687 2017-09-07

Note Title 9/7/2017

More about Sutton & Barto book 2nd Ed. in progress

$M = (\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma)$

MDP $t = \{0, 1, \ldots\}$ time step

$\mathcal{S}$: set of states can be finite, discrete, infinite

$\mathcal{A}$: possible actions usually finite

P : transition fn how states transition $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$

$P(s,a,s') \triangleq Pr(S_{t+1} = s' \mid S_t = s \wedge A_t = a)$

Can be deterministic.

R : reward fn $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ (most general)

$R(s,a,s') \triangleq E[R_t \mid S_t = s \wedge A_t = a \wedge S_{t+1} = s']$

d_0 : initial state distribution $d_0 : \mathcal{S} \to [0,1]$ (probabilities)

$d_0(s) \triangleq Pr(S_0 = s)$

γ : reward discount parameter $\gamma \in [0,1]$

Formalating the Agent for an MDP

Policy: Mechanism that determines which action to take in a state.

Learning: Changing the agent's policy

$\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ probabilities $\pi(s,a) \triangleq Pr(a_t = a \mid s_t = s)$

Can be deterministic or not; **deterministic** $\pi(s,a) \in \{0,1\}$ (else stochastic)

$\pi(s,a)$ is **stationary** if it does not depend on t (?)

$M = (\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma)$

① $s_0 \sim d_0$ ~ means "sampled from"

② $a_0 \sim \pi(s_0, \cdot)$

③ $s_1 \sim P(s_0, a_0, \cdot)$

④ $r_0 := E[R_0] = R(s_0, a_0, s_1)$

etc. $a_1, s_2, r_1, \ldots$

can define $P^\pi(s,s')$ – transition probs.

If $|\mathcal{S}| + |\mathcal{A}|$ are finite + R_t is bounded then an optimal policy exists.

deterministic means $a_0 \sim \pi(s_0, \cdot)$, $a_1 \sim \pi(s_1, \cdot)$ etc.

Goal: Find policy that maximizes **expected** reward

Objective fn: $J : \Pi \to \mathbb{R}$ (Π: policies, $\mathbb{R}$: quality)

$J(\pi) \triangleq E\left[\sum_{t=0}^{\infty} R_t \mid \pi\right]$

Optimal $\pi^* \in \arg\max_{\pi \in \Pi} J(\pi)$ can be more than one