Sutton & Barto Reinforcement Learning Notes

Stan Tuznik

May 17, 2022

Contents

Chapter 1

Introduction

Chapter 2

Multi-armed Bandits

- Reinforcement learning evaluates states and actions, it doesn't just tell what action to take.
  - Evolutionary methods, by contrast, take a policy, evaluate it as a whole, and modify it. It does this holistically, not by inspecting and evaluating individual actions!
  - Evaluating actions requires *taking* them; this requires exploration.
- **Bandit problems** use only a single state in which one of k actions may be taken. The goal is to evaluate each action.

2.1 A k-armed Bandit Problem

- **Bandit problem**: we have k possible actions, and we can take one action in each of N time steps. Taking actions results in an immediate reward. **Goal** is to learn how to behave in order to maximize the total reward over all time steps.
- When an action A_t is taken, a reward R_t is obtained. In general, this reward will be a random variable, and so comes from some unknown distribution, which we assume is *stationary* over time.
  - We can consider the expected reward of taking action $A_t = a$:

    $$q_*(a) \coloneqq \mathbb{E}\left[R_t \mid A_t = a\right] \tag{2.1}$$

    We call this the **value of taking action a at time t.**
  - The function $q_* : \mathcal{A} \to \mathbb{R}$ so defined is the **action-value function**.
- If we knew this action-value function, we could "beat probability" by always choosing the action with the highest expected reward, i.e., the highest action-value.
- Unfortunately, we don't actually know what the distribution of rewards looks like, so we can't exactly compute this action-value function.

  - Instead, we approximate it with a function Q_t such that $Q_t(a)$ is our estimated value of action a at time step t.
    * Does this mean the estimated value of taking action a at time t, or the estimated value at time t of taking action a? It seems from the notation "q_*" not indicating t that it must be the latter.
    * We want $Q_t(a) \approx q_*(a)$, so this seems to support my conclusion.
    * We also assumed that the distributions are stationary and so the value of any particular action should not depend on when we take it.

- A **greedy action** is one with the maximum expected total reward; an action is **exploratory** if it is not greedy.

- Strategizing about whether to act greedily depends on uncertainties, estimates, and the finite time horizon (if we don't have much time left, it's probably not worth exploring).

2.2 Action-value Methods

- **Action-value methods** estimate the value of actions and use those estimates to make action selection decisions.
  - The *true value* of an action is the expected (mean) reward when that action is taken. We can't possibly know this, though!

- One simple approach is the **sample-average** method, where take a large number $(t-1)$ of actions, observe the reward obtained each time, and compute the empirical average reward observed after taking action a:

$$Q_t(a) := \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{I}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{I}_{A_i=a}} \tag{2.2}$$

  - This actually converges to q_* in the limit as $t \to \infty$.

- How can we use these estimates to select actions?
  - The greedy action is the action (or set thereof) which maximizes the expected reward:

$$A_t = \arg\max_a Q_t(a) \tag{2.3}$$

  - "Greediness always exploits current knowledge to maximize immediate reward."
  - Performing greedily is entirely exploitation and no exploitation!

- **ϵ-greedy methods** incorporate exploration by setting a parameter $\epsilon \in (0,1)$ — generally to a small value — and taking an exploratory action $\epsilon \times 100\%$ of the time.
  - This method theoretically guarantees that all actions will be sampled an infinite number of times, so that $Q_t(a) \to q_*(a)$.

2.3 The 10-armed Testbed

- The authors build a simple bandit test problem. There are $k = 10$ actions to take, and for each action, the rewards are normally distributed.
  - Thus, taking an action multiple times can give different rewards, but the rewards all come from the same distribution.
  - The authors use a simple normal distribution for the reward distribution for each action.

    $$r \sim \Pr(r\,;\, A = a) = \mathcal{N}(q_*(a), 1) \tag{2.4}$$

    where the action-values $q_*(a)$ are generated from a standard normal distribution:

    $$q_*(a) \sim \mathcal{N}(0, 1) \tag{2.5}$$

  - This is fairly easy to implement, and gives us a "black box" environment for experimenting with different bandit algorithms, such as the aforementioned ϵ-greedy method.