동적계획법과 강화학습 강의 노트

작성자 : 김현민(2021314209)

Lecture 4. MC and TD / Simulation

1. Model Free Prediction

   모르는 MDP에 대한 value function을 추정하기 위한 것으로, Monte-Carlo Learning(MC)와 Temporal Difference Learning(TD)가 있음.

2. Model based vs. Model Free

   두 방식의 차지는 model의 존재 여부로 나누어지며, model을 사용하는 agent를 model-based라고 부르고 그렇지 않은 agent를 model-free라고 부름. model을 갖는 것은 장점과 단점이 있습니다.

   1-1. Model을 갖는 것의 장점은 Planning(계획)을 가능하게 한다는 것임. 즉, 자신의 action에 따라서 environment가 어떻게 바뀔지 안다면 실제로 행동하기 전에 미리 변화를 예상해보고 최적의 행동을 계획하여 실행할 수 있음.

   1-2. Model을 갖는 것의 단점은 environment의 정확한 model은 보통 알아내기가 어렵거나 불가능하다는 점임. 혹시라도 Model이 environment를 제대로 반영하지 않는다면, 오류는 그대로 agent의 오류로 이어지게 되며, 정확한 model을 만드는 것은 좋은 agent를 만드는 것만큼 또는 더 어려울 수 있음.

3. Monte Carlo Learning

   MC는 샘플링 된 에피소드로부터 학습하는 방법임. MC는 에피소드가 종료가 되야 value를 업데이트 할 수 있다는 특징을 가지며, 에피소드가 종료된 후에 받게 되는 보상(Gain)에 평균값을 value로 사용함.

   에피소드가 완전히 끝나야 업데이트가 가능하다는 단점과 큰 분산을 가진다는 단점과 반대로 unbias하다는 장점을 동시에 가짐.

   1-1. first visit MC

   에피소드에서 하나의 state를 여러번 지나갈 수 있을 것이며, 이때 해당 state에 첫번째 방문했을 때의 value만을 사용하는 방식임. 에피소드가 여러번 진행될 때 각 에피소드에 대한 평균으로 value를 추정함.

1-2. every visit MC

하나의 states를 두 번 이상 지나갔다면 이때의 모든 value를 평균 내어 추정하는 방식이며, 나머지는 first visit MC 방식과 비슷하게 진행됨.

4. Temporal Difference Learning

TD의 경우 MC와 달리 bootstrapping을 통해 time-step별로 이전의 value들을 통해 현재 value를 추정함.

TD의 경우 다음 step만 고려하여 업데이트를 진행하기 때문에 bias한 단점을 가지고 있음.

1-1. N-step TD

n-step TD는 update를 할 때, 하나의 step이 아니라 n개의 step을 보고 update를 하자는 것d이며. 만약 n을 Terminal state까지 가져간다면 MC가 됨. 이렇게 n-step을 취하여 update를 하면 MC와 TD의 장점을 모두 가져갈 수 있음.

5. Simulation

가상으로 시스템을 복제한 것으로, 실제 실험의 risk를 줄이기 위해 가상으로 실험을 하기 위한 것.

1-1. System

관련된 내용들의 집합

1-2. State

특정 시점에서의 시스템의 상태

1-3. Event

시스템의 상태를 바꾸는 시뮬레이션 기본 단위

1-4. Event scheduler

시간 순서대로 이루어지는 event들을 순차적으로 처리해주는 장치

[참고문헌]

- https://dreamgonfly.github.io/blog/rl-taxonomy/