Agentes Inteligentes

Franz Mayr
mayr@ort.edu.uy

Universidad ORT Uruguay
17 de abril de 2023

6. Métodos de Diferencias Temporales

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Inicializar:

$V(s) \in \mathbb{R}$ arbitrariamente, $\forall s \in \mathcal{S}$

$Retornos(s) \leftarrow$ lista vacía, $\forall s \in \mathcal{S}$

Repetir:

Generar un episodio según π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Para cada paso del episodio, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma\, G + R_{t+1}$

Si S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$:

Agregar G a $Retornos(S_t)$

$V(S_t) \leftarrow$ promedio($Retornos(S_t)$)

$V(s)$ converge a $v_\pi(s)$ (para los estados visitados).

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Inicializar:

$\quad V(s) \in \mathbb{R}$ arbitrariamente, $\forall s \in \mathcal{S}$

$\quad \mathit{Retornos}(s) \leftarrow$ lista vacía, $\forall s \in \mathcal{S}$

Repetir:

$\quad$ Generar un episodio según π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$\quad G \leftarrow 0$

$\quad$ Para cada paso del episodio, $t = T-1, T-2, \ldots, 0$:

$\quad\quad G \leftarrow \gamma\, G + R_{t+1}$

$\quad\quad$ Si S_t no aparece en $S_0, S_1, \ldots, S_{t-1}$:

$\quad\quad\quad$ Agregar G a $\mathit{Retornos}(S_t)$

$\quad\quad\quad V(S_t) \leftarrow \text{promedio}(\mathit{Retornos}(S_t))$

$V(s)$ converge a $v_\pi(s)$ (para los estados visitados).

También vimos estimación MC de $q_\pi(s,a)$, control MC, con exploración inicial y con políticas ε-greedy.

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Después de k episodios que pasaron por un estado $s \in \mathcal{S}$, el valor estimado de s es:

$$V = \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)}}{k}$$

donde $G^{(i)}$ es el retorno medido desde v en la i-ésima iteración.

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Después de k episodios que pasaron por un estado $s \in \mathcal{S}$, el valor estimado de s es:

$$V = \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)}}{k}$$

donde $G^{(i)}$ es el retorno medido desde v en la i-ésima iteración.

Al terminar el episodio $k+1$, el valor estimado de s es:

$$V' = \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)} + G^{(k+1)}}{k+1}$$

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Después de k episodios que pasaron por un estado $s \in \mathcal{S}$, el valor estimado de s es:

$$V = \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)}}{k}$$

donde $G^{(i)}$ es el retorno medido desde v en la i-ésima iteración.

Al terminar el episodio $k+1$, el valor estimado de s es:

$$\begin{aligned} V' &= \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)} + G^{(k+1)}}{k+1} \\ &= \frac{V \cdot k + G^{(k+1)}}{k+1} \end{aligned}$$

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Después de k episodios que pasaron por un estado $s \in \mathcal{S}$, el valor estimado de s es:

$$V = \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)}}{k}$$

donde $G^{(i)}$ es el retorno medido desde v en la i-ésima iteración.

Al terminar el episodio $k+1$, el valor estimado de s es:

$$\begin{aligned}
V' &= \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)} + G^{(k+1)}}{k+1} \\
&= \frac{V \cdot k + G^{(k+1)}}{k+1} \\
&= \frac{V \cdot k}{k+1} + \frac{G^{(k+1)}}{k+1} + \frac{V}{k+1} - \frac{V}{k+1}
\end{aligned}$$

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Después de k episodios que pasaron por un estado $s \in \mathcal{S}$, el valor estimado de s es:

$$V = \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)}}{k}$$

donde $G^{(i)}$ es el retorno medido desde v en la i-ésima iteración.

Al terminar el episodio $k+1$, el valor estimado de s es:

$$\begin{aligned}
V' &= \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)} + G^{(k+1)}}{k+1} \\
&= \frac{V \cdot k + G^{(k+1)}}{k+1} \\
&= \frac{V \cdot k}{k+1} + \frac{G^{(k+1)}}{k+1} + \frac{V}{k+1} - \frac{V}{k+1} \\
&= V \cdot \frac{k+1}{k+1} + \frac{G^{(k+1)}}{k+1} - \frac{V}{k+1}
\end{aligned}$$

Repaso: Estimación Monte Carlo de $v_\pi(s)$

Después de k episodios que pasaron por un estado $s \in \mathcal{S}$, el valor estimado de s es:

$$V = \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)}}{k}$$

donde $G^{(i)}$ es el retorno medido desde v en la i-ésima iteración.

Al terminar el episodio $k+1$, el valor estimado de s es:

$$\begin{aligned}
V' &= \frac{G^{(1)} + G^{(2)} + G^{(3)} + \ldots + G^{(k)} + G^{(k+1)}}{k+1} \\
&= \frac{V \cdot k + G^{(k+1)}}{k+1} \\
&= \frac{V \cdot k}{k+1} + \frac{G^{(k+1)}}{k+1} + \frac{V}{k+1} - \frac{V}{k+1} \\
&= V \cdot \frac{k+1}{k+1} + \frac{G^{(k+1)}}{k+1} - \frac{V}{k+1} \\
&= V + \frac{1}{k+1} \cdot \left(G^{(k+1)} - V\right)
\end{aligned}$$

Métodos de Diferencias Temporales

Recordemos la asignación que vimos hace un tiempo:

$$NuevaEst \leftarrow ViejaEst + TasaAct \left[\color{blue}{Objetivo} - ViejaEst \right]$$

Métodos de Diferencias Temporales

Recordemos la asignación que vimos hace un tiempo:

$$NuevaEst \leftarrow ViejaEst + TasaAct \left[Objetivo - ViejaEst \right]$$

Los métodos MC actualizan su estimación V de v_π al final de cada episodio, una vez que conocen el retorno del mismo:

$$V(S_t) \leftarrow V(S_t) + \alpha \Big[G_t - V(S_t) \Big]$$

Métodos de Diferencias Temporales

Recordemos la asignación que vimos hace un tiempo:

$$\mathit{NuevaEst} \leftarrow \mathit{ViejaEst} + \mathit{TasaAct}\left[\mathit{Objetivo} - \mathit{ViejaEst}\right]$$

Los métodos MC actualizan su estimación V de v_π al final de cada episodio, una vez que conocen el retorno del mismo:

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[G_t - V(S_t)\Big]$$

En contraste, los métodos de diferencias temporales (TD) no esperan hasta el final del episodio, sino que actualizan V en cada paso del episodio:

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\Big]$$

Métodos de Diferencias Temporales

Recordemos la asignación que vimos hace un tiempo:

$$NuevaEst \leftarrow ViejaEst + TasaAct \left[Objetivo - ViejaEst \right]$$

Los métodos MC actualizan su estimación V de v_π al final de cada episodio, una vez que conocen el retorno del mismo:

$$V(S_t) \leftarrow V(S_t) + \alpha \Big[G_t - V(S_t) \Big]$$

En contraste, los métodos de diferencias temporales (TD) no esperan hasta el final del episodio, sino que actualizan V en cada paso del episodio:

$$V(S_t) \leftarrow V(S_t) + \alpha \Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \Big]$$

Al término $G_t - V(S_t)$ se lo denomina error MC, y al término $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$, error TD.

Estimación TD de $v_\pi(s)$: Algoritmo TD(0)

Inicializar $V(s)$ arbitrariamente $\forall s \in \mathcal{S}$

Estimación TD de $v_\pi(s)$: Algoritmo TD(0)

Inicializar $V(s)$ arbitrariamente $\forall s \in \mathcal{S}$

Repetir:

    Inicializar S

    Repetir:

        $A \leftarrow$ acción desde S según π

        Ejecutar la acción A; observar R, S'

        $S \leftarrow S'$

    hasta que S sea terminal

Estimación TD de $v_\pi(s)$: Algoritmo TD(0)

Inicializar $V(s)$ arbitrariamente $\forall s \in \mathcal{S}$

Repetir:
 Inicializar S
 Repetir:
  $A \leftarrow$ acción desde S según π
  Ejecutar la acción A; observar R, S'
  $V(S) \leftarrow V(S) + \alpha \left[R + \gamma V(S') - V(S) \right]$
  $S \leftarrow S'$
 hasta que S sea terminal

Estimación TD de $v_\pi(s)$: Algoritmo TD(0)

Inicializar $V(s)$ arbitrariamente $\forall s \in \mathcal{S}$

Repetir:
  Inicializar S
  Repetir:
    $A \leftarrow$ acción desde S según π
    Ejecutar la acción A; observar R, S'
    $V(S) \leftarrow V(S) + \alpha \left[R + \gamma V(S') - V(S)\right]$
    $S \leftarrow S'$
  hasta que S sea terminal

La tasa de aprendizaje $\alpha \in (0, 1]$ es un parámetro del algoritmo.

Métodos de Diferencias Temporales

<u>Observación</u>:

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] && (1)\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma\, G_{t+1} \mid S_t = s] && (2)\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s] && (3)
\end{aligned}
$$

[1] Sin relación con el método estadístico de muestreo con reemplazo.

Métodos de Diferencias Temporales

Observación:

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] && (1)\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma\, G_{t+1} \mid S_t = s] && (2)\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s] && (3)
\end{aligned}
$$

En la línea (1) vemos el objetivo de MC: G_t.
En la línea (3) vemos el objetivo de TD: $R_{t+1} + \gamma\, V(S_{t+1})$.
Esto muestra que ambos métodos convergen a la misma estimación de $v_\pi(s)$.

[1] Sin relación con el método estadístico de muestreo con reemplazo.

Métodos de Diferencias Temporales

Observación:

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] && (1) \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma\, G_{t+1} \mid S_t = s] && (2) \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s] && (3) \end{aligned}$$

En la línea (1) vemos el objetivo de MC: G_t.
En la línea (3) vemos el objetivo de TD: $R_{t+1} + \gamma\, V(S_{t+1})$.
Esto muestra que ambos métodos convergen a la misma estimación de $v_\pi(s)$.

La gran diferencia entre ambos es que MC basa su estimación en un valor real de G_t, pero TD la basa en otro valor estimado. La literatura de Aprendizaje Reforzado se refiere a esta característica como bootstrapping.[1]

[1] Sin relación con el método estadístico de muestreo con reemplazo.

MC vs. TD - Ejemplo: Conduciendo a casa

Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30

MC vs. TD - Ejemplo: Conduciendo a casa

Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30
En el auto, llueve	18:05	35	40

MC vs. TD - Ejemplo: Conduciendo a casa

Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30
En el auto, llueve	18:05	35	40
Saliendo de autopista	18:20	15	35

MC vs. TD - Ejemplo: Conduciendo a casa

Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30
En el auto, llueve	18:05	35	40
Saliendo de autopista	18:20	15	35
Ruta local, atrás de camión	18:30	10	40

MC vs. TD - Ejemplo: Conduciendo a casa

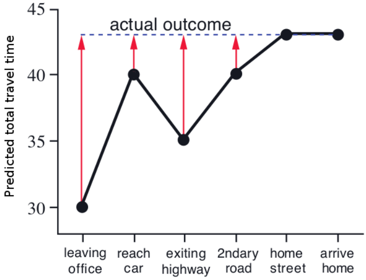
Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30
En el auto, llueve	18:05	35	40
Saliendo de autopista	18:20	15	35
Ruta local, atrás de camión	18:30	10	40
Calle de casa	18:40	3	43

MC vs. TD - Ejemplo: Conduciendo a casa

Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30
En el auto, llueve	18:05	35	40
Saliendo de autopista	18:20	15	35
Ruta local, atrás de camión	18:30	10	40
Calle de casa	18:40	3	43
Entrando en casa	18:43	0	43

MC vs. TD - Ejemplo: Conduciendo a casa

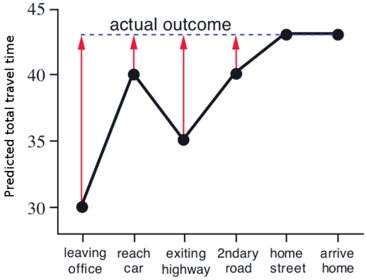
Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30
En el auto, llueve	18:05	35	40
Saliendo de autopista	18:20	15	35
Ruta local, atrás de camión	18:30	10	40
Calle de casa	18:40	3	43
Entrando en casa	18:43	0	43



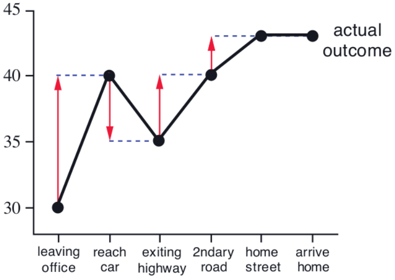
Métodos Monte Carlo

MC vs. TD - Ejemplo: Conduciendo a casa

Estado	Hora	Tiempo restante predicho	Tiempo total predicho
Saliendo del trabajo	18:00	30	30
En el auto, llueve	18:05	35	40
Saliendo de autopista	18:20	15	35
Ruta local, atrás de camión	18:30	10	40
Calle de casa	18:40	3	43
Entrando en casa	18:43	0	43



Métodos Monte Carlo



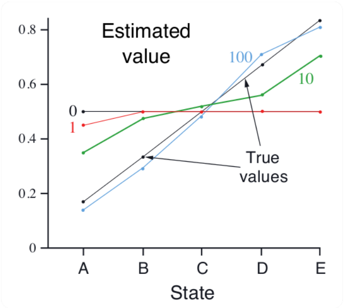
Métodos TD

MC vs. TD - Ejemplo: Random walk

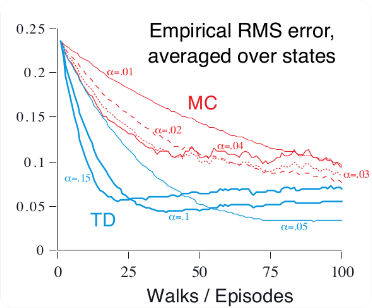
$$\pi(\leftarrow | \, s) = \pi(\rightarrow | \, s) = 0{,}5 \ \forall s \in \{\mathsf{A,B,C,D,E}\}; \ \gamma = 1$$

MC vs. TD - Ejemplo: Random walk

$$\pi(\leftarrow | s) = \pi(\rightarrow | s) = 0{,}5 \ \forall s \in \{\mathsf{A,B,C,D,E}\};\ \gamma = 1$$



Eje y: valores de los estados, aprendidos por TD(0) después del número indicado de episodios.



Eje y: RMSE de los valores aprendidos respecto de los reales, promediado para los 5 estados.

Estimación TD de $q_\pi(s, a)$: Algoritmo TD(0)

El algoritmo TD(0) para estimar la función de valor de los pares estado-acción es análogo:

Inicializar $Q(s, a)$ arbitrariamente $\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

Repetir:

  Inicializar S

  $A \leftarrow$ acción desde S según π

  Repetir:

    Ejecutar la acción A; observar R, S'

    $A' \leftarrow$ acción desde S' según π

    $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma\, Q(S', A') - Q(S, A)\right]$

    $S \leftarrow S'$

    $A \leftarrow A'$

  hasta que S sea terminal

La tasa de aprendizaje $\alpha \in (0, 1]$ es un parámetro del algoritmo.

On-policy TD control: Sarsa

En MC, la actualización de la estimación de $q_\pi(s, a)$ se realiza al final de cada episodio, una vez conocido el objetivo G_t:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[G_t - Q(S_t, A_t) \Big]$$

On-policy TD control: Sarsa

En MC, la actualización de la estimación de $q_\pi(s, a)$ se realiza al final de cada episodio, una vez conocido el objetivo G_t:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[G_t - Q(S_t, A_t) \Big]$$

En TD, esta estimación se actualiza después de tomar cada acción A_t y observar (S_{t+1}, R_{t+1}) (bootstrapping):

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma\, Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \Big]$$

donde A_{t+1} es una acción desde S_{t+1} elegida según una política basada en Q (p.ej., ε-greedy).

On-policy TD control: Sarsa

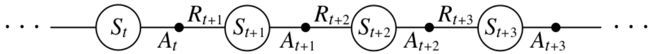
En MC, la actualización de la estimación de $q_\pi(s,a)$ se realiza al final de cada episodio, una vez conocido el objetivo G_t:

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\Big[G_t - Q(S_t,A_t)\Big]$$

En TD, esta estimación se actualiza después de tomar cada acción A_t y observar (S_{t+1}, R_{t+1}) (bootstrapping):

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\Big[R_{t+1} + \gamma\, Q(S_{t+1},A_{t+1}) - Q(S_t,A_t)\Big]$$

donde A_{t+1} es una acción desde S_{t+1} elegida según una política basada en Q (p.ej., ε-greedy).



Esta regla usa la quíntupla $S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}$, de ahí su nombre.

On-policy TD control: Sarsa

Inicializar $Q(s,a)$ arbitrariamente $\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

Repetir:

  Inicializar S

  $A \leftarrow$ acción desde S según política ε-greedy basada en Q

  Repetir:

    Ejecutar la acción A; observar R, S'

    $A' \leftarrow$ acción desde S' según política ε-greedy basada en Q

    $Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma\, Q(S',A') - Q(S,A)\right]$

    $S \leftarrow S'$

    $A \leftarrow A'$

  hasta que S sea terminal

La tasa de aprendizaje $\alpha \in (0,1]$ es un parámetro del algoritmo.

Off-policy TD control: Q-learning

Sarsa usa la misma política ε-greedy basada en Q para elegir la siguiente acción A_{t+1} y para estimar el retorno G_t (en rojo) en la regla de actualización de Q:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \, Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \Big]$$

Por eso, SARSA es un método on-policy.

Con este cambio a la regla de actualización de Q:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \, \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \Big]$$

desacoplamos la estimación del retorno G_t (ahora se supone una política greedy pura) de la política seguida para elegir la siguiente acción del episodio (ε-greedy).

Por eso, Q-learning es un método off-policy.

Off-policy TD control: Q-learning

Inicializar $Q(s,a)$ arbitrariamente $\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

Repetir:

  Inicializar S

  Repetir:

    $A \leftarrow$ acción desde S según política ε-greedy basada en Q

    Ejecutar la acción A; observar R, S'

    $Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$

    $S \leftarrow S'$

  hasta que S sea terminal

La tasa de aprendizaje $\alpha \in (0,1]$ es un parámetro del algoritmo.

Resumen - Métodos de Diferencias Temporales

Los métodos TD no esperan al final del episodio para actualizar sus estimaciones, sino que lo hacen luego de cada paso en el episodio.

- Algoritmo TD(0) para estimar $v_\pi(s)$ o $q_\pi(s, a)$.
- On-policy TD control: algoritmo Sarsa.
- Off-policy TD control: algoritmo Q-learning.

Próximas clases:

- Introducción a redes neuronales y deep learning.
- Métodos de aproximación de funciones; deep reinforcement learning.