Chapter. 05

어깨넘어 배워서 세상 조종하기

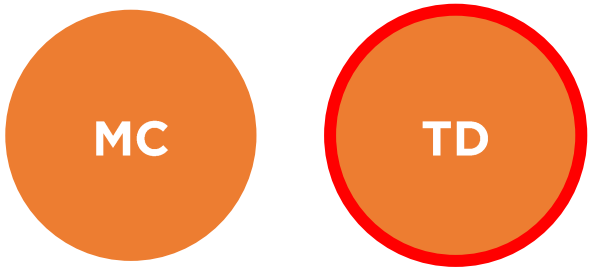
Off-policy TD control

FAST CAMPUS ONLINE

강화학습 A-Z I

강사. 박준영

지난 이야기 …



환경에 대해서 모를 때

On-policy learning

1. 현재 정책 π 로 세상에서, 데이터를 얻어 $Q^{\pi}(s,a)$를 추산.
2. 정책 개선기법 (e.g. $\epsilon-$ 탐욕적 정책개선) 으로 개선된 정책 π' 을 구함
3. (1), (2) 를 π' 가 수렴할 때까지 반복.

e.g.) Monte-Carlo Control, SARSA Control

- 이전 정책을 평가하기 위해 사용한 데이터를 재활용할 수 있을까?
- 다른 정책으로 얻어진 데이터를 학습에 사용할 수 있을까?

Off-policy learning

1. 임의의 행동 정책 μ 로 데이터를 얻음
2. Importance sampling 을 활용해서 타겟 정책 $Q^{\pi}(s,a)$을 추산
3. 정책 개선 수행

e.g) off-policy MC control

TD(0) (복습)

$$V(s) \leftarrow V(s) + \alpha\big(G_t - V(s)\big)$$

가장 간단한 TD 학습 알고리즘, **TD(0)** 의 경우: $G_t \overset{\text{def}}{=} R_{t+1} + \gamma V(S_{t+1})$

- $R_{t+1} + \gamma V(S_{t+1})$ 를 **TD target** 이라 부름. → $V(s)$ 를 **TD target** 에 가까워지게 조정
- $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ 를 **TD error** 라고 부름. → $V(s)$ 과 **TD target** 얼마나 차이 나는가?

Off-policy MC (복습)

- μ 임의의 행동 정책함수로 에피소드 생성
- $G_t^{\pi/\mu} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} G_t$ 계산
- MC update 수행

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(G_t^{\pi/\mu} - Q(s_t, a_t) \right)$$

- $\prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$ 때문에, $G_t^{\pi/\mu}$ 의 분산이 커질 수 있음.
- $\pi(A_k|S_k) \neq 0$ 이면, $\mu(A_k|S_k) \neq 0$ 이어야만 함.

Importance sampling for Off-Policy TD

- μ 임의의 행동 정책함수로 에피소드 생성
- TD target $R_{t+1} + \gamma V(S_{t+1})$ 계산
- Importance sampling corrected TD 업데이트 수행:

$$V(s_t) \leftarrow V(s_t) + \alpha \left(\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} \left(R_{t+1} + \gamma V(s_{t+1}) \right) - V(s_t) \right)$$

I No god please no, Importance sampling

Importance sampling 을 활용한 off-Policy control 들의 단점.

1. Sampling distribution (우리의 경우, 행동 정책 $\mu(a_t|s_t)$) 의 선택에 따라 추산치의 분산이 커질 수 있다.

- 예를 들어, 현재 평가하는 정책 π 과 상이하게 다른 $\mu(a_t|s_t)$를 사용하면 분산이 매우 커짐.
- 따라서, 아무 분포로 exploration 해도 되지만 분산 문제가 생길 가능성이 있음.

2. 구현이 귀찮음. 추가적으로 $\mu(a_t|s_t)$ 도 기록해야 함.

"Importance sampling 없는 off-policy 학습방법은 없을까?"

Q-Learning: An Off-Policy TD Control

Q-Learning 의 평가정책 π

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

Q-Learning (Watkins, 1989)

초기화 $Q(s,a) \leftarrow 0$ 모든 $(s,a) \in \mathcal{S} \times \mathcal{A}$

반복 (에피소드 0,1,2, …):

초기상태 s 관측

반복: (에피소드 내에서)

행동 a 를 $Q(s,a)$ 를 활용해서 결정 (e.g. ϵ – 탐욕적 정책) **행동정책 μ**

행동 a 를 환경에 가한 후, r, s' 를 관측.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$$

평가 정책 π

$s \leftarrow s'$

까지 s이 종결 상태

어...? Importance sampling 어떻게 없앴음?

Bellman 최적 방정식, 가치반복 알고리즘.

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

가치 반복 (Value iteration)

꼭 0 이 아니라 어떤 값으로 시작해도 하나의 값으로 알고리즘의 결과는 수렴함.

입력: 임의의 가치 함수 $V_0(s) \leftarrow 0$ 모든 $s \in \mathcal{S}$

출력: 최적 가치 함수 $V^*(s)$

반복: $(k = 0, \dots,)$:

- 모든 상태 s 에 대해서, $V_{k+1}(s) = \max_{a \in \mathcal{A}}(R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V_k(s'))$ 적용 <Bellman optimal backup>
- 모든 상태 s 에 대해서, $V_{k+1}(s) \sim V_k(s)$ 이면, 반복문 탈출

반환: $V^*(s) \leftarrow V_{k+1}$

가치반복 (Value iteration) 은 Bellman 최적 방정식의 해를 구하는 기법!

- 어떻게? 반복적으로 Bellman optimal backup 을 진행.

Bellman 최적 방정식의 샘플기반 추산 == Q-Learning objective

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

$$= R_s^a + \gamma \mathbb{E}_{s' \sim P_{ss'}^a} \left[\max_{a' \in \mathcal{A}} Q^*(s',a') \right]$$

$R_s^a, P_{ss'}^a$ 은 환경에 대한 정보!
RL 에이전트에게는 주어져 있지 않음.

샘플기반 추산 (sample이 1개)
+
Incremental update

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right)$$

추산치가 행동정책 μ 및 평가정책 π의 영향을 받지 않는다!

- Importance sampling 을 할 이유 자체가 사라짐.

Q-learning ain’t silver bullet

- Q-learning 은 Importance sampling 없이 off-policy learning 이 가능
- 언뜻 보기에는 On-policy TD Control 인 SARSA에 비해서 장점만 있어 보이지만
  사실, Maximization bias 라는 단점이 존재
- **Maximization bias:** Q-Learner 가 가치함수 $Q(s,a)$ 를 실제값보다 높게 평가하는 문제.
  Part 5 <심층 강화학습> 좀 더 자세히 이야기해봅시다!