Chapter 1: Introduction - Highlights

Jacob Taylor Cassady

November 22, 2022

1 Reinforcement Learning

- reinforcement learning is trying to maximize a reward signal instead of trying to find hidden structure.
- 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.
- All reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments.
- When reinforcement learning involves planning, it has to address the interplay between planning and real-time action selection, as well as the question of how environment models are acquired and improved. When reinforcement learning involves supervised learning, it does so for specific reasons that determine which capabilities are critical and which are not.

2 Examples

- All involve interaction between an active decision-making agent and its environment, within which the agent seeks to achieve a goal despite uncertainty about its environment.
- Correct choice requires taking into account indirect, delayed consequences of actions, and thus may require foresight or planning.
- The knowledge the agent brings to the task at the start—either from previous experience with related tasks or built into it by design or evolution—influences what is useful or easy to learn, but interaction with the environment is essential for adjusting behavior to exploit specific features of the task.

3 Elements of Reinforcement Learning

- one can identify four main subelements of a reinforcement learning system: a policy, a reward signal, a value function, and, optionally, a model of the environment.
- A policy defines the learning agent's way of behaving at a given time. Roughly speaking, a policy is a mapping from perceived states of the environment to actions to be taken when in those states.
- A reward signal defines the goal of a reinforcement learning problem. On each time step, the environment sends to the reinforcement learning agent a single number called the reward.
- In a biological system, we might think of rewards as analogous to the experiences of pleasure or pain.
- Whereas the reward signal indicates what is good in an immediate sense, a value function specifies what is good in the long run. Roughly speaking, the value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state.
- Whereas rewards determine the immediate, intrinsic desirability of environmental states, values indicate the long-term desirability of states after taking into account the states that are likely to follow and the rewards available in those states.
- To make a human analogy, rewards are somewhat like pleasure (if high) and pain (if low), whereas values correspond to a more refined and farsighted judgment of how pleased or displeased we are that our environment is in a particular state.
- We seek actions that bring about states of highest value, not highest reward, because these actions obtain the greatest amount of reward for us over the long run.
- In fact, the most important component of almost all reinforcement learning algorithms we consider is a method for efficiently estimating values.
- final element of some reinforcement learning systems is a model of the environment. This is something that mimics the behavior of the environment, or more generally, that allows inferences to be made about how the environment will behave.
- Models are used for planning, by which we mean any way of deciding on a course of action by considering possible future situations before they are actually experienced. 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.

4 Limitations and Scope

- The policies that obtain the most reward, and random variations of them, are carried over to the next generation of policies, and the process repeats. We call these evolutionary methods because their operation is analogous to the way biological evolution produces organisms with skilled behavior even if they do not learn during their individual lifetimes.
- If the space of policies is sufficiently small, or can be structured so that good policies are common or easy to find—or if a lot of time is available for the search—then evolutionary methods can be effective.
- In addition, evolutionary methods have advantages on problems in which the learning agent cannot sense the complete state of its environment.
- Methods able to take advantage of the details of individual behavioral interactions can be much more efficient than evolutionary methods in many cases.
- Evolutionary methods ignore much of the useful structure of the reinforcement learning problem: they do not use the fact that the policy they are searching for is a function from states to actions; they do not notice which states an individual passes through during its lifetime, or which actions it selects.

5 An Extended Example: Tic-Tac-Toe

- Here, a policy is a rule that tells the player what move to make for every state of the game—every possible configuration of Xs and Os on the three-by-three board.
- For each policy considered, an estimate of its winning probability would be obtained by playing some number of games against the opponent.
- This evaluation would then direct which policy or policies were considered next.
- First we would set up a table of numbers, one for each possible state of the game. Each number will be the latest estimate of the probability of our winning from that state. We treat this estimate as the state's value, and the whole table is the learned value function. State A has higher value than state B, or is considered "better" than state B, if the current estimate of the probability of our winning from A is higher than it is from B. Assuming we always play Xs, then for all states with three Xs in a row the probability of winning is 1, because we have already won. Similarly, for all states with three Os in a row, or that are filled up, the correct probability is 0, as we cannot win from them. We set the initial values of

all the other states to 0.5, representing a guess that we have a 50% chance of winning.

- Most of the time we move greedily, selecting the move that leads to the state with greatest value, that is, with the highest estimated probability of winning. Occasionally, however, we select randomly from among the other moves instead. These are called exploratory moves because they cause us to experience states that we might otherwise never see.
- We attempt to make them more accurate estimates of the probabilities of winning. To do this, we "back up" the value of the state after each greedy move to the state before the move
- This can be done by moving the earlier state's value a fraction of the way toward the value of the later state.
- alpha is a small positive fraction called the step-size parameter, which influences the rate of learning.
- This update rule is an example of a temporal-difference learning method, so called because its changes are based on a difference, V (St+1)−V (St), between estimates at two successive times.
- For example, if the step-size parameter is reduced properly over time, then this method converges, for any fixed opponent, to the true probabilities of winning from each state given optimal play by our player.
- If the step-size parameter is not reduced all the way to zero over time, then this player also plays well against opponents that slowly change their way of playing.
- In the end, 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.
- 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 and without conducting an explicit search over possible sequences of future states and actions.
- It is just as applicable when behavior continues indefinitely and when rewards of various magnitudes can be received at any time.
- 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 with reinforcement learning.
- We also have access to the true state in the tic-tac-toe example, whereas reinforcement learning can also be applied when part of the state is hidden, or when different states appear to the learner to be the same.

- A model is not required, but models can easily be used if they are available or can be learned
- Because models have to be reasonably accurate to be useful, model-free methods can have advantages over more complex methods when the real bottleneck in solving a problem is the difficulty of constructing a sufficiently accurate environment model.
- Model-free methods are also important building blocks for model-based methods.
- In hierarchical learning systems, reinforcement learning can work simultaneously on several levels.

6 Summary

- Reinforcement learning is a computational approach to understanding and automating goal-directed learning and decision making.
- 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.
- We take the position that value functions are important for efficient search in the space of policies.