동적계획법과 강화학습 강의 노트

작성자 : 이진우(2021324002)

작성일시 : 2022.09.19

Lecture 2. Value Evaluation and Policy Iteration

1. Optimal value function $V_*(s)$: 주어진 state s에서 모든 가능한 policy들에 따른 value function 값들 중 가장 큰 값을 지칭한다.

$$V_*(s) = \max_{\pi} V_\pi(s)$$

2. Optimal action-value function $q_*(s,a)$: 주어진 state s와 action a에서 모든 가능한 policy들에 따른 value function 값들 중 가장 큰 값을 지칭한다.

$$q_*(s,a) = \max_{\pi} q_\pi(s,a)$$

3. Markov Decision Process(MDP)의 해를 찾는 것은 모든 state 에 대해 최적의 optimal value 값을 주는 policy π_*를 찾는 것과 동일하다.

4. $V_*(s)$를 또 다르게 표현할 수 있는데 – action value function $q_*(s,a)$에 대한 식으로 나타낼 수 있다 :

$$V_*(s) = \max_a q_*(s,a) = \max_a \{R(s,a) + \gamma \sum_{s'} P^a_{ss'} V_*(s')\}$$

   4.1. 직관적으로 보자면, 주어진 state s 에서 취할 수 있는 action 들 중 action value function 값을 최대화하는 action a 를 greedy 하게 찾는 것으로 볼 수 있다.

5. Finite MDP 하에서는 unique optimal policy, 즉 모든 state 에 대한 optimal value function 들을 찾을 수 있는데, state 갯수 만큼의 unknown value function 값들과 각각에 대한 linear equation 이 존재하기 때문에 matrix inversion 같은 방법으로 direct solution 을 찾을 수 있다.

   5.1. 그러나 아주 많은 개수의 state 들과 action 들에 대해서는 direct solution 찾는 computation cost 가 매우 크기 때문에 Policy Iteration 과 같은 approximation 방법을 활용한다.

6. Policy Iteration 을 하기 위해 먼저 임의로 주어진 policy π 에 대한 다음의 expected value function 값들이 필요하다 :

$$V_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} P^{\pi(s)}_{ss'} V_*(s')$$

   6.1. 이때 다음의 iterative process 를 통해 각 state s 에 대한 fixed point 을 찾는다 :

$$V_{k+1}(s) = \sum_a \pi(a|s)\{R_s^a + \gamma \sum_{s'} P_{ss'}^a V_k(s')\}$$

적절한 초기값 $V_0(s)\ \forall\ s \in S$ 들과 함께 위의 과정을 반복하여 $V_1 \rightarrow V_2 \rightarrow \cdots \rightarrow fixed\ point\ V_\pi$을 구할 수 있다. 이러한 과정을 Policy evaluation 이라 부른다.

7. 이후 현재 구해진 value function 들을 바탕으로 Policy improvement 를 시행한다.

[참고문헌]

- Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.