동적계획법과 강화학습 강의 노트

작성자 : 공지환(2021321342)

작성일시 : 2022.10.19

1. Monte Carlo learning

   - Monte Carlo learning은 Episode를 전부 진행하고나서 최종적으로 얻은 gain값을 기준으로 가치 함수 값을 update한다.
   - gain값이 추출되어야 update가 가능하므로 episode의 끝이 있어야 한다.
   - $V(S) \leftarrow V(S) + \alpha(G_t - V(S'))$의 수식을 기준으로 가치함수 값을 update한다.

2. TD learning

   - TD learning은 Episode 진행 도중 step마다 가치함수 값을 update한다. 따라서 Episode가 끝나지 않은 상태에서도 update가 가능하다.
   - $V(S) \leftarrow V(S) + \alpha(R + \gamma V(S') - V(S))$의 수식을 기준으로 가치함수 값을 update한다.
   - TD leaning 방식을 사용하는 알고리즘에는 SARSA와 Q-learning이 있다.

3. SARSA

   - SARSA는 하나의 에피소드에서 현재 state, action, reward, 다음 순서의 state, action(S, A, R, S', A')까지 고려하여 q value를 update 하는 것을 의미한다.
   - SARSA에서 value값 update시, episode의 현재 상태의 q value를 update할 때, 다음 상태 action 까지 고려하는데 이 때, action은 episode에서 직접 실행하는 action 기준으로 생각하기 때문에 SARSA는 on-policy learning이다.
   - $Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A))$ 와 같은 수식을 기준으로 q value를 update한다.

4. Q-learning

   - Q-learning은 SARSA와 마찬가지로 (S, A, R, S', A')를 고려하여 Episode가 진행되는 동안 각 step에서 q value를 update 하는데 다음 시점의 action인 A'가 해당 가치함수를 최대로 만들도록 하는 action이 되도록 학습을 시킨다. 실제 episode에서 진행된 A'이 아닌 가치 함수 값을 최대로 만드는 A'을 사용하기 때문에 off-policy learning이다.
   - $Q(S,A) \leftarrow Q(S,A) + \alpha\left(R + \gamma \max_{A'} Q(S',A') - Q(S,A)\right) = Q(S,A) + \alpha\left(R + \gamma Q(S', \operatorname{argmax}_{A'} Q(S',A')) - Q(S,A)\right)$ 을 기준으로 q value를 update한다.

5. ε-greedy exploration

   - 모든 action이 non-zero 확률로 action 수행을 한다.
   - $\pi(a|s) = \begin{cases} \varepsilon/m + 1 - \varepsilon & if\ a^* = \underset{a}{\operatorname{argmax}}\, Q(s,a) \\ \varepsilon/m & otherwise \end{cases}$ 의 확률로 정책을 설정하여 다른 m-1개의 action이 간헐적으로 동작하게 한다.

6. ε-greedy exploration policy iteration

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_{a \in \mathcal{A}} \pi'(a|s) q_\pi(s,a) \\
&= \epsilon/m \sum_{a \in \mathcal{A}} q_\pi(s,a) + (1-\epsilon) \max_{a \in \mathcal{A}} q_\pi(s,a) \\
&\geq \epsilon/m \sum_{a \in \mathcal{A}} q_\pi(s,a) + (1-\epsilon) \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \epsilon/m}{1-\epsilon} q_\pi(s,a) \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a) = v_\pi(s)
\end{aligned}
$$

Therefore, from policy improvement theorem $v_{\pi'}(s) \geq v_\pi(s)$

7. Monte Carlo policy iteration

   - Policy iteration을 진행할 때, evaluation 과정에서 Q값이 완전히 수렴할 때까지 evaluation 진행한다.

8. Monte Carlo Control

   - Episode 종료시점에 바로 improvement 진행, Q가 수렴하지 않아도 improvement를 진행한다.

9. GLIE(Greedy in the Limit with Infinite Exploration)

- Sample kth episode using π: $\{S_1, A_1, R_2, ..., S_T\} \sim \pi$
- For each state S_t and action A_t in the episode,

$$
\begin{aligned}
N(S_t, A_t) &\leftarrow N(S_t, A_t) + 1 \\
Q(S_t, A_t) &\leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)} (G_t - Q(S_t, A_t))
\end{aligned}
$$

- Improve policy based on new action-value function

$$
\begin{aligned}
\epsilon &\leftarrow 1/k \\
\pi &\leftarrow \epsilon\text{-greedy}(Q)
\end{aligned}
$$

   - GLIE Monte-Carlo control 최적 행동가치 함수 $q^*(s,a)$로 수렴한다.

10. n-step SARSA

- Consider the following n-step returns for $n = 1, 2, \infty$:

$$
\begin{array}{lll}
n = 1 & (Sarsa) & q_t^{(1)} = R_{t+1} + \gamma Q(S_{t+1}) \\
n = 2 & & q_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2}) \\
\vdots & & \vdots \\
n = \infty & (MC) & q_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-1} R_T
\end{array}
$$

- Define the n-step Q-return

$$q_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n Q(S_{t+n})$$

- n-step Sarsa updates $Q(s, a)$ towards the n-step Q-return

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(q_t^{(n)} - Q(S_t, A_t) \right)$$

  - 바로 다음 시점보다 더 과거의 가치 함수 값을 고려하여 현재 q value를 산정할 필요가 있을 시 n-step SARSA를 사용한다.