Aprendizado por Reforço

Métodos de diferenças temporais – parte 1

Introdução

- Métodos de diferenças temporais (TD) são algoritmos **sem modelo**, assim como os métodos de Monte Carlo (MC).
- A principal vantagem dos métodos TD está em sua forma **incremental**, que permite atualizações a cada passo.
- Aprendizado por TD pode ser visto como uma classe de algoritmos estocásticos para resolver as equações de Bellman (ou equações de otimalidade de Bellman).

Panorama dos algoritmos TD

- TD (valores de estado)
  - Estima os valores de estado de uma dada política.
  - É o algoritmo base para os demais métodos TD.
- Sarsa
  - Estima valores de ação de uma dada política.
  - Pode ser combinado com melhoria de política para encontrar políticas ótimas.
- n-step Sarsa
  - Generalização do Sarsa.
  - Sarsa e MC são casos particulares do n-step Sarsa.
- Q-learning
  - Algoritmo clássico de aprendizado por reforço.
  - Busca resolver diretamente a equação de otimalidade de Bellman para encontrar políticas ótimas.

Aprendizado por TD

- Aprendizado por TD na realidade pode se referir a uma ampla classe de algoritmos de aprendizado por reforço.
- Consideremos o algoritmo clássico de TD para a estimação de valores de estado.
  - Objetivo: dado uma política π, queremos estimar os valores $v_\pi(s)\ \forall s \in \mathcal{S}$.
  - A estimativa é feita a partir de amostras obtidas (experiência) ao seguir a política π:

$$(s_0, r_1, s_1, \dots, s_t, r_{t+1}, s_{t+1}, \dots)$$

Atualização incremental dos valores de estado

- O algoritmo TD atualiza apenas o valor do estado visitado s_t:

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)[v_t(s_t) - (r_{t+1} + \gamma v_t(s_{t+1}))]$$

- Para todos os outros estados não visitados, os valores permanecem inalterados:

$$v_{t+1}(s) = v_t(s), \qquad \forall s \neq s_t$$

  - $t = 0,1,2,\ldots$ (tempo)
  - $v_t(s_t)$: estimativa de $v_\pi(s_t)$ no tempo t
  - $\alpha_t(s_t)$: taxa de aprendizado para s_t no tempo t

Conexão com a equação de Bellman

- O algoritmo TD pode ser interpretado como um algoritmo de aproximação estocástica para resolver a equação de Bellman:

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid S_t = s], \qquad s \in \mathcal{S}$$

  - A equação acima pode ser reescrita como (equação de expectativa de Bellman):

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s], \qquad s \in \mathcal{S}$$

  - O algoritmo de TD pode ser derivado ao aplicar o método de Robbins-Monro para resolver a equação de expectativa de Bellman.

TD como aproximação estocástica

- Definimos

$$g\big(v_\pi(s_t)\big) \triangleq v_\pi(s_t) - \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

- Assim, encontrar $v_\pi(s_t)$ equivale a resolver $\boxed{g\big(v_\pi(s_t)\big) = 0}$ (problema de encontrar raiz).

TD como aproximação estocástica

- Como apenas amostras r_{t+1} e s_{t+1} estão disponíveis, observamos a versão ruidosa

$$\tilde{g}\big(v_\pi(s_t)\big) = v_\pi(s_t) - \big(r_{t+1} + \gamma v_\pi(s_{t+1})\big)$$

$$\tilde{g}\big(v_\pi(s_t)\big) = \underbrace{\big(v_\pi(s_t) - \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]\big)}_{g(v_\pi(s_t))} + \underbrace{\big(\mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] - [r_{t+1} + \gamma v_\pi(s_{t+1})]\big)}_{\eta}$$

TD como aproximação estocástica

Atualização Robbins-Monro com Amostras

- O método de Robbins-Monro ajusta a estimativa usando a observação ruidosa:

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)\tilde{g}\big(v_t(s_t)\big)$$

- Substituindo $\tilde{g}\big(v_t(s_t)\big)$:

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)[v_t(s_t) - (r_{t+1} + \gamma v_\pi(s_{t+1}))]$$

  - $v_t(s_t)$: estimativa de $v_\pi(s_t)$ no tempo t
  - $\alpha_t(s_t)$: taxa de aprendizado para s_t no tempo t

TD como aproximação estocástica

Generalização para todos os estados

- A forma a seguir supõe que os demais valores de estado já são conhecidos e deseja-se estimar somente o valor de estado de s_t:

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)[v_t(s_t) - (r_{t+1} + \gamma v_\pi(s_{t+1}))]$$

- Para estimar todos os estados simultaneamente, substitui-se $\gamma v_\pi(s_{t+1})$ por $v_t(s_{t+1})$:

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)[v_t(s_t) - (r_{t+1} + \gamma v_t(s_{t+1}))]$$

Alvo e Erro em TD

Definições

1. Algoritmo TD

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)[v_t(s_t) - (r_{t+1} + \gamma v_t(s_{t+1}))]$$

2. Alvo TD

$$\bar{v}_t \triangleq r_{t+1} + \gamma v_t(s_{t+1})$$

2. Erro TD

$$\delta_t \triangleq v_t(s_t) - \bar{v}_t = v_t(s_t) - (r_{t+1} + \gamma v_t(s_{t+1}))$$

- A nova estimativa $v_{t+1}(s_t)$ combina a atual $v_t(s_t)$ com o erro δ_t:

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)\delta_t$$

Por que $\overline{v}$ é o alvo?

Por que $\bar{v}$ é o alvo?

- Algoritmo TD

$$v_{t+1}(s_t) = v_t(s_t) - \alpha_t(s_t)[v_t(s_t) - \bar{v}_t]$$

- Subtraindo $\bar{v}_t$ de ambos os lados da regra de atualização:

$$v_{t+1}(s_t) - \bar{v}_t = [v_t(s_t) - \bar{v}_t] - \alpha_t(s_t)[v_t(s_t) - \bar{v}_t]$$

$$v_{t+1}(s_t) - \bar{v}_t = [1 - \alpha_t(s_t)][v_t(s_t) - \bar{v}_t]$$

- Como $0 < 1 - \alpha_t(s_t) < 1$ ($\alpha_t(s_t)$ é um pequeno número positivo), obtém-se

$$v_{t+1}(s_t) - \bar{v}_t < [v_t(s_t) - \bar{v}_t]$$

Conclusão: cada passo aproxima o valor de estado à $\bar{v}_t$; por isso $\bar{v}_t$ é tratado como o alvo que a estimativa persegue.

Interpretação do Erro TD

- Erro TD (δ_t)
  - Reflete a discrepância entre dois instantes consecutivos t e $t+1$:

$$\delta_t = v_t(s_t) - (r_{t+1} + \gamma v_t(s_{t+1}))$$

  - Quando v_t coincide com o valor verdadeiro v_π,

$$\mathbb{E}[\delta_t \mid S_t = s] = 0$$

  - **Por quê?**

Interpretação do Erro TD

- Erro TD (δ_t)
  - Reflete a discrepância entre dois instantes consecutivos t e $t+1$:

$$\delta_t = v_t(s_t) - (r_{t+1} + \gamma v_t(s_{t+1}))$$

  - Quando v_t coincide com o valor verdadeiro v_π,

$$\mathbb{E}[\delta_t \mid S_t = s] = 0$$

  - **Por quê?**

$$\mathbb{E}[\delta_t \mid S_t = s] = \mathbb{E}\left[v_\pi(S_t) - \left(R_{t+1} + \gamma v_\pi(S_{t+1})\right) \mid S_t = s\right]$$
$$\mathbb{E}[\delta_t \mid S_t = s] = v_\pi(s_t) - \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$
$$\mathbb{E}[\delta_t \mid S_t = s] = 0$$

Demais Propriedades

- O algoritmo TD apresentado estima apenas valores de estado para uma dada política;
- Encontrar **políticas ótimas** exige estimar **valores de ação** e realizar a melhoria de política.

TD vs. MC?

TD vs. MC

Aprendizado por TD

- Incremental
  - Atualiza valores de estado ou de ação imediatamente após receber uma amostra de experiência.
- Tarefas
  - Lida tanto com tarefas episódicas quanto com tarefas contínuas.
- Bootstrapping
  - A atualização recorre à estimativa anterior;
  - Exige um palpite inicial para os valores.
- Baixa variância de estimação
  - Envolve menos variáveis aleatórias.
  - Para estimar $q_\pi(s_t, a_t)$, o Sarsa usa apenas $R_{t+1}, S_{t+1}, A_{t+1}$.

Aprendizado por MC

- Não-incremental
  - Precisa aguardar o término completo do episódio para calcular o retorno descontado.
- Tarefas episódicas
  - Restrito a tarefas episódicas (número finito de passos).
- Sem bootstrapping
  - Estima diretamente valores de estado ou ação sem depender de estimativas iniciais.
- Alta variância de estimação
  - Usa mais variáveis aleatórias.
  - Para $q_\pi(s_t, a_t)$ requer $R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$
  - Se o episódio tem comprimento L e cada estado possui $|\mathcal{A}|$ ações, há $|\mathcal{A}|^L$ possíveis trajetórias, elevando a variância quando poucas amostras são usadas.

Aprendizado TD

- Convergência

Dada uma política π, pelo algoritmo TD, $v_t(s)$ converge quase certamente para o valor de estado $v_\pi(s)$ quando $t \to \infty$, para todo $s \in \mathcal{S}$, se

$$\sum_t \alpha_t(s) = \infty \qquad e \qquad \sum_t \alpha_t^2(s) < \infty, \qquad \forall s \in \mathcal{S}$$

Referências

- S. Zhao. *Mathematical Foundations of Reinforcement Learning*. Springer Singapore, 2025. [capítulo 7]
  - **disponível em:** https://github.com/MathFoundationRL/Book-Mathematical-Foundation-of-Reinforcement-Learning

- R. S. Sutton e A. G. Barto. *An Introduction Reinforcement Learning*, Bradford Book, 2018. [capítulos 6 e 7]
  - **disponível em:** http://incompleteideas.net/book/the-book-2nd.html

Slides construídos com base nos livros supracitados, os quais estão disponibilizados publicamente pelos seus respectivos autores.