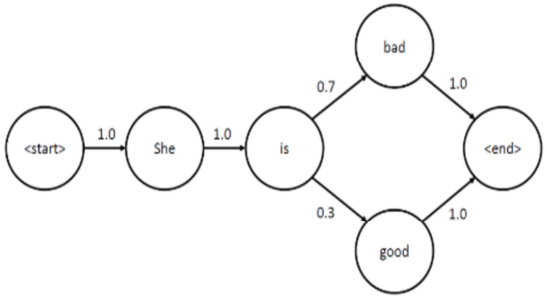
Decision Making

정민우

09/22/2022

메모리를 갖지 않는 이산 시간 확률 과정
확률과정
시간의 흐름에 따라 상태가 확률적으로 변화하는 과정
확률 분포를 따르는 임의변수(random variable)가 discrete한 time interval 마다 값을 생성하는 것
현재 상태(state)가 이전 상태에만 영향을 받는 확률 과정
미래는 오로지 현재에 의해 결정됨
상태가 변화하는 과정은 확률 계산에 영향을 미치지 않음
단일 상태 정보만으로 정보가 충분하도록 상태를 잘 구성해야 함



Markov Property

어떤 시간에 특정 상태에 도달하든 그 이전에 어떤 상태를 거쳐왔든 다음 상태로 갈 확률은 항상 같음
미래는 과거와 독립적인 것
Memoryless property

$$Pr(S_{t+1} = s' | S_0, S_1, ..., S_{t-1}, S_t) = Pr(S_{t+1} = s' | S_t) \quad (1)$$

Terminal state
Stationary distribution

State Transition Probability Matrix

Transition : state 간 이동
State transition probability : 확률적 transition

$$P_{ss'} = Pr(S_{t+1} = s' | S_t = s) \tag{2}$$

Transition probability를 행렬 형태로 정리한 것

Reward

Markov process에 reward 개념을 추가하는 것
Reward는 transition에 대한 가중치를 추가하는 것

$$R_s = E[r_{t+1}|S_t = s] \tag{3}$$

Discounting factor

감가율
현재와 미래의 가치 차이가 발생함

Return

미래에 얻을 수 있는 total reward 예측 가능함
각 시점에서 immediate reward들을 현재가치로 환산하여 합한 값

$$G_t = R_{t+1} + R_{t+2} + ... = \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1}) \tag{4}$$

Value Function of MRP

state의 가치를 표현하는 함수
특정 state에서 미래에 얻을 수 있는 모든 reward를 더한 것에 대한 expectation
state에서 이동 가능한 state들의 시나리오들을 따라 reward에 discounting factor를 적용하여 합한 값

$$V(s) = E[G_t|S_t = s] \tag{5}$$

어떤 문제를 컴퓨터로 풀기 위해서 수학적으로 정의되어야 함
Markov decision process는 의사결정 과정을 모델링하는 틀을 제공함

Action

MRP에 agent 개념이 추가됨
Agent는 각 상황마다 Action

$$MDP \equiv (S, A, P, R, \gamma) \tag{6}$$

S : 상태집합
A : 액션집합
P : 전이확률행
R : 보상함수
gamma : 감쇠인자

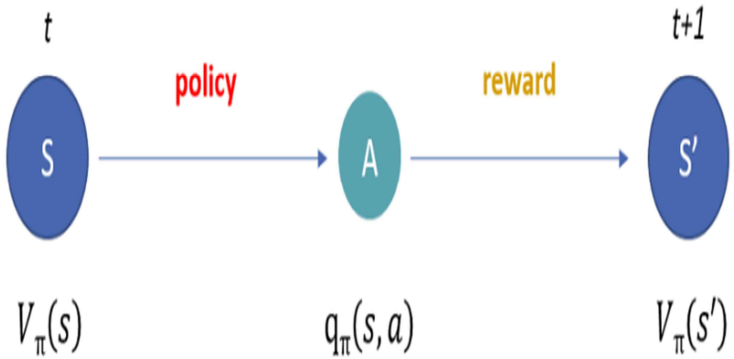
Policy

state에서 action을 mapping하는 함수
해당 state에서 어떤 action을 할 지 정하는 것

$$\phi(a|s) = Pr(A_t = a|S_t = s) \tag{7}$$

강화학습의 목적은 return을 최대화 할 수 있는 policy를 찾는 것
Return을 최대화하는 action을 선택하는 함수를 찾는 것

Value Function of MDP



t 시점에 state s에 놓인 agent가 policy에 따라 action a를 수행함
state s에서 action a를 수행하면 reward를 받음
transition probability에 따라 state s'으로 전이함
state-value 함수의 경우 state s에서 시작해서 policy를 따라서 나오는 return 값들의 기대값

$$V_\phi(s) = E_\phi[G_t|S_t = s] \tag{8}$$

action value function은 state s에서 action a를 실행하고 policy에 따라서

Bellman Expectation Equation

Markov process에 reward 개념을 추가하는 것
Reward는 transition에 대한 가중치를 추가하는 것

$$R_s = E[r_{t+1}|S_t = s] \tag{10}$$

Deep Q Learning Netwrok

강화학습의 일종
딥러닝의 인식 능력과 자체 의사결정 능력을 결합하여 복잡한 상태에서 인지적 의사결정 문제에 대한 솔루션
강화학습은 예측, 지능형 제어, 의사결정을 지원함
MDP 기반 알고리즘