TD-Camino más corto

Ejercicio 1: SARSA

Supongamos que tienes un agente que se mueve en una cuadrícula de 4x4. El objetivo del agente es llegar a la esquina inferior derecha partiendo de la esquina superior izquierda. Las recompensas son -1 por cada movimiento, excepto los movimientos que llevan al estado objetivo, que tienen una recompensa de 0. Los movimientos posibles en cada estado son arriba, abajo, izquierda y derecha. Los movimientos que llevarían al agente fuera de la cuadrícula lo dejan en su posición actual. Utiliza el enfoque de SARSA para encontrar la política óptima, asumiendo una política inicial aleatoria y los siguientes parámetros:

- Tasa de aprendizaje (α): 0.1
- Factor de descuento (γ): 0.9
- Política de elección de acciones: ε-greedy con $\epsilon = 0.1$

Realiza una iteración de actualización para una secuencia de estados-acción-recompensa que elijas, comenzando desde el estado inicial.

Ejercicio 2: Q-Learning

Utilizando el mismo escenario de la cuadrícula de 4x4 del ejercicio 1, aplica esta vez el algoritmo de Q-Learning para encontrar la política óptima. Al igual que antes, las recompensas son -1 por cada movimiento, excepto los movimientos que llevan al estado objetivo, que tienen una recompensa de 0, y los movimientos posibles en cada estado son arriba, abajo, izquierda y derecha, con la misma regla para movimientos fuera de la cuadrícula.

Utiliza los mismos parámetros:

- Tasa de aprendizaje (α): 0.1
- Factor de descuento (γ): 0.9

Sin embargo, esta vez, asume que siempre se selecciona la acción con el mayor Q-value (política greedy). Realiza una iteración de actualización para una secuencia de estados y acciones que elijas, comenzando desde cualquier estado.

Ejercicio 3: Evitando el Abismo

Utiliza la misma cuadrícula de 4x4 de los ejercicios anteriores, pero esta vez añade un "abismo" en la celda (3, 2) (utilizando la notación (fila, columna) con base en 1). La

recompensa de caer en el abismo es -10, mientras que alcanzar el estado objetivo sigue teniendo una recompensa de 0, y moverse incurre en una recompensa de -1 como antes. Las reglas de movimiento siguen siendo las mismas.

Parte A: SARSA con Abismo

Realiza una iteración de actualización del algoritmo SARSA teniendo en cuenta el abismo. Supongamos que la política inicial lleva al agente a una ruta que incluye el abismo:

1. Describe el efecto esperado del abismo en la política aprendida con SARSA, considerando que SARSA es un algoritmo on-policy.
2. Realiza y describe una iteración de actualización específica que involucre el abismo, usando los mismos parámetros ($\alpha = 0.1$, $\gamma = 0.9$, ε-greedy con $\epsilon = 0.1$).

Parte B: Q-Learning con Abismo

Ahora aplica el algoritmo Q-Learning teniendo en cuenta el abismo:

1. Describe cómo el abismo afectaría la política aprendida con Q-Learning en comparación con SARSA, especialmente en términos de exploración y explotación.
2. Realiza y describe una iteración de actualización específica que involucre el abismo, asumiendo una selección greedy de acciones basada en los Q-values, con los mismos parámetros de aprendizaje.

Reflexión sobre el Abismo

- **Comparación de Estrategias**: Después de completar ambas partes, reflexiona sobre cómo el abismo afecta las decisiones de política en SARSA y Q-Learning. Considera cómo cada algoritmo se adapta a las recompensas negativas severas y qué esto puede decir sobre su uso en entornos con penalizaciones significativas.
- **Riesgo vs. Seguridad**: Piensa en cómo la presencia del abismo puede cambiar la estrategia de exploración del agente. ¿El agente se vuelve más cauteloso con SARSA en comparación con Q-Learning, o viceversa?

Para más info:

- Example 6.6: Cliff Walking (cap 6.5, Reinforcement Learning. An Introduction", R.S. Sutton & A.G. Barto (2018)
- Temporal Difference Learning (including Q-Learning) | Reinforcement Learning Part 4