REINFORCEMENT LEARNING

Lecture 3 : Monte Carlo and Temporal Difference

Ibrahim Sammour

December | 2023

Types of Reinforcement Learning

- Model-Based
  - Uses an explicit model of the environment.
  - Example: Grid World

- Model-Free
  - Learn by directly interacting with the environment without prior knowledge
  - Example: Robot navigation

- On-Policy and Off-Policy.. (Later)

Monte Carlo Methods

- Based on experience sampling
- Sequence of state, action, reward pairs from direct interaction with the environment
- No prior knowledge of the Model of the environment
  - We do not know the probability distribution of the actions environment.
  - We do not necessarily have static rewards.

Monte Carlo Methods

- Solving reinforcement learning based on averaging returns
- They work on an episode to episode sense and only applicable for episodic tasks
- Aims to learn the state-value V_π(s) or action-value Q_π(s,a) functions

Recap

- Value function:
  - Value of a state or the expected return from a state

- Action value function:
  - Value of each action in a state

- Cumulative reward (return)
  - Discounted future reward starting from a state

RL Prediction and Control

- RL Prediction
  - We already have a policy and we need to figure out how does it perform
  - Predict the expected return of a state

- RL Control
  - We are learning the policy until reaching an optimal one
  - Find π that maximizes the expected return in any state

Monte Carlo Prediction

- **Model-free method.**
- **Value of a state = mean of the returns (return === Cumulative reward).**
- **Applies only to episodic tasks.**
- **Steps:**
  1. **Policy evaluation.**
  2. **Policy improvement.**
  3. **Control.**

Monte Carlo Prediction

First Visit Method

Input: a policy π

Initialize:

Initialize $V(s)$, for all s in $\mathcal{S}$

$Returns(s) \leftarrow$ an empty list, for all s in $\mathcal{S}$

Loop over episodes:

- Generate an episode following π: $s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{t+1}, a_{t+1}, r_{t+1}$
- $\bar{r} \leftarrow 0$
    - Loop for each step of the episode, t = T-1, T-2, ..., 0:

        If we did not pass through s_t before in this episode:

        - $\bar{r}_{st} \leftarrow r_t + \gamma \bar{r}_{st+1}$
        - Append $\bar{r}$ to $Returns(s_t)$

$V(s_t) \leftarrow average(Returns(s_t))$

Monte Carlo Prediction

Every Visit Method

Input: a policy π

Initialize:

Initialize $V(s)$, for all s in $\mathcal{S}$

$Returns(s)$ ← an empty list, for all s in $\mathcal{S}$

Loop over episodes:

- Generate an episode following π: $s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{t+1}, a_{t+1}, r_{t+1}$
- $\bar{r} \leftarrow 0$
  - Loop for each step of the episode, $t = T-1, T-2, \ldots, 0$:
    - $\bar{r}_{st} \leftarrow r_t + \gamma \bar{r}_{st+1}$
    - Append $\bar{r}$ to $Returns(s_t)$
  - Average $\bar{r}$ for each state in the episode
  - $V(s_t) \leftarrow average(Returns(s_t))$

Monte Carlo Control

Exploring Starts Method

Initialize:

Initialize π arbitrarily

Initialize Q(s,a), for all states and actions

Returns(s,a) ← an empty list, for all states and actions

Loop over episodes:

- Start from a random state s_0 and pick a random action a_0

- Generate an episode following π: $s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{t+1}, a_{t+1}, r_{t+1}$

- $\bar{r} \leftarrow 0$

  - Loop for each step of the episode, t = T-1, T-2, ..., 0:

    If we did not pass through s_t before in this episode:

    - $\bar{r}_{st} \leftarrow r_t + \gamma\bar{r}_{st+1}$

    - Append $\bar{r}$ to Returns(s_t, a_t)

  - Q(s_t,a_t) ← average(Returns(s_t, a_t))

  - Update policy as the best action given a state