Multi-Armed Bandit for Solving Quadratic Assignment Problem

Reinforcement Learning

Milad Khademi Nori

November 14, 2018

Contents

Introduction to Reinforcement Learning

- Origin
  - "A gazelle calf struggles to its feet minutes after being born. Half an hour later it is running at 20 miles per hour." Sutton and Barto

Introduction to Reinforcement Learning

- Goals
  - Agents interacts dynamically with its environment, moves from one state to another.
  - Based on the actions taken by the agent, rewards are given.
  - Guidelines for which action to take in each state is called a policy.
  - Try to efficiently find an optimal policy in which rewards are maximized.
- Achievement
  - Google's AlphaGo used deep reinforcement learning in order to defeat world champion Lee Sedol at Go. In Go number of possible games is larger than the number of atoms in the universe and it is much more challenging than chess.

Reinforcement Learning vs Supervised Learning

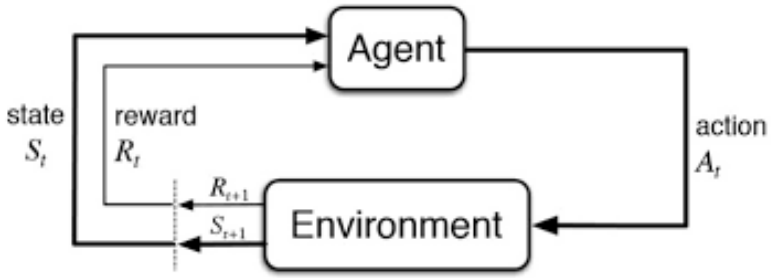
- Supervised Learning
  - Learning from examples (Dataset) provided by knowledgeable external supervisor.
  - For any state that the agent may be in, the supervisor can supply enough relevant examples of the outcomes which result from similar states so that we may make an accurate prediction.
- Reinforcement Learning
  - No supervisor exists.
  - Agent must learn from experience as it explore the range of possible states.

Reinforcement Learning vs Supervised Learning

- Supervised Learning
  - Learning from examples (Dataset) provided by knowledgeable external supervisor.
  - For any state that the agent may be in, the supervisor can supply enough relevant examples of the outcomes which result from similar states so that we may make an accurate prediction.
- Reinforcement Learning
  - No supervisor exists.
  - Agent must learn from experience as it explore the range of possible states.

Introduction to Reinforcement Learning

- Reinforcement Learning

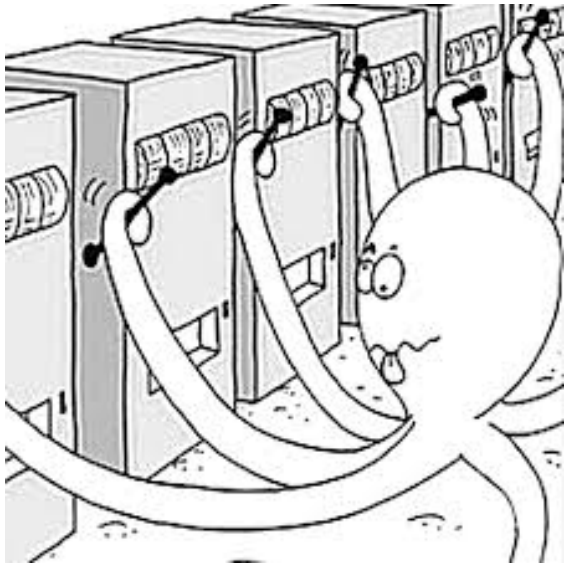


- Examples

Agent	Environment	Actions	Rewards	Policy
Board game player	Set of all game configs.	Legal	Winning the game	Optimal strategy
Mouse	Maze	Running & turning	Cheese	Most direct path to cheese

Introduction to Reinforcement Learning

- Exploration & Exploitation
  - In the absence of a supervisor, the agent must explore the environment in order to gain information about rewards, while exploiting its current information to maximize its rewards.
  - Balancing this tradeoff is a common theme
- A Single State Example: Multi-Armed Bandit Problem



Multi-Armed Bandit (MAB) Problem

- Multi-Armed Bandit Problem
  - Given N different arms to choose from, each with an unknown reward, what strategy should we use to explore and learn the values of each arm, while exploiting our current knowledge to maximize profit?
  - This is a very common approach for optimizing online marketing campaigns.
  - This can be thought of as a single-state reinforcement learning problem.
- MAB Solvers
  - Epsilon Greedy
  - Upper Convergence Bound (UCB)

Multi-Armed Bandit (MAB) Problem

- Upper Convergence Bound (UCB)

$$Score_j^t = \bar{x}_j^t + \sqrt{\frac{c \times ln \sum_k p_k^t}{p_j^t}} \tag{1}$$

- The first term in the formula, $\bar{x}_j^t$, encodes the expected average reward for arm j according to knowledge available in time-step t.
- Always choosing the arm with the highest expected reward would result in a purely exploitative algorithm, so the formula includes a second term to deal with exploration.
- The variable p_j^t represents the number of times arm j has been pulled at time-step t, making the value of the second term in formula inversely proportional to the arm popularity.