Advanced Deep Learning and Reinforcement Learning

January 13, 2019

Core concepts

- Environment
- Reward signal
- Agent
  - Agent state
  - Policy
  - Value function
  - Model

Reward

A reward Rt is a scalar feedback signal

The agent's job is to maximize cumulative reward

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots$$

Value

Expected cumulative reward, from a state s

$$\begin{aligned} v(s) &= \mathbb{E}[G_t | S_t = s] \\ &= \mathbb{E}[R_{t+1} + R_{t+2} + R_{t+3} + \dots | S_t = s] \end{aligned}$$

The actual value function is the Expected return:

$$\begin{aligned} v(s) &= \mathbb{E}[G_t | S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s] \end{aligned}$$

the discount factor $\gamma \in [0, 1]$ trades off importance of immediate vs long-term rewards.

That leads to Bellman equation

$$\begin{aligned} v(s) &= \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t \sim \pi(s)] \\ &= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t \sim \pi(s)] \end{aligned}$$

Actions in sequential problems

Goal: select actions to maximize value

A mapping from states to actions is called a Policy

$$\begin{aligned} q(s, a) &= \mathbb{E}[G_t | S_t = s, A_t = a] \\ &= \mathbb{E}[R_{t+1} + R_{t+2} + R_{t+3} + \ldots | S_t = s, A_t = a] \end{aligned}$$

Agent State

The state including agent state and environment state

A history is a sequence of observations, actions, rewards

$$H_t = O_0, A_0, R_1, O_1, \ldots, O_{t-1}, A_{t-1}, R_t, O_t$$

This history can be used to construct an agent state S_t

Fully Observable Environments

Observation = environment state

The agent state could just be this observation:

S_t = O_t=environment state

Then the agent is in a Markov decision process:

$$p(r, s | S_t, A_t) = p(r, s | H_t, A_t)$$

"The future is independent of the past given the present"

Partially Observable Environments

Policy

Defines the agent's behaviour

Deterministic policy: $A = \pi(S)$

Stochastic policy: $\pi(A|S) = p(A|S)$