Aprendizaje por refuerzo
Licenciatura en ciencia de datos
2023-II

Nombre:

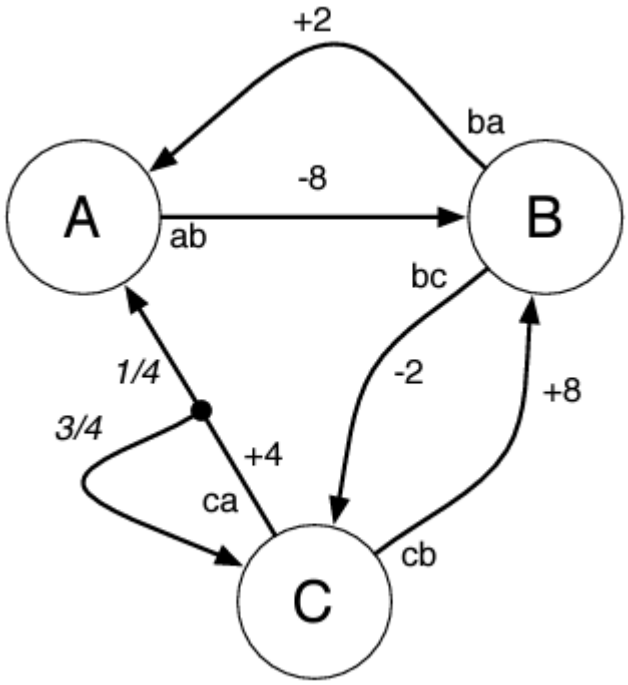
29 de marzo de 2023

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

1. (25 puntos)**Básico I:** Explique lo siguiente
   - (5 puntos) ¿Cuál es el objetivo de aprendizaje por refuerzo?
   - (5 puntos) Explique con sus propias palabras que es un MDP
   - (5 puntos) ¿Qué representan los estado-valor y estado acción?
   - (5 puntos) ¿Cuál es la diferencia entre explorar y explotar?
   - (5 puntos) Explique el principio de control de Monte-Carlos, SARSA y aprendizaje Q
2. (25 puntos) **Básico II:** Considere una tienda de sándwiches en el IIMAS para la población de N personas. Los clientes llegan a tiempos gobernados por una distribución de probabilidad desconocida. Cada cliente puede ordenar un sándwich con un cierto tipo de pan (5 opciones) y relleno (4 tipos). Los clientes pagan un precio dado por cada sándwich.

   Si un cliente no puede obtener el sándwich de su elección, nunca regresará. Los ingredientes deben ser descartados 3 días después de su compra. El dueño quiere encontrar una política para comprar los ingredientes de forma que maximice su ganancia a largo plazo.
   - (5 puntos) Defina los estados, acciones y recompensas
   - (5 puntos) ¿Utilizaría una versión con o sin descuento? y ¿por qué?
   - (5 puntos) ¿Utilizaría planeación o aprendizaje por refuerzo? y ¿por qué?
   - (5 puntos) Entre Monte-Carlo y aprendizaje de diferencia temporal, ¿cuál método preferiría?
   - (5 puntos) ¿Es necesario utilizar aproximación de funciones? ¿por qué?
3. (25 puntos) **Programación dinámica:** Considere el siguiente MDP con un factor de descuento $\gamma = 0{,}5$. Las letras A, B, C representan estados; los arcos transiciones: las letras ab, ba, bc, ca, cb representan acciones; los

enteros con signo representan recompensas; y las fracciones representan probabilidades



- (5 puntos) Defina la función estado-valor $v^{\pi}(s)$ para un MDP con descuento
- (5 puntos) Considere una política aleatoria uniforme $\pi_1(s, a)$ que toma todas las acciones desde el estado s con la misma probabilidad. Iniciando con un valor de función $v_1(A) = v_1(B) = v_1(C) = 2$, aplicar una actualización de evaluación de política iterativa para calcular $v_2(s) \forall s \in S$.
- (5 puntos) Aplicar una mejora de política voraz para calcular una nueva política $\pi_2(s)$
- (5 puntos) Iniciando con un valor de función $v_1(A) = v_1(B) = v_1(C) = 2$, aplicar una actualización de iteración de valor para calcular $v_2(s) \forall s \in S$.
- (5 puntos) ¿Es $v_2(s)$ óptimo? justifique su respuesta

4. (30 puntos) **RL libre de modelo:** considere el siguiente proceso de decisión de Markov con dos estados A y B y acciones a y b. La matriz de transición y la función de recompensa son desconocidos, pero se han observado dos episodios:

$$A, a, 3, A, b, 2, B, b, -4, A, b, 4, B, a, -3$$

$$B, a, -2, A, b, 3, B, a, -3$$

- (4 puntos) Utilice evaluación de Monte-Carlos con primera visita para estimar $v(A), v(B)$

- (4 puntos) Utilice evaluación de Monte-Carlos con cada visita para estimar $v(A), v(B)$
- (4 puntos) Dibuje el diagrama que mejor represente el MDP que mejor explique esos dos episodios
- (4 puntos) Resuelva la ecuación de Bellman para encontrar $v(A), v(B)$
- (4 puntos) ¿Cuál función de valor se encontraría con TD(0) usando la información de esos dos episodios?
- (5 puntos) ¿Qué valor encontraría diferencia temporal con mínimos cuadrados (LSTD(0))?
- (5 puntos) Aplique un paso de REINFORCE iniciando con una política aleatoria uniforme

5. (10 puntos) **Aproximación de funciones:** 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.

   - (2 puntos) Representar el estado del ratón s por dos características $bell(s) \in 0, 1$ y $light(s) \in 0, 1$. Escriba la secuencia de vectores estados correspondiente a este episodio.
   - (2 puntos) Aproxime la función estado-valor por medio de una combinación lineal de estas características con dos parámetros $b(bell(S)+ l(light(s)$. Si $b = 2$ y $l = -2$ escriba la secuencia de valores aproximados correspondiente al estado.
   - (2 puntos) Define el retorno λ v_t^λ
   - (2 puntos) Escriba la secuencia de retornos λ v_t^λ correspondientes al episodio para $\lambda = 0{,}5, b = 2, l = -2$
   - (2 puntos) Utilizando $TD(\lambda)$ y la función de aproximación lineal, ¿Cuáles son las secuencias de actualizaciones al peso b? utilice $\lambda = 0{,}5, \gamma = 1, \alpha = 0{,}5$ e inicie con $b = 2, l = -2$.

6. (10 puntos) 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 para transitar en un estado se definen como $R(s_i) = i$. La recompensa máxima es 3. El agente sigue la trayectoria

$$s_1, a_1, 1, s_1, a_2, 2, s_2$$

   - (2 puntos) Aplique el algoritmo de Q-learning para el episodio dado inicializando q a cero
   - (2 puntos) Dado que el agente se encuentra en s_1, y siguiendo una política épsilon-voraz. ¿Cuáles son las probabilidades para tomar la siguiente acción?
   - (2 puntos) Aplique el algoritmo de Q-learning para el episodio dado inicializando q de forma optimista a rmax

- (2 puntos) Dado que el agente se encuentra en s_1, y siguiendo una política épsilon-voraz. ¿Cuáles son las probabilidades para tomar la siguiente acción?
- (2 puntos) Genere el MDP a partir del episodio dado, muestree 3 episodios más siguiendo una política voraz y realice las actualizaciones a q.