Date created: 2025-07-20-Sun

2.9 Associative Search (Contextual Bandits)

So far we've only considered *nonassociative* tasks, i.e. tasks in which there is no association between actions and situations. In these tasks, the learner tries to track a single "best" action. However, in the general RL problem the goal is to learn a *policy* that maps from different situations to actions that are best in those situations. To set the stage for the full problem, we briefly discuss the simplest way in which nonassociative tasks extend to the associative setting.

Contextual bandits

Suppose there are several different k-armed bandit tasks, and that on each step you confront one of these at random. Naively, this might appear as a single, nonstationary k-armed bandit task where $q_*(a)$ changes at random. However, if you were to naively apply a nonstationary bandit algorithm, it would not work very well unless the values changed very slowly.

Now suppose there is some hint as to which bandit is selected at each step, e.g. the color of the screen. Now you can learn a policy associating each task (signaled by color) with the best action to take when faced with that task.

This is an example of an *associative search* task, because it involves both:

- trial-and-error learning to *search* for the best actions, and
- *association* of those actions with the situations in which they are best.

Often called *contextual bandits* in literature, associative search tasks is an intermediate between k-armed bandits and the full RL problem. Contextual bandits are like RL in that they require learning a policy, but are like k-armed bandits in that each action only affect the immediate reward. If actions are allowed to affect the *next situation* as well as the reward, then we have the full RL problem.

Exercises

2.10:

Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B). If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it? Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associasive search task. What is the best expectation os success you can achieve in this task, and how should you behave to achieve it?

Solution Case 1: bandit unknown

$$\begin{aligned}\mathbb{E}[R|a=1] &= 0.5 \cdot 0.1 + 0.5 \cdot 0.9 = 0.5 \\ \mathbb{E}[R|a=2] &= 0.5 \cdot 0.2 + 0.5 \cdot 0.8 = 0.5 \\ &\implies \forall \pi(a), \mathbb{E}(R) = 0.5\end{aligned}$$

Case 2: bandit known

$$\begin{aligned}\mathbb{E}[R|b=A] = \pi(1,A) \cdot 0.1 + \pi(2,A) \cdot 0.9 &\implies \pi(1,A) = 0, \pi(2,A) = 1 \\ \mathbb{E}[R|b=B] = \pi(1,B) \cdot 0.2 + \pi(2,B) \cdot 0.8 &\implies \pi(1,B) = 1, \pi(2,B) = 0 \\ \implies \mathbb{E}(R) = 0.5 \cdot 0.2 + 0.5 \cdot 0.9 &= 0.55\end{aligned}$$