Introduction to Reinforcement Learning – Part 1

Introduction:

Two key distinguishing features of reinforcement learning:

1. Exploration-Exploitation dilemma (trial and error search)
    - Exploration: Exploring potential hypotheses for how to choose actions
    - Exploitation: Exploiting limited knowledge about what is already known should work well
2. Delayed reward (maximizing rewards over time)
    - Actions now can impact rewards not just right now but in future time steps as well

The problem of reinforcement learning can be formalized as the optimal control of incompletely-known Markov decision processes

- Overview:
    - Capture the most important aspects of the real problem facing a learning agent interacting over time with its environment to achieve a goal
    - The learning agent must be able to sense the state of its environment to some extent and must be able to take actions that affect the state
    - The agent also must have a goal or goals relating to the state of the environment
- Markov decision processes are intended to include just three aspects:
    - Sensation
    - Action
    - Goal

Reinforcement learning is different from supervised learning as well as unsupervised learning

- In supervised learning, learning is from a training set of labeled examples. However, in interactive problems, it is often impractical to obtain examples of desired behavior that are both correct and representative of all the situations in which the agent has to act.
- In unsupervised learning, learning is by finding structure hidden in collections of unlabeled data. This is different from reinforcement learning, as 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.
- Thus, reinforcement learning is third machine learning paradigm, alongside supervised and unsupervised learning.
- All reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments

Examples of reinforcement learning:

1. A master chess player makes a move
2. An adaptive controller adjusts parameters of a petroleum refinery's operation in real time
3. A calf learning to run
4. A mobile robot decides whether it should enter a new room in search of more trash to collect or start trying to finds its way back to its battery recharging station
5. Anyone preparing a meal

All these examples 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. The agent's actions are

permitted to affect the future state of the environment, thereby affecting the actions and opportunities available to the agent at later times. Correct choice requires taking into account indirect, delayed consequences of actions, and thus may require foresight or planning.

At the same time, in all these examples the effects of actions cannot be fully predicted; thus, the agent must monitor its environment frequently and react appropriately. Furthermore, the agent can use its experience to improve its performance over time. The knowledge the agent brings to the tasks 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.

Elements of reinforcement learning:

1. Agent
2. Environment state
    - Signal conveying to the agent some sense of "how the environment is" at a particular time
    - It is an input to the policy and value function, as well as an input to and output from the model
    - The state signal is produced by some preprocessing system that is nominally part of the agent's environment
3. 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
    - Analogous to "a set of stimulus-response rules or associations" (in psychology)
    - Can either be a simple function or lookup table OR it may involve extensive computation such as a search process
    - It is core to a reinforcement learning agent as it alone is sufficient to determine behavior
    - In general, policies may be stochastic, specifying probabilities for each action
4. Reward signal
    - Defines the goal of a reinforcement learning problem
    - One each time step, the environment sends to the reinforcement learning agent a single number called the reward
    - The agent's sole objective is to maximize the total reward it receives over the long run
    - The reward signal thus defines what are the good and bad events for the agent
    - Analogous to "experiences of pleasure or pain" (in biological systems)
    - They are the immediate and defining features of the problem faced by the agent
    - The reward signal is the primary basis for altering the policy; if an action selected by the policy is followed by low reward, then the policy may be changed to select some other action in that situation in the future
    - In general, reward signals may be stochastic functions of the state of the environment and the actions taken
5. Value function
    - While the reward function indicates what is good in the immediate term, a value function specifies what is good in the long run
    - The value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state
    - While 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

- Analogous to "humans having a more refined and farsighted judgement of how pleased or displeased they are that their environment is in a particular state"
- Rewards are primary, whereas values as predictions of rewards are secondary
    i. Without rewards there could be no values, and the only purpose of estimating values is to achieve more reward
    ii. Nevertheless, it is values with which we are most concerned when making and evaluating decisions
    iii. Action choices are made based on value judgements
    iv. We seek actions that bring about states of highest value, not highest reward, because these actions obtain the greatest amount of reward over the long run
    v. However, it is much harder to determine values vs. rewards; values must be estimated and re-estimated from the sequence of observations an agent makes over its entire lifetime
- As a result, the most important component of almost all reinforcement learning algorithms is a method for efficiently estimating values

6. Model of the environment (optional)
    - 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 – 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

Limitations and scope:

- The focus is on reinforcement learning methods that learn from interacting with the environment, which evolutionary methods do not do
    - Evolutionary methods have advantages on problems in which the learning agent cannot sense the complete state of its environment

Reference:

- *Sutton, Richard S. and Barto, Andrew G., Reinforcement Learning – An Introduction, Second Edition, The MIT Press, 2018 Draft*