Lecture 12. RL and Control.

- Introduction:

  (1) no explicit supervision.

  (2) only reward function.

  (3) formalism: MDP

- Markov Decision Process:

  (1) Definition:

  - S: state set
  - A: action set
  - $P_{(s,a)}$: transition probability.

    in state s, take action a.
  - $\gamma \in [0,1)$: discount factor.
  - $R: S \times A \to \mathbb{R}$. reward function.

  (2) dynamics:

  $$S_0 \xrightarrow{a_0} S_1 \xrightarrow{a_1} S_2 \cdots$$

  $$R = R(S_0, a_0) + \gamma R(S_1, a_1) + \gamma^2 R(S_2, a_2) + \cdots$$

  or only write the states:

  $$R = \sum_{k=0}^{+\infty} \gamma^k R(S_k)$$

  *: maximize target:

  $$E\left[\sum_{k=0}^{+\infty} \gamma^k R(S_k)\right].$$

(3) policy: $\pi : S \to A$. $a = \pi(s)$

$$V^\pi(s) = E[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \mid s_0 = s, \pi].$$

*: Bellman Equation:

$$V^\pi(s) = R(s_0) + \gamma[\sum P(s' \mid s, \pi(s)) V^\pi(s')].$$

two terms:

① immediate reward.

② expected sum. (after first step)

*: $|S|$ is finite.

(4) optimal value function:

$$V^*(s) = \max_\pi V^\pi(s)$$

$$= R(s) + \max_{a \in A} \gamma \sum P(s' \mid s, a) V^*(s')$$

*: maximum over a.

$$V^*(s) = V^{\pi^*}(s) \geq V^\pi(s)$$

$$\pi^*(s) = \arg\max_{a \in A} \sum_{s' \in S} P(s' \mid s, a) V^*(s')$$

◦ Value Iteration and Policy Iteration:

(1) assumption:

$|S| < \infty$. $|A| < \infty$.

(2) value iteration:

$$V(s) := R(s) + \gamma \max \sum P(s' \mid s, a) V(s')$$

① synchronous update.

② asynchronous update.

converge to V^*

use V^* to find optimal policy.

(3) policy iteration:

$V := V^{\pi}$

$\pi(s) = \arg\max_{a \in A} \sum P(s'|a,s) V(s')$

*: updating policy with current value function.

(4) for small MDP:

policy iteration is faster.

for larger MDP: $|S| \uparrow$ $|A| \uparrow$

value iteration is preferred.

◦ Learning a model for MDP:

(1) so far discussion: known:

transition probability / reward function

*: estimate from data.

(2) trials → MLE.

estimate $P(s,a)$.

average reward → $R(s)$

(3) sampling ↔ optimizing

exploration ↔ exploitation.