

## 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 es un Markov Decision Process (MDP)?
2. ¿Cuáles son los componentes principales de un MDP?
3. ¿Cuál es el objetivo principal del aprendizaje por refuerzo con MDPs?

## Task 2

El objetivo principal de este ejercicio es que simule un MDP que represente un robot que navega por un laberinto de cuadrículas de 3x3 y evalúe una política determinada.

Por ello considere, a un robot navega por un laberinto de cuadrícula de 3x3. El robot puede moverse en cuatro direcciones: arriba, abajo, izquierda y derecha. El objetivo es navegar desde la posición inicial hasta la posición de meta evitando obstáculos. El robot recibe una recompensa cuando alcanza la meta y una penalización si choca con un obstáculo.

El laberinto es el siguiente

	S			G	
		X			
			X		

Donde:

- S = punto de inicio
- G = punto de meta
- X = son obstáculos

Instrucciones:

- Defina los componentes del MDP:
  - Estados:  $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$ , donde cada número representa una celda del laberinto.
  - Acciones:  $A = \{\text{arriba, abajo, izquierda, derecha}\}$
  - Probabilidades de transición:  $P(s' | s, a)$
  - Recompensas:  $R(s, a, s')$
- Matriz de transición:
  - Defina las probabilidades de transición  $P$  como un diccionario donde  $P[s][a]$  asigna los siguientes estados  $s'$  a sus probabilidades.
- Función de recompensa:
  - Defina las recompensas  $R$  como un diccionario donde  $R[s][a][s']$  da la recompensa por la transición del estado  $s$  al estado  $s'$  mediante la acción  $a$ .
- Política:

- Defina una política  $\pi$  como un diccionario que asigna cada estado a una acción.
- Simular la política:
  - Escriba una función para simular la política en el MDP para una cierta cantidad de pasos.
  - Realice un seguimiento de la recompensa acumulada obtenida siguiendo la política.
- Evaluar la Política:
  - Simule la política varias veces para estimar la recompensa acumulada promedio.

### Task 3

En clase hemos dicho que una vez tengamos  $v^*$  o  $q^*$  sabemos la política óptima  $\pi^*$  ¿Por qué?  
Puede consultar el libro en la sección 3.8 en adelante

### Entregas en Canvas

1. Documento PDF con las respuestas a cada task
2. Código de la implementación del Task 2
  - a. Si trabaja con JN deje evidencia de la última ejecución
  - b. Caso contrario, deje en comentarios el valor resultante

### Evaluación

1. [0.75 pts] Task 1 (0.25 cada pregunta)
2. [3.25 pts] Task 2
3. [1 pts] Task 3