Chapter. 01

마르코프 결정과정

| 강화학습의 놀이터: MDP

FAST CAMPUS ONLINE
강화학습 A-Z I

강사. 박준영

I 지난 이야기…

- 마르코프 특성 (Markov property)
- 마르코프 과정 (Markov processes)
  - 상태 집합 $\mathcal{S}$
  - 상태 천이 행렬 P
- 마르코프 보상 과정 (Markov reward process: MRP)
  - 보상함수 R
  - 감가율 γ
  - 리턴 G_t / 가치 함수 $V(s)$
  - Bellman 방정식
  - Bellman 방정식 풀이기법

하지만 MRP 에서는 어떤 방법으로도 우리는 상황을 좋게 만들 방법이 없었다!

마르코프 결정과정 (Markov decision processes: MDP)

마르코프 결정과정은 MRP에 행동을 추가한 확률 과정:

마르코브 결정 과정 (MDP)는 $<\mathcal{S}, \mathcal{A}, P, R, \gamma>$ 인 튜플이다.

- $\mathcal{S}$ 은 (유한한) 상태의 집합
- $\mathcal{A}$ 는 (유한한) 행동의 집합
- P 는 상태 천이 행렬, $P_{ss'}^{a} = P[S_{t+1} = s' | S_t = s, A_t = a]$ 더 이상 매트릭스로 표현 불가능! 3d 구조로 표현해야 함.
- R 는 보상 함수, $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$
- γ 는 감소율, $\gamma \in [0,1]$ 0이상 1이하의 모든 실수 중 하나!

비로소, 환경에 **행동을 가함**으로써 미래의 상태와, 보상을 바꿀 수 있게 됨!

I 정책 함수 (Policy function)

정책함수 π는 현재 상태에서 수행 할 행동의 확률 분포이다.

$$\pi(a|s) = P(A_t = a|S_t = s)$$

- 강화학습의 에이전트는 현재 상태 s_t 를 활용하여, 현재의 행동 a_t 를 결정한다.
- s_t 를 아는 것이 역사를 아는것과 동일하다는 Markov 특성을 가정하였으므로, 현재 상태만을 가지고 의사결정을 해도 충분.

I MDP, MRP 와 MP 의 관계

MDP $<\mathcal{S}, \mathcal{A}, P, R, \gamma>$ 와 정책 π가 결정 됐을 때,

- $S_0, S_1, S_2, \dots$ 는 마르코프 과정이다.
- $S_0, R_1, S_1, R_2, S_2, \dots$ 는 마르코프 보상 과정 $<\mathcal{S}, P^\pi, R^\pi, \gamma>$ 이다.

$$P^\pi_{ss'} = \sum_{a \in \mathcal{A}} \pi(a|s) P^a_{ss'}$$

$$R^\pi_s = \sum_{a \in \mathcal{A}} \pi(a|s) R^a_s$$

"제어 이론의 feedback control과 동일"

즉, 좋은 π 를 가지고 있다면, 최대한 많은 이득을 얻는 것이 가능하다!

I MDP의 가치함수

상태 가치함수: $V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$

(State value function)

c.f., MRP 의 가치함수 $V(s) = \mathbb{E}[G_t \mid S_t = s]$

현재 t 상태 s 에서 정책 π 를 따른다면 얻을 미래의 가치의 감가 총합

행동 가치함수: $Q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$

'상태-행동 가치함수' 라고도 불림

(State-action value function)

현재 t 상태 s 에서 a라는 행동을 취한 후, 정책 π 를 따른다면 얻을 미래의 가치의 감가 총합

네 맞습니다. Q-Learning 의 그 Q 입니다.

Bellman 기대값 방정식 (Bellman expectation equation)

상태 가치함수: $V_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t = s]$

행동 가치함수:
$$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ \dots | S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \ \dots) | S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \end{aligned}$$

$R_{t+2}, R_{t+3}, \dots$ 는 정책 π 를 따라서 행동할 때, 얻어지는 보상들입니다!

I 상태가치 함수 V 와 행동 가치함수 Q의 관계

행동 가치함수
↓
상태 가치함수

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q_\pi(s,a)$$

의미를 받아들여보세요!
현재 상태 s에서 각각의 행동 a를 할 확률이 $\pi(a|s)$ 이므로, 각 상태와 행동에 대한 가치 $Q_\pi(s,a)$를 $\pi(a|s)$로 평균을 내면 현재 상태의 가치 $V_\pi(s)$가 된다.

상태 가치함수
↓
행동 가치함수

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V_\pi(s')$$

의미를 받아들여보세요!
현재 s 상태에서 a를 행동하면, R_s^a 만큼의 보상을 받고 확률적으로 어떤 s'들에 도달하게 된다. 각 도달한 s' 에서 각각의 가치는 $V_\pi(s')$ 이니,
다음에 기대적으로 얻을 가치는 각각의 s' 에대한 확률 $P_{ss'}^a$ 과 가치를 $V_\pi(s')$ 활용해 평균을 낸 값이다. 그 후 γ 만큼 감가한다.

| 상태가치 함수 V 와 행동 가치함수 Q의 관계

1. $V_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q_\pi(s,a)$
2. $Q_\pi(s,a) = R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a V_\pi(s')$

(1)에 (2)를 대입

$$V_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\left(R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a V_\pi(s')\right)$$

$$= \sum_{a\in\mathcal{A}} \pi(a|s) R_s^a + \gamma \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} \pi(a|s) P_{ss'}^a V_\pi(s')$$

$$= R_s^\pi + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^\pi V_\pi(s')$$

$$R_s^\pi = \sum_{a\in\mathcal{A}} \pi(a|s) R_s^a$$

$$P_{ss'}^\pi = \sum_{a\in\mathcal{A}} \pi(a|s) P_{ss'}^a$$

(2)에 (1)를 대입

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a \sum_{a'\in\mathcal{A}} \pi(a'|s') Q_\pi(s',a')$$

- 현재의 상태 / 행동 s, a
- 한 스텝 뒤의 상태 / 행동 s', a'

사실 8페이지의 수식에서도 직접유도가 가능합니다 ☺

| Bellman (expectation) equation 의 행렬표현과 직접해

$$v = R^{\pi} + \gamma P^{\pi} v$$

$$v = (I - \gamma P^{\pi})^{-1} R^{\pi}$$

I Value function 를 계산하면 끝?

MDP $<\mathcal{S}, \mathcal{A}, P, R, \gamma>$ 와 정책 π가 결정 됐을때,

- $S_0, S_1, S_2, \ldots$ 는 마르코프 과정이다.
- $S_0, R_1, S_1, R_2, S_2, \ldots$ 는 마르코프 보상 과정 $<\mathcal{S}, P^\pi, R^\pi, \gamma>$ 이다.

$$P^\pi_{ss'} = \sum_{a \in \mathcal{A}} \pi(a|s) P^a_{ss'}$$
$$R^\pi_s = \sum_{a \in \mathcal{A}} \pi(a|s) R^a_s$$

"제어 이론의 feedback control과 동일"

즉, 좋은 π 를 가지고 있다면, 최대한 많은 이득을 얻는 것이 가능하다!

- 무엇이 좋은 π 의 기준일까
- 무엇이 최적의 π 를 결정지을까?

l 최적 가치 함수 (Optimal Value Function)

최적 상태 가치 함수: $V^*(s) = \max_{\pi} V_\pi(s)$

$V^*(s)$: 존재하는 모든 정책들 중에 모든 상태 s 에서 가장 높은 $V_\pi(s)$ 를 만드는 π 를 적용했을 때 얻는 $V_\pi(s)$

최적 행동 가치 함수: $Q^*(s,a) = \max_{\pi} Q_\pi(s,a)$

$Q^*(s,a)$: 존재하는 모든 정책들 중에 모든 상태 /행동 조합 (s,a) 에서 가장 높은 $Q_\pi(s,a)$ 를 만드는 π 를 적용했을 때 얻는$Q_\pi(s,a)$

최적 가치함수 (혹은 그것을 만드는 $\boldsymbol{\pi}$ 를 찾았을 때) MDP 를 '풀었다' 라고 표현.

I 최적 정책 (Optimal policy)

정책함수 π 간의 대소관계를 다음과 같이 정의 한다.

$$\pi' \geq \pi \leftrightarrow V_{\pi'}(s) \geq V_{\pi}(s), \forall\, s \in \mathcal{S}$$

최적 정책 정리 (Optimal policy theorem)
어떠한 MDP 에 대해서도 다음이 성립한다.

- 최적 정책 π^* 가 존재하며, 모든 존재하는 정책 π 에 대하여 $\pi^* \geq \pi$ 를 만족한다. (최적 정책의 존재성)
- 최적 정책들은 π^* 최적 상태 가치함수를 성취한다. 즉, $V_{\pi^*}(s) = V^*(s)$ 이다.
- 최적 정책들은 π^* 최적 행동 가치함수를 성취한다. 즉, $Q_{\pi^*}(s,a) = Q^*(s,a)$ 이다.

1) 최적 정책이 **유일**하게 존재하는 것은 아님.
2) 최적 정책을 찾으면 최적 가치함수를 찾을 수 있음.

I 최적 가치함수 중 하나를 찾는 방법

만약 최적 행동 가치함수 $Q^*(s,a) = \max_{\pi} Q_\pi(s,a)$ 를 안다면,

(존재하는 여러가지 최적 정책 중 하나를) 찾는 방법:

$$\pi^*(a|s) = \begin{cases} 1, & if\ a = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^*(s,a) \\ 0, & otherwise \end{cases}$$

Argument **Max**

$$\underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^*(s,a)$$

$Q^*(s,a)$ 들 중에서 $Q^*(s,a)$를 최대화 하는 a 를 찾아서 내놓아라.

Bellman 최적 방정식 (Bellman Optimality Equation: BOE)

$$V^*(s) = \sum_{a \in \mathcal{A}} \pi^*(a|s) Q^*(s,a)$$
$$= \max_{a \in \mathcal{A}} Q^*(s,a)$$

$$\pi^*(a|s) = \begin{cases} 1, & if\ a = \operatorname{argmax}_{a \in \mathcal{A}} Q^*(s,a) \\ 0, & otherwise \end{cases}$$

$$Q^*(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V^*(s')$$

유도방법: 벨만 기댓값 방정식에 π 자리에 π^* 대입해보세요.

Bellman 최적 방정식 (Bellman Optimality Equation: BOE)

1. $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a)$
2. $Q^*(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V^*(s')$

(1)에 (2)를 대입

$$V_\pi(s) = \max_{a \in \mathcal{A}} \left(R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V^*(s') \right)$$

(2)에 (1)를 대입

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a \max_{a' \in \mathcal{A}} Q^*(s',a')$$

- 현재의 상태 / 행동 s, a
- 한 스텝 뒤의 상태 / 행동 s', a'

I ‘Bellman optimality equation 직접해’ 는 존재하지 않습니다.

- Bellman Optimality Equation (BOE) 는 선형방정식이 아니다.
- BOE 는 일반해가 존재하지 않는다.
- 대신 반복적 알고리즘을 통해서 해를 계산할 수 있다
    - Policy evaluation / Policy iteration
    - Value iteration
    - Q-Learning
    - SARSA
    - …

ı 마무리!

- 마르코프 결정과정 (Markov decision process: MDP)
  - 정책함수 π
  - 상태 가치 함수 V_π
  - 행동 가치 함수 Q_π
- Bellman 기대값 방정식 (Bellman expectation equation: BEE)
  - V_π 와 Q_π 의 관계
- Bellman 최적 방정식 (Bellman Optimality equation: BOE)
- 최적 가치 함수 V^*, Q^*
- 최적 정책 함수 π^*
- 최적 가치함수와 최적 정책함수의 관계