Note for Reinforcement Learning 2nd Edition

July 24, 2018

Finite Markov Decision Process

MDP framework consists of an environment and agent. For $t = 0, 1, 2...t$, agent receives observed state $S_t \in \mathcal{S}$ based on which the agent perform an action $A_t \in \mathcal{A}$. The dynamic of the environemtn then returns a reward $R_{t+1} \in \mathcal{R}$. This form a trajectory $S_0, A_0, R_1, S_1, A_1, R_2, \ldots$. The dynamic for MDP is defined to be

$$p(s', r \mid s, a) = Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

Several commonly use quantities:
state-transition probability

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

expected reward for state-action pair

$$\begin{aligned} r(s, a) &= E[R_t \mid S_{t-1} = s, A_{t-1} = a] \\ &= \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a) \end{aligned}$$

expected reward for state-action-next-state triples

$$\begin{aligned} r(s, a, s') &= E[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] \\ &= \sum_{r \in \mathcal{R}} r p(r \mid s, a, s') \\ &= \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a, s')} \end{aligned}$$