Chapter 2: Multi-armed Bandits - Highlights

Jacob Taylor Cassady

November 24, 2022

- The most important feature distinguishing reinforcement learning from other types of learning is that it uses training information that *evaluates* the actions taken rather than *instructs* by giving correct actions.
- Purely evaluative feedback indicates how good the action taken was, but not whether it was the best or the worst action possible.
- one that does not involve learning to act in more than one situation. This *nonassociative* setting
- *associative*, that is, when actions are taken in more than one situation.

1 A k-armed Bandit Problem

- You are faced repeatedly with a choice among k different options, or actions. After each choice you receive a numerical reward chosen from a stationary probability distribution that depends on the action you selected. Your objective is to maximize the expected total reward over some time period, for example, over 1000 action selections, or *time steps*.
- This is the original form of the *k-armed bandit problem*, so named by analogy to a slot machine, or "one-armed bandit," except that it has k levers instead of one.
- Through repeated action selections you are to maximize your winnings by concentrating your actions on the best levers.
- Today the term "bandit problem" is sometimes used for a generalization of the problem described above, but in this book we use it to refer just to this simple case.
- In our k-armed bandit problem, each of the k actions has an expected or mean reward given that that action is selected; let us call this the *value* of that action. We denote the action selected on time step t as A_t, and the corresponding reward as R_t. The value then of an arbitrary action a, denoted $q_*(a)$, is the expected reward given that a is selected:

$$q_*(a) = \mathbb{E}[R_t | A_t = a]$$

- We assume that you do not know the action values with certainty, although you may have estimates. We denote the estimated value of action a at time step t as $Q_t(a)$. We would like $Q_t(a)$ to be close to $q_*(a)$.
- If you maintain estimates of the action values, then at any time step there is at least one action whose estimated value is greatest. We call these the *greedy* actions. When you select one of these actions, we say that you are *exploiting* your current knowledge of the values of the actions. If instead you select one of the nongreedy actions, then we say you are *exploring*, because this enables you to improve your estimate of the nongreedy action's value.
- Reward is lower in the short run, during exploration, but higher in the long run because after you have discovered the better actions, you can exploit them many times.

2 Action-value Methods

- .

3 The 10-armed Testbed

- .

4 Incremental Implementation

- .

5 Tracking a Nonstationary Problem

- .

6 Optimistic Initial Values

- .

7 Upper-Confidence-Bound Action Selection

- .

8 Gradient Bandit Algorithms

- .

9 Associative Search (Contextual Bandits)

- .

10 Summary

- .