Reinforcement learning - part II

Announcements

1. Gradiance is due on Sunday night
2. For finals, a practice quiz and list of topics will be posted soon
3. Assignment 3 due this week on Friday

RL

↳ Markov Decision Processes (MDP)

$$p(s', r \mid s, a)$$

$$G_t = R_{t+1} + \cdots$$

Agent

Policy (π)

$$\pi(a|s) \rightarrow A_t = a \mid S_t = s$$

value function at state s

$$V_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$$

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = s]$$

Complete model of the environment.

$$p(s', r \mid s, a)$$

Dynamic Programming (DP) based methods

Init: Start with a policy

Step 1. Find $V_\pi(s)$ $\forall s \in S$

Policy prediction

Step 2 $\pi \Longrightarrow V_\pi(s) \longrightarrow \pi'$

Policy Improvement

$p(s', r \mid s, a)$

DP → works if we ~~have a perfect~~ know the environment model completely $p(s', r | s, a)$

→ Data-free

→ not very useful

Monte-carlo methods.

Policy prediction

Policy Control

---

first visit MC prediction

S_1 $V(S_1) = 1$ $Returns(S_1) = [\,]$

S_2 $V(S_2) = 1$ $Returns(S_2) = [\,]$

$S_1 A_0 R_1 S_1 A_1 R_2 S_2 A_1 R_2 S_1 A \ldots$

$V(S_t)$

$Q(S_t, a_t)$

PlayerTotal, AceUsable, HouseCard

S_1	11	0	6
S_2	13	0	6
S_3	17	0	6
S_4	18	0	6
S_5	22	0	6

Returns $(22, 0, 6) = [-1]$

Returns $(18, 0, 6) = [\ -\ ]$

$V(\_\_) = -1$,

+1	0	6