Temporal Difference (TD) Learning

1. khái niệm

Temporal Difference (TD) learning là một phương pháp học tăng cường cho phép agent học từ kinh nghiệm một cách liên tục, thay vì đợi đến cuối một tập (episode) mới đánh giá hiệu quả. Phương pháp này kết hợp các ý tưởng từ lập trình động và phương pháp Monte Carlo.

2.1. Nguyên Lý Cơ Bản

Nguyên lý cơ bản của Temporal Difference (TD) Learning trong Reinforcement Learning (RL) là sự kết hợp giữa:

- Learning from experience (học từ trải nghiệm thực tế), như trong Monte Carlo methods, và
- Bootstrapping (ước lượng giá trị hiện tại từ giá trị trong tương lai gần), như trong Dynamic Programming.

Mục tiêu của TD Learning:

Ước lượng giá trị kỳ vọng (value function) của một trạng thái nào đó theo chính sách hành động π, tức là:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid S_0 = s\right]$$

Cập Nhật Hàm Giá Trị(Cập nhật theo TD(0) – đơn giản nhất)

$$V(S_t) \leftarrow V(S_t) + \alpha \cdot (R_t + \gamma V(S_{t+1}) - V(S_t))$$

Trong đó:

$V(S_t)$: Giá trị ước tính của trạng thái hiện tại

R_t: Phần thưởng nhận được sau khi thực hiện hành động

α : Kích thước bước (Learning rate)

γ: Hệ số chiết khấu, xác định tầm quan trọng của phần thưởng tương lai

$V(S_{t+1})$: Giá trị dự đoán của trạng thái tiếp theo

TD error:

Được ký hiệu là δ_t

$$\delta_t = R_t + \gamma V(S_{t+1}) - V(S_t)$$

Lỗi này đo lường sự khác biệt giữa giá trị dự đoán và kết quả thực tế.

Ứng dụng :

- Là nền tảng cho các thuật toán như:
  - **SARSA** (on-policy TD control)
  - **Q-learning** (off-policy TD control)
- Dùng trong **Deep Reinforcement Learning** để cập nhật mạng neural