UVG
UNIVERSIDAD DEL VALLE DE GUATEMALA

Reinforcement Learning
- Laboratorio 1 -

Semestre II - 2024

Instrucciones:

- Esta es una actividad en grupos de 3 personas máximo
- No se permitirá ni se aceptará cualquier indicio de copia. De presentarse, se procederá según el reglamento correspondiente.
- Tendrán hasta el día indicado en Canvas.

Task 1

Responda a cada de las siguientes preguntas de forma clara y lo más completamente posible.

1. ¿Qué pasa si algunas acciones tienen probabilidades de cero?
2. ¿Qué pasa si la póliza es determinística?
    a. $\pi_1(a) = 1$ para algún a
3. Investigue y defina a qué se le conoce como cada uno de los siguientes términos, asegúrese de definir en qué consiste cada una de estas variaciones y cómo difieren de los k-armed bandits
    a. Contextual bandits
    b. Dueling bandits
    c. Combination bandits

Task 2

Implemente el algoritmo épsilon-codicioso para el problema de multi-armed bandits para maximizar la recompensa acumulativa en una serie de pruebas. Utilice un entorno con 10 brazos, cada uno de los cuales proporciona una recompensa de una distribución de probabilidad fija diferente. Para esto considere el siguiente set de instrucciones

1. Crea una clase *Bandit* para representar el entorno. Esta clase debería inicializar 10 brazos, cada uno con una probabilidad de recompensa elegida al azar entre 0 y 1.
2. La clase debe tener un método *pull(arm)* que devuelva una recompensa de 1 con la probabilidad específica del brazo elegido y 0 en caso contrario.
3. Cree una clase de *Agent* para implementar la estrategia de épsilon-greedy.
4. Inicialice el agente con un valor épsilon específico para la exploración, una matriz para almacenar las recompensas estimadas para cada brazo (inicializada en cero) y una matriz para contar la cantidad de veces que se ha extraído cada brazo.
5. Implemente un método *select_arm()* en la clase Agent que:
    a. Con probabilidad épsilon, selecciona un brazo aleatorio.
    b. Con probabilidad $1-\epsilon$, selecciona el brazo con la recompensa estimada más alta.
6. Implemente un método *update_estimates(arm, recompensa)* para actualizar la recompensa estimada para el brazo elegido usando la fórmula vista en clase
7. Inicialice *Bandit* y *Agent* con épsilon configurado en 0.1.
8. Ejecute la simulación para 1,000 iteraciones.
9. En cada prueba, seleccione un brazo usando *select_arm*(), tire del brazo en el entorno Bandit para obtener una recompensa y actualice las recompensas estimadas usando *update_estimates*().
10. Realice un seguimiento e imprima la recompensa acumulada al final de la simulación.
11. Grafique la recompensa acumulada en las pruebas para visualizar la mejora del desempeño del agente.
12. Grafique los valores estimados de cada brazo versus las probabilidades reales para evaluar la precisión de las estimaciones.

UVG
UNIVERSIDAD DEL VALLE DE GUATEMALA

Reinforcement Learning
- Laboratorio 1 -

Semestre II - 2024

Tareas a realizar:

1. Implemente las clases *Bandit* y *Agent* siguiendo la estructura proporcionada.
2. Ejecute la simulación y observe la recompensa acumulada y los valores estimados de cada brazo.
3. Experimente con diferentes valores de épsilon (por ejemplo, 0,01, 0,5) y observe el efecto en el equilibrio de exploración y explotación.
4. Trace la recompensa acumulada en las pruebas para visualizar el progreso de aprendizaje del agente.
5. Trazar los valores estimados de cada brazo versus las probabilidades reales para evaluar la precisión de las estimaciones.

Entregas en Canvas

1. Documento PDF con las respuestas a cada task

Evaluación

1. [1 pts] Task 1 (0.2 cada pregunta)
2. [4 pts] Task 2