Introduction to Reinforcement Learning

Diego Ascarza-Mendoza

Escuela de Gobierno y Transformación Pública

Outline

Reinforcement Learning

Elements of Reinforcement Learning

Tic-Tac-Toe

Introduction

- Learning by interacting with our environment is the first to occur to us when thinking about learning.
- This connection produces a wealth of information about cause and effect, consequences of actions, and what to do to achieve goals.
- If we are learning to drive a car, a videogame, or to hold a conversation, we are aware of how our environment responds to what we do.
- Learning from interaction is one of the most basic ideas behind learning and intelligence. We take a computational approach on this.

Reinforcement learning

- Reinforcement learning is learning what to do: mapping situations to actions.
- The goal is to maximize a numerical reward signal.
- The learner must discover which actions yield the most reward by trying them. Actions might affect not only what happens today but also what happens tomorrow.
- In other words, the most important distinguishing features of reinforcement learning are:
    1. Trial-and-error search.
    2. Delayed reward.

Reinforcement learning - formalization

- We formalize the problem of reinforcement learning using ideas from optimal control of incompletely-known Markov decision processes (MDPs) (wtf is that?).
- The basic idea is to capture the problem of an agent that interacts with its environment to achieve a goal.
- The agent must be able to know the environment to some extent and be able to affect the situation through its actions.
- The agent must also have a goal related to the state of the environment.
- Any method that is intended to deal with: sensation, action, and goal will be considered a reinforcement learning method.

Exploration vs Exploitation

- A feature that characterizes reinforcement learning is the presence of the trade-off between **exploration and exploitation**.
- To obtain more reward, a RL agent must preferred actions chosen in the past and found to be effective in producing reward (exploitation).
- BUT, to discover those actions, it has to try actions that has not selected before (exploration).
- Dilemma: neither exploration nor exploitation can be pursued exclusively without failing the task.

Reinforcement learning departs from the whole thing

- RL explicitly considers the whole problem of a goal-directed agent interacting with an uncertain environment.
- All RL agents have explicit goals, can sense aspects of their environment, and can choose actions to influence it.
- Furthermore, it is usually assumed that from the beginning, the agent has to operate despite significant uncertainty.
- Importantly, a complete, interactive, goal-seeking agent does not always mean something like a complete organism or robot. It can also be a component of a larger behaving system.
- For instance, it can be an agent that monitors the charge level of the robot's battery and sends commands to the robot's control architecture.

Why is it worth it learning RL?

- It has lots of interactions with other engineering and scientific disciplines, including economics!
- RL is part of decades-long trend within AI and ML toward greater integration with statistics, optimization, and other cool mathematical subjects.
- For instance, the ability of RL methods to learn with parameterized approximators addresses the classical ’curse of dimensionality’ often found in complex macroeconomic models.
- RL is also part of a larger trend in AI back toward simple general principles. Before, methods based on search learning were characterized as ’weak methods’. This is changing; we just did not put enough effort into it.

Examples

- A master chess player makes a move: choice informed by planning-anticipating replies and counterreplies, and by immediate, intuitive judgments of the desirability of particular positions and moves.
- A gazelle calf struggles to its feet minutes after being born. Half an hour later it is running at 20 miles per hour.
- A mobile robot decides whether it should enter a new room in search of more trash to collect or start trying to find its way back to its battery recharging station.
- An individual who faces uncertainty in terms of income and medical expenses, with a certain level of wealth, decides how much to consume and how much to save.

What are common features in these examples?

- All of them involve interaction between an active decision-making agent and its environment.
- Within the environment, 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 effects of actions cannot be fully predicted (it must monitor the environment frequently and react appropriately).
- The agent can use its experience to improve its performance over time. It learns to identify what is useful and what is not from interacting with its environment.

Outline

Reinforcement Learning

Elements of Reinforcement Learning

Tic-Tac-Toe

Elements of Reinforcement Learning

There are four main subelements of a reinforcement learning system beyond the agent and the environment:

1. Policy
2. Reward signal
3. Value function
4. Model of the environment (optional)

Elements of Reinforcement Learning - Policy

- A policy defines the learning agent's way of behaving at a given time.
- It is a mapping from perceived states of the environment to actions to be taken when in those states.
- It is the core of a reinforcement learning agent in the sense that it alone is sufficient to determine behavior.
- Policies may be stochastic, specifying probabilities for each action.

Elements of Reinforcement Learning - Reward signal

- Defines the goal of a reinforcement learning problem.
- At each time step, the environment sends the agent a single number, called the reward.
- Agent's sole objective is to maximize the total reward it receives over the long run.
- Reward signal defines what the good and bad events are for the agent. They are the immediate and defining features of the problem faced by the agent.
- In the biological system, this could be analogous to pleasure and pain.
- In general, Rewards may be stochastic functions of the state of the environment and the actions taken.

- While reward indicates what is good in the immediate sense, a value function specifies what is good in the long run.
- The value of a state is more or less the total amount of reward an agent can expect to accumulate over the future, starting from that state.
- 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 in those states.
- A state might always yield a low immediate reward but still have a high value because it offers higher yields in the future.

- Without rewards, there could be no values. The only purpose of estimating values is to achieve more reward.
- However, it is values with which we are most concerned when making and evaluating decisions.
- Action choices are made based on value judgments. We seek actions that bring about states of highest value NOT rewards.
- Sadly, it is much harder to determine values than it is to determine rewards.
- Rewards are directly given by the environment, but values have to be estimated and re-estimated from the sequences of observations and agent makes over its entire lifetime. Indeed, much of what we will study involves estimating values efficiently.

Elements of Reinforcement Learning - Model

- This mimics the behavior of the environment, or allows one to make inferences about how the environment will behave.
- In a model, for instance, given a state and action, it is possible to predict the resultant next state and next reward.
- For instance, in a neoclassical growth model, by choosing how much to consume, it is possible to determine how much utility we get in one period and with what capital we start next period.
- Models are used for planning: this means considering possible future situations before they are experienced, as a way of deciding on a course of action.
- Methods for solving RL problems that use models and planning are called *model-based* methods. Those who are explicitly trial-and-error learners are called *model-free* methods.

Outline

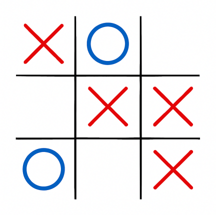
Reinforcement Learning

Elements of Reinforcement Learning

Tic-Tac-Toe

Extended example: Tic-Tac-Toe

- Consider the child's game tic-tac-toe.
- Two players take turns playing on a three-by-three board.



- One plays X's and the other plays O's. A player wins by placing three marks in a row, horizontally, vertically, or diagonally.
- If the board fills up with neither player getting three in a row, then the game is a draw.

Tic-tac-toe

- A skilled player can play so as never to lose. Let us assume we are playing against an imperfect player (he can screw it up occasionally).
- Let's assume that a draw is equally bad for us as losing. How might we construct a player that will find the imperfections in its opponent's play and learn to maximize its chances of winning?
- This is a simple problem, but it can not be solved with traditional methods such as "minimax" solutions. This is because minimax assumes a way of playing by the opponent.
- Other optimization methods, such as dynamic programming, can compute optimal policies for any opponent, but need a full specification of the opponent, including probabilities at which the opponent makes a move at each state.
- Suppose we do not have all this information, and it is not available for the vast majority of problems.

- The best we could do, perhaps, is to learn a model of the opponent's behavior, up to some level of confidence.
- We could then use dynamic programming to compute an optimal solution given our approximate model of the opponent.
- There is an important difference between how *evolutionary* methods would operate relative to RL methods.

Evolutionary methods

- These methods would search the space of possible policies for one with a high probability of winning against the opponent.
- A policy is a rule that tells the player what to move for every state of the game.
- For each policy, many games would be played to estimate the probability of winning. This would then tell which policies to select for each scenario.
- Hundreds of algorithms could be applied in this family. But you can see that it becomes tiring in the policy space.

How would RL approach it?

- A method using a value function would set up a table of numbers, one for each possible state.
- Each number would be our last estimate of the probability of winning the game from that state.
- This estimate is the state's *value*. The whole table is the learned value function.
- Assuming we play X's, we say that if we have three X's in a row, probability of winning is one by definition.
- If there are three O's in a row, the probability of winning is zero, and in the rest of the scenarios, we say the probability of winning is 0.5.

How would RL approach it?

- We then play many games against the opponent. To select our moves, we examine the states that would result from each possible move and consider their current values in the table.
- Most of the time we move greedily, selecting actions that lead to the states with the highest likelihood of winning.
- Occasionally, we select randomly from among the other moves instead (exploration). These are exploratory moves because they cause us to experience states that otherwise we might not see.
- While we play, we update the value of the states in which we find ourselves in the game.
- We attempt to make these estimates more accurate estimates of the probability of winning. We basically update the value of a state as follows:

$$V(S_t) \leftarrow V(S_t) + \alpha \left(V(S_{t+1}) - V(S_t) \right), \tag{1}$$

- S_t is thought as the state before the greedy move, and S_{t+1} as the move after the greedy move. α is a step size parameter that influences the rate of learning.

What do we learn from this?

- The example above shows the difference between RL methods and evolutionary methods.
- An evolutionary method keeps the policy fixed and plays many games, using a model of the opponent. Each policy change is made only after playing many games. Credit is given to moves that never occurred.
- RL emphasizes learning while interacting with the environment.
- Also, there is a clear goal, and correct behavior requires planning or foresight that takes into account delayed effects of one's choices.
- It is striking that RL can achieve the effects of planning and lookahead without a model of the opponent and without conducting an explicit search over possible sequences of future states and actions.

Summary

- RL is a computational approach to understanding and automating goal-directed learning and decision making.
- Direct interaction with environment is what makes this discipline different.
- It uses a formal framework of MDPs to define the interaction between agent and its environment in terms of states, actions and rewards.
- The key concepts are value and value function. Value functions are important for efficient search in the space of policies.