Machine Learning

Examen #3 Entrega: Diciembre 15 de 2019 11:59 p.m.

Considere el juego en el que se quiere alcanzar una de las casillas marcadas con **W**, en el siguiente tablero:

	7			**L**		-4			**W**		**L**		**W**

El juego opera de la siguiente forma:

- Inicialmente el jugador pone una ficha en la primera casilla a la izquierda.
- En cada jugada, el jugador decide si la ficha se va a mover hacia la izquierda o hacia la derecha. A continuación lanza un dado, y se desplaza el número de casillas indicado por el dado en la dirección que ha decidido antes de lanzar el dado.
- Si la ficha cae en una casilla donde hay un número, la ficha avanza ese número de casillas si el número es positivo o retrocede ese número de casillas si el número es negativo.
- Si la ficha cae en una casilla marcada con **L**, el jugador pierde.
- Si la ficha cae en una casilla marcada con **W**, el jugador gana
- En los extremos, la ficha rebota (si se supera el extremo, se avanza en la otra dirección la cantidad restante).

Se proveen 5 instancias de este juego, con 100 casillas en los archivos `m1.csv`, `m2.csv`, `m3.csv`, `m4.csv`, `m5.csv`, codificados así:

- 0: casilla vacía.
- $-1000 = \mathbf{L}$
- $1000 = \mathbf{W}$
- $\pm n \neq \pm 1000, 0$, avaance de $\pm n$.

Usted debe usar el archivo indicado en la siguiente tabla:

Archivo	Utimo dígito de su código de estudiante
`m1.csv`	0,1
`m2.csv`	2,3
`m3.tcsv`	4,5
`m4.csv`	6,7
`m5.csv`	8,9

1. Plantee este problema como un proceso de decisión de Markov (MDP). Especifique *todos* los elementos de su MDP.

2. Encuentre una política óptima que maximice la probabilidad de ganar. Note que dado que se conoce por completo la dinámica del ambiente, es posible encontrar exactamente una política óptima usando por ejemplo programación dinámica o value iteration.

3. Utilice Q learning para encontrar una política óptima. Evalúe el avance de su algoritmo con respecto a la política encontrada en la pregunta anterior.

4. Evalúe la política encontrada por Q Learning en 10.000 episodios del juego. Estime la probabilidad de ganar de su política y la duración esperada de un episodio.

En todos los casos usted debe elaborar su propio código, y usar la versión tabular de los algoritmos.