FPT Education

FPT UNIVERSITY

Temporal Difference Learning

Objectives

- Define temporal-difference learning
- Define the temporal-difference error
- Understand the TD(0) algorithm

Temporal Difference

- Temporal Difference (TD) learning is a reinforcement learning technique that combines elements of Monte Carlo methods and dynamic programming.
- It is a model-free approach used to estimate value functions or policies directly from experience, without requiring a model of the environment's dyna

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

Temporal Difference

- Temporal Difference Error:
  - At each time step, TD learning updates value estimates based on the temporal difference error, which is the difference between the expected return and the current estimate.
  - In other words, it updates the value estimate towards a better estimate of the true value, based on the difference between observed rewards and the predictions made by the current estimate.

FPT Education
FPT UNIVERSITY

Temporal Difference

- TD Target:
  - The TD target is the sum of the immediate reward plus the estimated value of the next state, discounted by a factor γ. It represents the agent's expected return from the current state-action pair.

Temporal Difference

- TD Error:
  - The TD error is the difference between the TD target and the current estimate of the value function.
  - It measures how much the current estimate needs to be adjusted to match the observed returns.
  - The temporal-difference (TD) error, often denoted as δt, is a key concept in temporal-difference learning in reinforcement learning.
  - It represents the discrepancy between the predicted value of a state or state-action pair and the observed return obtained from the environment at a given time step.

Temporal Difference

- TD Error:
  - The TD error at time step t is defined as:

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

Where:

- δ_t is the TD error at time step t.
- r_{t+1} is the immediate reward received after taking an action from state s_t and transitioning to state s_{t+1}.
- γ is the discount factor, which represents the importance of future rewards relative to immediate rewards.
- $V(s_t)$ is the estimated value of state s_t at time step t.
- $V(s_{t+1})$ is the estimated value of state s_{t+1} at time step $t+1$.

FPT Education
FPT UNIVERSITY

Temporal Difference

- TD Error:
  - The TD error measures the difference between the expected value of the current state and the sum of the immediate reward and the discounted value of the next state.
  - It indicates how much the current estimate of the value function needs to be adjusted to match the observed return.
  - Temporal-difference learning algorithms use the TD error to update value estimates iteratively, adjusting the estimates towards the observed returns in order to improve the accuracy of value function estimates and

Temporal Difference

- Temporal Difference Update Rule:
  - TD learning algorithms update value estimates iteratively based on TD errors.
  - The value estimates are adjusted towards the TD target by a small step size α, known as the learning rate.
  - The update rule is typically of the form:

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

Temporal Difference

- Policy Improvement:
  - TD learning can be used for policy improvement by estimating action values (Q-values) and selecting actions based on the estimated values. Q-learning is a popular TD learning algorithm that learns action values and selects actions greedily based on the estimated Q-values.

Temporal Difference

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

Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π, the policy to be evaluated

$V \leftarrow \vec{0}, V' \leftarrow \vec{0}$

Loop:

  $\Delta \leftarrow 0$

  Loop for each $s \in \mathcal{S}$:

    $V'(s) \leftarrow \sum_a \pi(a \mid s) \boxed{\sum_{s',r} p(s',r \mid s,a)} [r + \gamma V(s')]$

    $\Delta \leftarrow \max(\Delta, |V'(s) - V(s)|)$

  $V \leftarrow V'$

until $\Delta < \theta$ (a small positive number)

Output $V \approx v_\pi$

Temporal Difference

- The tabular TD zero algorithms.

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\textit{terminal}) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\left[R + \gamma V(S') - V(S)\right]$
        $S \leftarrow S'$
    until S is terminal

Temporal Difference

- Simple illustrative example:
  - We have a simplified grid world environment where an agent can move left or right. The agent starts at position A and the goal is to reach position B, where it receives a reward of +1. The agent receives a reward of 0 for all other states.

```
# 2.5 Simple Illustrative Example - HoaDNt@fe.edu.vn

A 0 0 0 0 0 B

```

Temporal Difference

- Simple illustrative example:
  - Episode of the agent's interaction with the environment using TD learning:

1. **Start**: The agent starts at state A with a value estimate of 0.
2. **Action**: The agent chooses to move right (e.g., based on a random policy).
3. **Transition**: The agent moves to state B and receives a reward of +1.
4. **Update Value**: The agent updates its value estimate for state A using the TD update rule:
   - New Value of A = Old Value of A + (Step Size) * (Reward + (Discount Factor) * Value of B - Old Value of A)
   - New Value of A = 0 + (Step Size) * (1 + 0 - 0) = Step Size
   - Let's assume we use a step size of 0.1, so the new value of A becomes 0.1.
5. **End of Episode**: The episode ends since the agent reached the goal state B.

Temporal Difference

- Simple illustrative example:
  - the agent's value estimate for state A has been updated based on the observed reward and the estimated value of the next state B.
  - The agent continues to interact with the environment over multiple episodes, updating its value estimates after each step using the TD update rule.
  - Over time, the agent's value estimates converge to the true values, allowing it to make better decisions and ultimately reach the goal more efficiently.

Summary

- Define temporal-difference learning
- Define the temporal-difference error
- Understand the TD(0) algorithm

FPT Education
FPT UNIVERSITY

Q & A