Author：Liu Jian

Time：2021-07-09

1 Introduction

强化学习是什么

- Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics—trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning.

强化学习与有监督学习、无监督学习的区别：

- Supervised learning：Each example is a description of a situation together with a specification—the label—of the correct action the system should take in that situation, which is often to identify a category to which the situation belongs. The object of this kind of learning is for the system to extrapolate, or generalize, its responses so that it acts correctly in situations not present in the training set. This is an important kind of learning, but alone it is not adequate for learning from interaction.
- Unsupervised learning：which is typically about finding structure hidden in collections of unlabeled data. Although one might be tempted to think of reinforcement learning as a kind of unsupervised learning because it does not rely on examples of correct behavior, reinforcement learning is trying to maximize a reward signal instead of trying to find hidden structure. Uncovering structure in an agent's experience can certainly be useful in reinforcement learning, but by itself does not address the reinforcement learning problem of maximizing a reward signal.
- 强化学习器需要：感知状态、采取行动、达成目标
    - One of the challenges that arise in reinforcement learning, and not in other kinds of learning, is the trade-off between exploration and exploitation. The agent has to exploit what it has already experienced in order to obtain reward, but it also has to explore in order to make better action selections in the future. For now, we simply note that the entire issue of balancing exploration and exploitation does not even arise in supervised and unsupervised learning, at least in the purest forms of these paradigms.
    - Another key feature of reinforcement learning is that it explicitly considers the whole problem of a goal-directed agent interacting with an uncertain environment.

强化学习的四要素：

- 策略 A policy: defines the learning agent's way of behaving at a given time.
- 奖赏 A reward signal: defines the goal of a reinforcement learning problem.
- 价值函数 A value function: specifies what is good in the long run, Whereas the reward signal indicates what is good in an immediate sense.

- 环境 A model of the environment: Methods for solving reinforcement learning problems that use models and planning are called model-based methods, as opposed to simpler model-free methods that are explicitly trial-and-error learners—viewed as almost the opposite of planning. Modern reinforcement learning spans the spectrum from low-level, trial-and-error learning to high-level, deliberative planning.

基于价值函数的强化学习方法与进化法 (evolutionary methods) 的区别：

- Our focus is on reinforcement learning methods that learn while interacting with the environment, which evolutionary methods do not do.
- Evolutionary and value function methods both search the space of policies, but learning a value function takes advantage of information available during the course of play:
    - To evaluate a policy, an evolutionary method holds the policy fixed and plays many games against the opponent or simulates many games using a model of the opponent. The frequency of wins gives an unbiased estimate of the probability of winning with that policy, and can be used to direct the next policy selection. But each policy change is made only after many games, and only the final outcome of each game is used: what happens during the games is ignored. For example, if the player wins, then all of its behavior in the game is given credit, independently of how specific moves might have been critical to the win. Credit is even given to moves that never occurred!
    - Value function methods, in contrast, allow individual states to be evaluated.

Some of the key features of reinforcement learning methods:

- there is the emphasis on learning while interacting with an environment.
- there is a clear goal, and correct behavior requires planning or foresight that takes into account delayed effects of one's choices.
- It is a striking feature of the reinforcement learning solution that it can achieve the effects of planning and lookahead without using a model of the opponent (这里的 opponent 也就是指环境) and without conducting an explicit search over possible sequences of future states and actions.

强化学习如何使用神经网络处理状态数量巨大甚至无穷的问题：

- The artificial neural network provides the program with the ability to generalize from its experience, so that in new states it selects moves based on information saved from similar states faced in the past, as determined by the network. How well a reinforcement learning system can work in problems with such large state sets is intimately tied to how appropriately it can generalize from past experience. It is in this role that we have the greatest need for supervised learning methods within reinforcement learning.

强化学习可以基于或不基于环境模型进行，它们分别称为 model-based methods 和 model-free methods。此外，Model-free methods are also important building blocks for model-based methods.

总结：

- Reinforcement learning is a computational approach to understanding and automating goal-directed learning and decision making. It is distinguished from other computational approaches by its emphasis on learning by an agent from direct interaction with its environment, without requiring exemplary supervision or complete models of the environment.
- Reinforcement learning uses the formal framework of Markov decision processes to define the interaction between a learning agent and its environment in terms of states, actions, and rewards.

- The use of value functions distinguishes reinforcement learning methods from evolutionary methods that search directly in policy space guided by evaluations of entire policies.

强化学习研究的三条发展轨迹：

- learning by trial and error
- optimal control
- temporal-difference learning methods

All three threads came together in the late 1980s to produce the modern field of reinforcement learning as we present it in this book.
