múltiples acciones influyen en la recompensa total.
        ii. La diferencia principal con los k-armed bandits es que en lugar de evaluar acciones individuales, el agente evalúa combinaciones de acciones, lo cual aumenta significativamente la complejidad del espacio de búsqueda y el proceso de toma de decisiones.