Facultad de prácticas de IA . de Matemáticas e Informática
Universitat de Barcelona

# Práctica 3: Aprendizaje por refuerzo

El objetivo de esta práctica es familiarizarnos con Q-learning, un algoritmo básico de aprendizaje por refuerzo (RL). Los algoritmos RL son una familia de métodos potentes para resolver problemas complejos de forma semisupervisada. Es decir, a diferencia de las redes neuronales, que aprenden a partir de ejemplos, el RL se produce mediante una señal escalar (recompensa).

Como regla general, se puede considerar que cualquier implementación de RL debe cumplir varios requisitos:

- En primer lugar, el problema se define como un problema de decisión de Markov, donde las acciones del agente en general no determinan el resultado (pueden tener un resultado estocástico).

- En segundo lugar, la definición del estado es crucial, ya que puede alterar enormemente el problema, facilitando su solución o volviéndolo intratable. Por ello, conviene dedicar algo de tiempo a pensar cuál es la mejor definición de estado que se puede utilizar para nuestro problema específico. Por ejemplo, en el caso del ajedrez (como el de nuestra práctica), el problema consiste en encontrar una política que lleve al jaque mate. Como estado, podríamos utilizar una lista que contenga las posiciones y tipos de piezas, o el tablero entero.
  El problema puede resolverse en cualquiera de los casos, pero la cantidad de memoria y cálculos necesarios varía en gran medida en cada caso.
- En tercer lugar, resolver el problema equivale a encontrar la política correcta para realizar la transición entre estados, partiendo de un estado inicial hacia un estado final (que no necesariamente se conoce de antemano).
- En cuarto lugar, dado que la política “correcta” (óptima) es la que conduce a la máxima recompensa, la forma en que asignamos la recompensa (la llamada función de recompensa) también es crucial para encontrar el estado deseado.
- Cinco, los algoritmos de RL aprenden sobre la base de una evaluación de ensayo y error, cuantificada en términos de recompensa. En otras palabras, el aprendizaje se produce cuando la predicción de la recompensa esperada asociada a una transición particular difiere de la recompensa real obtenida al ejecutar la transición. Por el contrario, se dice que el algoritmo converge cuando las recompensas obtenidas se aproximan a los valores esperados.
- Sexto, la definición de la función de asignación de recompensa (la que cuantifica la recompensa asociada a cada transición) a menudo requiere una introspección importante, ya que puede arrojar soluciones radicalmente diferentes para el mismo problema. Por ejemplo, si se quiere encontrar el camino más corto desde el punto A al punto B en una cuadrícula, se puede asignar un valor negativo en cada paso intermedio y un valor grande al llegar al punto B.

## El simulador

El mismo simulador que utilizaste para las prácticas anteriores, que consta de cuatro clases jerárquicas: Aichess, Ajedrez, Tablero y Pieza. Las tres últimas implementan la dinámica de una partida de ajedrez, que puede ser ejecutada por dos jugadores (ejecuta la clase principal en la clase Ajedrez para este fin). La clase Aichess es una cápsula de las otras tres, con el propósito de implementar algoritmos de IA para analizar y alterar la dinámica de la partida de ajedrez.

Práctica 3

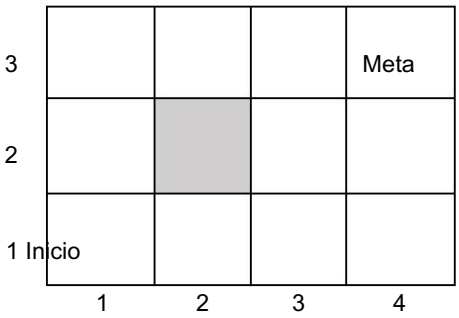
Facultad de prácticas de IA . de Matemáticas e Informática
Universitat de Barcelona

La definición del Estado

Aunque tienes plena autonomía en cuanto a la definición concreta de estado, te sugerimos que comiences definiendo el estado en función de las posiciones y tipos de piezas, tal y como hiciste en las prácticas anteriores.

Tu trabajo

1. (4 pts) Comenzaremos abordando el sencillo problema visto en clase de encontrar un camino desde el inicio hasta la meta en el siguiente escenario:




Recuerde proporcionar la primera tabla Q, las dos intermedias y la tabla Q final en todos los casos.

a. Implemente el algoritmo Q-learning para encontrar el camino óptimo considerando una recompensa de -1 en todas partes excepto en el objetivo, con recompensa 100.

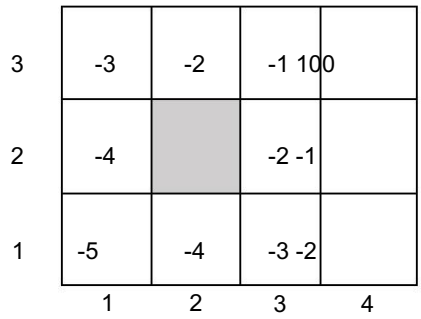
i. (0,4 pts) Imprima la primera tabla Q, las dos intermedias y la tabla Q final. ¿Qué secuencia de acciones obtiene?

ii. (0,4 pts) Después de probar un poco, ¿cuál es su elección de parámetro para alfa? ¿Gamma y épsilon? ¿Por qué?

iii. (0,4 pts) ¿Cómo se juzga la convergencia del algoritmo? ¿Cuánto tiempo

¿Cuánto tiempo se tarda en converger?

b. Intente implementar la recompensa más precisa que se ofrece de la siguiente manera:




que es una función de la distancia a la meta.

i. (0,4 pts) Responda las preguntas de la sección anterior para este caso. ii. (0,4 pts) ¿Cuál es el efecto de la nueva función de recompensa sobre el desempeño?

iii. (0,4 pts) ¿Cómo se relaciona esto con los algoritmos de búsqueda estudiados en P1? ¿Podrías aplicar uno de estos en este caso?

c. La principal novedad de los algoritmos RL con respecto a los algoritmos de búsqueda en P1 es que pueden aplicarse en entornos estocásticos, donde el agente no determina totalmente el resultado de sus acciones.

i. (0,6 pts) Marinero borracho. Tu agente es ahora un marinero borracho que intenta irse a la cama después de una buena ración de whisky y mariscadas: sus piernas no parecen obedecerle todo el tiempo. Introduce la estocasticidad (= aleatoriedad) imponiendo que solo el 99% de los pasos previstos por el marinero se realicen realmente, y el resto conduzca aleatoriamente en cualquier otra dirección posible.

ii. (1 pts) Utilice al menos una de las dos recompensas propuestas:
1. (0,15 pts) ¿Cuál es tu elección de parámetro? ¿Por qué?
2. (0,15 pts) Suponiendo que el marinero se encuentra en un estado que le permite aprender: ¿cuántas noches de borrachera son necesarias para que domine el peligroso camino hacia la cama? Compárese con el escenario determinista anterior.
3. (0,2 pts) ¿Cuál es el camino óptimo encontrado? Si viéramos al marinero intentar seguirlo, ¿seguiría siempre el mismo camino?
4. (0,5 pts) ¿Podríamos aplicar uno de los algoritmos de P1 aquí? ¿Por qué? Sugerencia: piense en las nociones de determinismo versus azar y de trayectoria. vs política.

2. (6+1 pts) Ahora, volvamos al escenario de ajedrez, es decir, a la primera configuración del tablero de P1. Recuerde que tenemos el rey negro, el rey blanco y una torre blanca, y que solo las blancas mueven. Recuerde proporcionar la primera tabla de damas, las dos intermedias y la tabla de damas final en todos los casos.

a. Adapte su implementación de Q-learning para encontrar el camino óptimo hacia un jaque mate considerando una recompensa de -1 en todas partes excepto en el objetivo (jaque mate para las blancas), con recompensa 100.

i. (0,5 pts) ¿Qué secuencia de acciones obtienes? ii. (0,5 pts) Después de probar un poco, ¿cuál es tu elección de parámetro para alfa? ¿Gamma y épsilon? ¿Por qué?
iii. (0,5 pts) ¿Cómo se juzga la convergencia del algoritmo? ¿Cuánto tiempo ¿Cuánto tiempo se tarda en converger?

b. Pruebe ahora con una función de recompensa más sensata adaptada de la heurística utilizada para la búsqueda A*:

i. (0,5 pts) Responda las preguntas de la sección anterior para este caso. ii. (0,5 pts) ¿Cuál es el efecto de la nueva función de recompensa sobre el desempeño?

c. Marinero borracho. De camino a la cama, nuestro marinero borracho ve un tablero de ajedrez sobre una mesa, casualmente configurado como en la sección anterior. Ha visto al capitán jugar con el primer oficial y quiere intentarlo, pero sólo tiene un conocimiento rudimentario de las reglas (sabe cómo se mueve cada pieza y qué es un jaque mate, pero no que las negras también se mueven).

i. (0,5 pts) Introduzca estocasticidad (= aleatoriedad) imponiendo que solo se realice un porcentaje determinado de los movimientos previstos por el navegante, y el resto se tome aleatoriamente entre todas las demás posibilidades.

ii. (1 pts) Utilice cualquier recompensa que prefiera:

1. (0,5 pts) ¿Cuál es tu elección de parámetro? ¿Por qué?
2. (0,5 pts) Suponiendo que nuestro marinero obsesivo se encuentra en un estado que le permite aprender: ¿cuántas partidas tiene que jugar antes de estar satisfecho de haber encontrado la mejor estrategia y poder irse a dormir? Compárese con el escenario determinista anterior.
3. (0,5 pts) ¿Cuál es el camino óptimo encontrado? Si viéramos al marinero intentar seguirlo, ¿seguiría siempre el mismo camino?

d. (0,5 pts) Compare la aplicación del aprendizaje Q en este escenario de ajedrez con la de la cuadrícula del ejercicio 1. ¿En qué se diferencian los dos escenarios? ¿Cómo se traduce eso en sus resultados?

e. (1 pts) Compare el uso de Q-learning con el de los algoritmos de búsqueda de P1 para el escenario de ajedrez visto aquí, tanto en el caso determinista como en el estocástico.

f. (Voluntario: 1 pts) Utilice su implementación de aprendizaje cuantitativo del marinero borracho en la segunda configuración de placa de P1 e intente encontrar la mejor combinación de parámetros para una convergencia rápida. ¿Qué tan sólida fue su elección de parámetros anterior? Analice sus resultados.