Aprendizaje por refuerzo
2023-II

Nombre:

5 de junio de 2023

Instrucciones: Para cada problema conteste lo que se le pide.

1. (20 puntos)**Básico I:** Explique lo siguiente
   - (4 puntos) Describa los elementos de un problema de aprendizaje por refuerzo
   - (4 puntos) ¿En qué se diferencia a un problema de aprendizaje supervisado o no supervisado?
   - (4 puntos) ¿Cuáles son los problemas centrales en aprendizaje por refuerzo?
   - (4 puntos) ¿Cuál es la diferencia entre los métodos en política y fuera de política?'
   - (4 puntos) ¿Cuál es la diferencia entre los métodos de aprendizaje Q y gradiente de política?'
2. (20 puntos)**Programación dinámica:** Usted se encuentra en un casino, tiene \$200.00 para este ejercicio y jugará hasta que pierda todo o hasta que duplique su dinero. Puede elegir entre dos juegos. El primero A, cuesta \$100.00 y da como retorno \$200.00 con probabilidad 0.05 y \$0.00 de lo contrario. El primero A, cuesta \$200.00 y da como retorno \$300.00 con probabilidad 0.1 y \$0.00 de lo contrario. Hasta terminar, se elegirá jugar el juego A o B.
   - (10 puntos) Provea un MDP que describa la situación
   - (5 puntos) Aplique el algoritmo de iteración de valor para resolver el problema
   - (5 puntos) ¿Cuál es la política óptima?
3. (15 puntos)**RL libre de modelo:** un agente explora un MDP $M = (S, A, R, P, \gamma)$ donde $S = \{s_1, s_2, s_3\}$ y $A = \{a_1, a_2, a_3\}$, $\gamma = 0{,}5$ y $P(s, a_i, s_i) = 1$ para cualquier s para todo i. Las recompensas por transicionar a un estado son definidas como $R(s_i) = i$. La recompensa máxima es de 3.
   - (5 puntos) Dibuje en un diagrama el MDP propuesto
   - (5 puntos) El agente sigue la trayectoria: $(s_1, a_1, 1, s_1, a_2, 2, s_2)$. Considere aprendizaje Q utilizando ϵ-vóraz, donde una acción aleatoria nunca elije la acción vóraz, los empates se deciden eligiendo a_i con

la menor i. El factor de aprendizaje $\alpha = 0{,}5$. Q se inicializa con ceros. ¿Podría el agente generar esa trayectoria para $\epsilon \neq 0$? de ser así, etiquete las acciones que son voraces y aquellas que son aleatorias

- (5 puntos) Considere un algoritmo de inicialización llamado $Rmax$ donde todos los estados son inicializados con 6. ¿podría haber generado este método la trayectoria del inciso anterior? ¿por qué?

4. (15 puntos)**Aproximación de funciones:**

   - (5 puntos) Explique el algoritmo de DDQN y como se diferencia de DQN
   - (5 puntos) ¿Cuál es el rol de la repetición de experiencia? ¿en cuáles métodos se utiliza? ¿Cuál es la alternativa?
   - (5 puntos) Un ratón es involucrado en un experimento. Experimenta un episodio, en el primer paso escucha una campana. En el segundo paso ve una luz. En el tercer paso escucha una campana y ve una luz. Después recibe un pedazo de queso que vale 1 de recompensa y el episodio termina. Todas las otras recompensas fueron cero y el experimento no tiene descuento. Representar el estado del ratón s por dos características $bell(s) \in 0, 1$ y $light(s) \in 0, 1$. Aplique el algoritmo de Q-learning para actualizar la tabla Q inicializanda en ceros.

5. (15 puntos)**Exploración y explotación:**

   - (5 puntos) ¿Por qué es díficil el problema de exploración y explotación?
   - (5 puntos) ¿Describa el método de UCB?
   - (5 puntos) Suponga que tenemos un problema de bandido multibrazo donde existen 3 acciones disponibles. Siguiendo el principio de optimismo bajo incertidumbre, ¿Cuáles serían las primeras 4 acciones a tomar? y ¿por qué?

6. (15 puntos)**Temas avanzados:**

   - (5 puntos) Considere el siguiente problema y utilice el método de métricas ponderada con $w = [0{,}5, 0{,}5]^T$, $p = 2$ y encontrar el óptimo

$$\min_{x \in [-10,10]^2} (x_1^2 + x_2^2, (x_1 - 5)^2 + (x_2 - 5)^2)$$

   - (5 puntos) Lea y critique el siguiente artículo CS de Witt et al. (2020). Deep Multi-Agent Reinforcement Learning for Decentralized Continuous Cooperative Control
   - (5 puntos) Lea y critique el siguiente artículo A Mahajan et al. (2022). Generalization in Cooperative Multi-Agent Systems