Práctica final – Aplicación de conceptos generales de RL a varios entornos.
Sesión segunda. Operación en almacén

1. Introducción

En esta sesión, tras habernos iniciado en el RL con aproximación de función en un entorno muy simple, vamos a complicar progresivamente el entorno con el que interactuamos.

Nos moveremos en el mismo recinto, con los mismos obstáculos, que ahora van a representar estanterías donde puede haber objetos. Esto da lugar a un entorno de almacén como el representado en la Figura 1. El agente se representa en color naranja y la zona en verde representa la zona de entrega. Aparecen en azul tres objetos, uno por estantería.

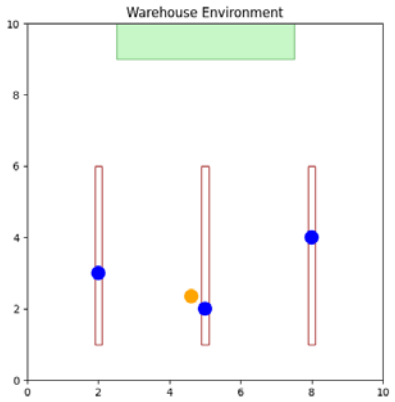

Figura 1: El entorno almacén

2. Trabajo previo

Simplemente lee este documento. Asegúrate de que comprendes cuáles son las acciones que puede elegir el agente en cada uno de los entornos propuestos. En base a tus conclusiones sobre la sesión anterior, reflexiona sobre cuáles serían diseños adecuados de la observación también para cada entorno.

No hay que realizar ninguna entrega. Simplemente plantearemos un breve debate al inicio de la sesión entre todos sobre las posibilidades para cada situación.

3. Entornos

Trabajaremos sobre tres variantes del entorno que son, en orden creciente de complejidad:

- **Entorno 1: Objetos fijos, recogida de un objeto**. En este caso, los objetos están en tres posiciones fijas. El agente debe aproximarse a uno cualquiera de ellos e implementar la recogida del objeto de forma exitosa. La acción de entrega no tiene nada asociado, por lo que podemos obviarla. El episodio termina (con fracaso) si el agente toca las paredes del perímetro o alguna estantería y también (con éxito) en el momento en el que el agente recoge un objeto.
- **Entorno 2: Objetos fijos, recogida y entrega de un objeto**. En este caso, los objetos están en tres posiciones fijas. El objetivo del agente es entregar un objeto en la zona de entrega. Lógicamente, para ello tiene que haber cogido un objeto antes. El episodio termina con fracaso si el agente toca las paredes del perímetro o alguna estantería o si suelta el objeto fuera del área de recogida, y con éxito en el momento en el que el agente recoge un objeto.
- **Entorno 3: Objetos en posición aleatoria, recogida y entrega de un objeto.** Este entorno es igual que Entorno 2, pero la posición de los objetos en las estanterías cambia de episodio a episodio.

Detalles "finos" de los entornos

Como puede observarse en la *Figura 1*, el entorno tiene forma cuadrada con 10 metros de lado. El área objetivo es un rectángulo cuyo vértice inferior izquierdo se encuentra situado en (x = 2.5, y = 9), con dimensiones (l_x = 5, l_y = 1). Por otro lado, hay tres estanterías con dimensiones (l_x = 1, l_y = 5) y cuyos vértices inferiores izquierdos se encuentran situados respectivamente en (x = 1.5, y = 1), (x = 4.5, y = 1) y (x = 7.5, y = 1).

El entorno devuelve una observación con la siguiente estructura:

- obs[0]: posición del agente en el eje x.
- obs[1]: posición del agente en el eje y.
- obs[2]: posición del objeto 1 en el eje x.
- obs[3]: posición del objeto 1 en el eje y.
- obs[4]: posición del objeto 2 en el eje x.
- obs[5]: posición del objeto 2 en el eje y.
- obs[6]: posición del objeto 3 en el eje x.
- obs[7]: posición del objeto 3 en el eje y.
- obs[8]: *agent_has_object*. Representa si el agente porta un objeto o no.
- obs[9]: *collision*. A 1 si el agente ha chocado con pared o estantería.
- obs[10]: *delivery*. A 1 si el agente suelta un objeto en la zona de entrega.

El entorno recibe una acción del agente que, para todos los estados, puede ser $\mathcal{A}(s) = \{arriba, abajo, izquierda, derecha, coger, soltar\}$. Entonces, si el agente elige una acción de movimiento, se desplaza 0.25 metros en la dirección elegida; y si elige una acción sobre un objeto, aplica lo siguiente:

- coger: si el agente elige esta acción y hay un objeto a menos de 30 cm, entonces recoge con éxito dicho objeto, lo que se reflejará en un flanco a 1 en la variable *agent_has_object*. A partir de este momento, salvo que termine el episodio, la posición de ese objeto será la del agente.
- soltar: sólo aplica a los entornos 2 y 3. En estos, hay un flanco a 0 de la variable *agent_has_object*, aunque importa poco, dado que en todos los casos (fracaso o éxito) se termina el episodio.

4. Ingeniería de variables, diseño de la recompensa y entrenamiento de los agentes

Todo esto lo tienes que resolver apoyándote en lo desarrollado en la sesión anterior y con los nuevos conocimientos adquiridos en las sesiones de teoría.

La única cuestión relevante para poner en marcha los procesos es que hay que pasar unos argumentos al constructor de la clase **WarehouseEnv()** que tenéis en el script almacen_all.py que os hemos compartido. En concreto, hay que especificar dos variables al instanciar la clase, que son autoexplicativas: random_objects=True/False y drop=True/False. La relación entre estas binarias y los entornos es:

- Entorno 1: random_objects = False, drop = False
- Entorno 2: random_objects = False, drop = True
- Entorno 3: random_objects=True, drop = True

5. Evaluación de los agentes

Vamos a medir, para cada entorno, el rendimiento de dos tipos de agente: 1) un agente entrenado con sólo 10000 episodios de experiencia y 2) un agente entrenado con toda la experiencia que consideres oportuna.
En ambos casos, el banco de ensayos constará de 1000 episodios con punto de partida pseudoaleatorio (por lo tanto, el mismo para todos los agentes).
La evaluación la haremos los profesores (no conocerás esas posiciones iniciales) una sola vez, con los dos agentes que nos envíes.

6. Memoria de trabajo

Aunque este trabajo lo vamos a evaluar en la presentación final, debes recoger en un cuaderno de trabajo todo lo que vayas haciendo. Eventualmente podemos pedírtelo para comprobar alguno de los resultados/desarrollos tras la presentación final.