발표자료

CONTENTS

1

마르코프 결정 과정 – 문제의 정의

마르코프 결정 과정 (MDP) – 문제의 정의

마르코프 가정 – 상태가 연속적인 시간에 따라 이어질 때, 어떤 시점의 상태는 바로 직전의 상태에만 영향을 받는다.

$$P(S_t \mid S_1, S_2, \cdots, S_{t-1}) = P(S_t \mid S_{t-1})$$

마르코프 과정 – 마르코프 가정을 만족하는 연속적인 일련의 상태

마르코프 과정은 일련의 상태<S1,S2,…,St>와 상태 전이 확률 P로 정의된다.

상태 전이 확률 - Pss’ = P[St+1=s’|St=s] : 현 상태가 다음 상태로 바뀌는 확률

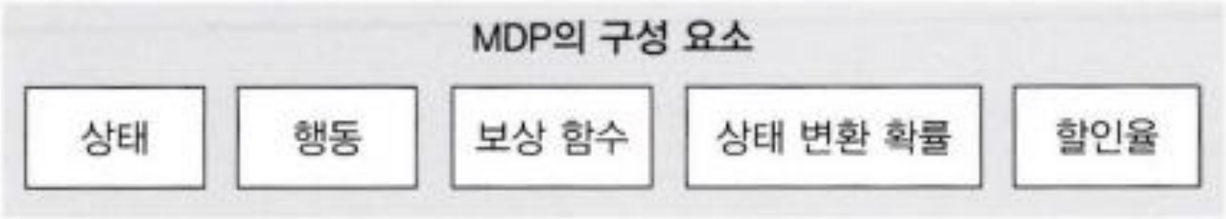
마르코프 결정 과정 (MDP) – 문제의 정의

MDP – 마르코프 과정을 기반으로 한 순차적 행동 결정 문제의 수학적 정의

구성 요소 – 상태(State), 행동 (Action), 보상 함수(Reward Function)
상태 변환 함수(Pss), 할인율 (r)

표현 – tuple(S,A,R,P,r)

강화학습에서는 사용자가 문제를 정의해야 하며, 문제의 정의는 에이전트가 학습하는데 가장 중요한 단계 중 하나이다. 학습하기에 많지도 않고, 적지도 않은 적절한 정보를 에이전트가 알 수 있도록 정의해야 한다.



마르코프 결정 과정 (MDP) – 문제의 정의

상태(State) – 에이전트가 관찰 가능한 모든 상태(State)의 집합 S
사용자가 정의하며, 학습에 충분한 정보를 주는지 검토 필요

MDP에서 상태는 시간 t에 따라 확률적으로 변한다. (확률 변수의 개념)

$S_t = s$: 시간 t에서의 특정 상태 s

행동(Action) – 에이전트가 할 수 있는 모든 행동의 집합 A
보통 행동들의 집합은 한 문제 내에서 변하지 않는다.

$A_t = a$: 시간 t에서의 특정한 행동 a

마르코프 결정 과정 (MDP) – 문제의 정의

보상함수 – 에이전트가 “학습할 수 있는 유일한 정보“
환경이 에이전트에게 주는 정보

$$r(s, a) = E[R_{t+1} \mid S_t = s, A_t = a]$$: 보상함수의 정의

시간 t 일 때, 상태와 행동에 대한 보상의 기댓값 (다변수 함수)
시간 t일 때의 데이터를 반영하나, 보상은 t+1의 시점에 주어진다.

기댓값의 의미 – 환경에 따라 같은 상태에서 같은 행동을 취해도
보상을 달리 줄 수 있다.

참고 – 환경에서 하나의 시간 단위를 타임 스탭(time step)이라고 한다.

마르코프 결정 과정 (MDP) – 문제의 정의

상태 변환 함수 – 상태의 “변화 " 를 확률의 개념을 이용하여 정의
환경의 모델이라고 부르며, 에이전트가 행동 a를 취하면
상태 변환 확률을 통해 “환경 " 이 에이전트의 다음 상태를 결정

$$P_{ss'}^{a} = P[S_{t+1} = s' | S_t = s, A_t = a]$$: 상태 변환 함수의 정의

마르코프 결정 과정 (MDP) – 문제의 정의

같은 양의 보상이라도 에이전트가 수령하는 시점에 따라 그 가치가 다르다.
에이전트는 보상이 얼마나 시간을 지나서 받는지를 고려해 현재의 가치로 환산한다.

이자는 나중에 받을 보상에 추가적인 보상을 더하여 현재의 보상과 같게 조정한다.
이는 같은 보상이면, 나중에 받을수록 가치가 감소함을 의미한다.

이를 반영하기 위한 수학적 개념이 할인율(r)discount factor이다

미래의 가치를 현재의 가치로 환산하는 것을 할인한다고 하며, 할인율은 “시간에 따라” 할인하는 비율이다.

$$\gamma \in [0,1]$$

수식 2.13 할인율의 정의

$$\gamma^{k-1} R_{t+k}$$

수식 2.14 할인율을 고려한 미래 보상의 현재 가치

마르코프 결정 과정 (MDP) – 문제의 정의

정책 – 상태 St=s를 입력으로 받고 행동 At=a를 출력으로 대응시키는 함수
정책은 일반적으로 각 상태에서 확률로 표현된다.

$$\pi(a \mid s) = P[A_t = a \mid S_t = s]$$

수식 2.15 정책의 정의

강화학습의 목적은 최적 정책을 결정하는 것이고, 최적 정책은 각 상태에서 “단 하나만의 행동 " 을 선택한다.

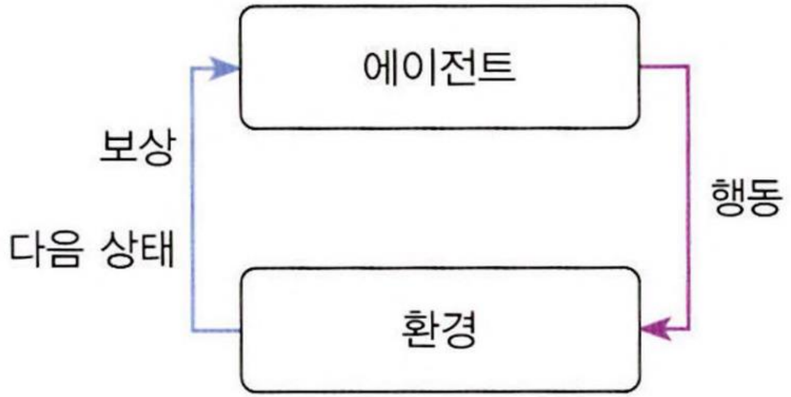
다만, 학습 단계에서는 확률적으로 여러 개의 행동을 선택한다.

2

마르코프 결정 과정 – 최적 정책 결정

마르코프 결정 과정 (MDP) – 최적 정책 결정

에이전트는 현재 상태에서 앞으로 받을 보상들을 고려(예측)해서 행동을 결정한다.



환경은 에이전트의 행동에 대해 “실제 보상 " 과 다음 상태를 에이전트에 알려준다.
에이전트는 이러한 피드백을 통해 예상한 보상에 대해 틀렸다는 것을 알게 된다.

이를 통해 에이전트는 실제로 받은 보상을 토대로 예측한 보상과 정책을 수정하며,
가장 많은 보상을 받게 하는 정책을 학습한다.

마르코프 결정 과정 (MDP) – 최적 정책 결정

가치함수 – 앞으로 받을 보상에 대한 개념

보상은 행동을 한 시점 t가 아닌 t+1의 시점에서 환경으로부터 받는다. (Rt+1)

$$R_{t+1} + R_{t+2} + R_{t+3} + R_{t+4} + R_{t+5} + \cdots$$

수식 2.16 일련의 보상들의 단순합

가치함수를 보상들의 단순합으로 표현하면 아래와 같은 문제가 있다.

1. 같은 보상도 받는 시점에 따라 가치가 다르나, 이에 대한 반영이 안된다.
2. 같은 양의 보상을 1번 받는 것과 나누어 받는 것을 구분할 수 없다.
3. 시간이 무한대인 경우, 수치적인 구분이 불가능하다.

Chap 2

마르코프 결정 과정 (MDP) – 최적 정책 결정

반환값(return) – 할인율을 도입한 보상합으로 실제로 환경을 탐험하며 받은 보상의 합
단순합의 3가지 문제를 해결했다.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

수식 2.19 반환값의 정의

에피소드의 정의 – 에이전트와 환경 간 유한한 시간 동안의 상호작용
반환값은 에피소드 종료 이후, 에이전트가 계산한다.

MDP로 정의되는 세계는 에이전트와 환경의 상호작용은 확률적 개념을 통해 기술된다. 그러므로, 불확실성을 내포하고 있고 특정 상태의 반환값은 에피소드 마다 다를 수 있다.

반환값에 대한 수학적 접근

보상들의 단순합 $= \sum_{k=1}^{\infty} R_{t+k}$

반환값 $G_t = \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$ where $\gamma \in [0,1]$

문제점 1. 같은 보상도 받는 시점에 따라 가치가 다르다.

Let $R_{t+k_1} = R_{t+k_2}$ $k_1, k_2 \in \mathbb{N}$ and $k_1, k_2 < k$ and $k_1 \neq k_2$

by discount Factor, $R_{t+k_1} \rightarrow \gamma^{k_1-1} R_{t+k_1}$, $R_{t+k_2} \rightarrow \gamma^{k_2-1} R_{t+k_1}$

$\gamma^{k_1-1} R_{t+k_1} = \gamma^{k_1-1} R_{t+k_2} \neq \gamma^{k_2-1} R_{t+k_2}$ $(\because k_1 \neq k_2)$

반환값에 대한 수학적 접근

보상들의 단순합 $= \sum_{k=1}^{\infty} R_{t+k}$

반환값 $G_t = \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$ where $\gamma \in [0,1]$

문제점 2. 같은 보상을 1번 받는 것과 나눠서 받는 것을 구별할 수 없음.

Let $K_t = \sum_{k=1}^{\infty} R'_{t+k}$, $G_t = \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$

In this case, $\sum_{k=1}^{\infty} R'_{t+k} = \sum_{k=1}^{\infty} R_{t+k}$: Same Reward

$\sum_{k=1}^{\infty} R_{t+k}' = K_t$, $\sum_{k=1}^{\infty} R_{t+k} > \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} = G_t$ $(\because \gamma \in [0,1])$

$\therefore K_t > G_t$: Not same value

반환값에 대한 수학적 접근

문제점 3.

$G_t = \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$ if $\gamma = 1$, It means that G_t is series of R_{t+k}.

So, Let $\gamma \neq 1$.

Let R = Maximum value of $[R_{t+1}, R_{t+2}, \cdots, R_{t+n}, R_{t+n+1}, \cdots]$

$R_{t+k} < R$ where $R_{t+k} \in [R_{t+1}, R_{t+2}, \cdots, R_{t+n}, R_{t+n+1}, \cdots]$

So $R_{t+k} \cdot \gamma^{k-1} < R \cdot \gamma^{k-1}$ $(\gamma \in [0, 1))$

$$G_t = \sum_{k=1}^{\infty} R_{t+k} \cdot \gamma^{k-1} < \sum_{k=1}^{\infty} R \cdot \gamma^{k-1} = R \cdot \frac{1}{1-\gamma} < \infty$$

$\therefore$ G_t Converges to a specific value

마르코프 결정 과정 (MDP) – 최적 정책 결정

가치함수 – 특정 상태에서 반환값에 대한 기댓값으로 각 타임 스탭마다 받는 보상은 확률적이므로, 반환값은 확률변수이다. 다만, 가치함수는 확률변수가 아니라 특정 값이다.

$$v(s) = E[G_t \mid S_t = s]$$

수식 2.21 가치함수

$$v(s) = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots \mid S_t = s]$$

수식 2.22 앞으로 받을 보상에 대한 기댓값인 가치함수

$$v(s) = E[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} \cdots) \mid S_t = s]$$

$$v(s) = E[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

수식 2.23 반환값으로 나타내는 가치함수

마르코프 결정 과정 (MDP) – 최적 정책 결정

가치함수에서 타임 스탭이 t+2 이상인 항들은 실제 받은 보상이 아닌 추정값이다.

$$v(s) = E[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s]$$

수식 2.24 가치함수로 표현하는 가치함수의 정의

지금까지 정의한 가치함수에서는 정책을 배제하였으나, 실제 가치함수에서는 정책이 고려되어야 한다. 각 상태에서 행동을 선택하는 것이 정책이기 때문이다.

보상은 어떤 상태에서 어떤 행동을 할 때, 환경에서 에이전트에게 주어진다. 따라서, MDP로 정의되는 문제에서 가치함수는 항상 정책에 종속된다.

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

수식 2.25 정책을 고려한 가치함수의 표현

수식 2.25는 벨만 기대 방정식으로, 현재 상태의 가치함수와 다음 상태의 가치함수의 관계를 서술한다. 강화학습은 벨만 방정식을 풀어가는 과정이다.

마르코프 결정 과정 (MDP) – 최적 정책 결정

상태 가치함수 – 상태가 입력되면, 보상의 합을 출력하는 함수
행동 가치함수 – 행동이 입력되면, 보상의 합을 출력하는 함수

행동 가치함수(Q-Function) – 상태와 행동이 입력되면, 보상의 합을 출력하는 함수

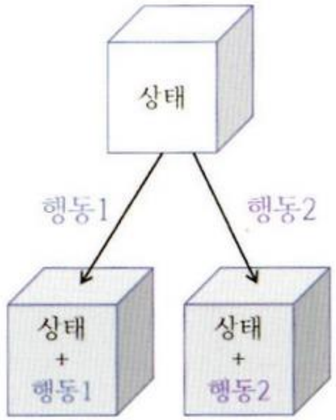


그림 2.16 큐함수의 다이어그램. 흰상자는 상태를 의미하고 회색상자는 행동 상태를 의미한다

마르코프 결정 과정 (MDP) – 최적 정책 결정

가치함수와 큐 함수 사이의 관계

1. 각 행동을 했을 때 앞으로 받을 보상인 큐함수 $q_\pi(s,a)$를 $\pi(a \mid s)$에 곱합니다.
2. 모든 행동에 대해 큐함수와 $\pi(a \mid s)$를 곱한 값을 더하면 가치함수가 됩니다.

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s)\, q_\pi(s,a)$$

수식 2.26 가치함수와 큐함수 사이의 관계식

에이전트가 행동을 선택하는 기준으로 보통 가치함수보다 큐함수를 사용한다.

수식 2.27은 큐함수의 벨만 기대 방정식으로, 가치함수와 달리 조건문에 행동이 들어간다.

$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a]$$

수식 2.27 큐함수의 정의

3

벨만 방정식

벨만 기대 방정식

수식 2.29를 통한 가치함수의 계산은 연산에 있어 비효율적이다.

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots \mid S_t = s]$$

수식 2.29 반환값으로 나타내는 가치함수

벨만 기대 방정식은 현 상태와 다음 상태 사이의 가치함수간 관계를 서술한다. 수식 2.28은 가치함수를 점화식 형태로 표현 가능함을 보여주며, Recursion을 이용하면, 연산의 효율성이 높아진다.

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

수식 2.28 벨만 기대 방정식

벨만 기대 방정식을 통한 가치함수의 연산에서는 정책과 상태 변환 확률을 포함해 계산해야 한다. (우변이 기댓값의 개념이기 때문이다.)

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s)\left(r_{(s,a)} + \gamma \sum_{s' \in S} P^a_{ss'} v_\pi(s')\right)$$

수식 2.32 계산 가능한 벨만 방정식

벨만 기대 방정식

벨만 기대 방정식을 통한 가치함수의 연산에서는 정책과 상태 변환 확률을 포함해 계산해야 한다. (우변이 기댓값의 개념이기 때문이다.)

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s)\left(r_{(s,a)} + \gamma \sum_{s' \in S} P^{a}_{ss'} v_\pi(s')\right)$$

수식 2.32 계산 가능한 벨만 방정식

벨만 기대 방정식 연산 예시)

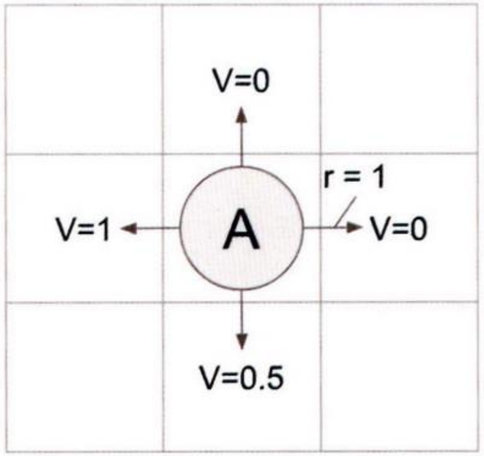


그림 2.17 그리드월드에서 가치함수의 업데이트

1	행동 = 상	$0.25\times(0+0.9\times0)=0$
2	행동 = 하	$0.25\times(0+0.9\times0.5)=0.1125$
3	행동 = 좌	$0.25\times(0+0.9\times1)=0.225$
4	행동 = 우	$0.25\times(1+0.9\times0)=0.25$
총합	기댓값 $=0+0.1125+0.225+0.25=0.5875$	

벨만 최적 방정식

벨만 기대 방정식에 대한 연산을 반복하면 특정 값으로 수렴하게 되는데, 이 값을 참 가치함수라 한다. 이 값은 “어떤 정책 " 을 따랐을 때 받게 되는 보상에 대한 값이다.

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

수식 2.34 벨만 기대 방정식

최적의 가치함수는 수많은 정책 중, 가장 높은 보상에 대한 가치함수이다.

$$v_*(s) = \max_\pi [v_\pi(s)] \quad q_*(s,a) = \max_\pi [q_\pi(s,a)]$$

수식 2.36 최적의 가치함수 수식 2.37 최적의 큐함수

가장 높은 가치함수(큐함수)를 찾았다고 가정하면, 최적 정책은 상태 s에서의 최적의 큐함수 중에서 가장 큰 큐함수를 가진 행동을 해야 한다. (s 고정, a 선택)

$$\pi_*(s,a) = \begin{cases} 1 & \text{if } a = \text{argmax}_{a \in A} q_*(s,a) \\ 0 & \text{otherwise} \end{cases}$$

수식 2.38 최적 정책

벨만 최적 방정식

최적 가치함수와 큐함수 구하기 = 순차적 행동 결정 문제의 해법

$$v_*(s) = \max_a [q_*(s,a) \mid S_t = s, A_t = a]$$

수식 2.39 큐함수 중 최대를 선택하는 최적 가치함수

$$v_*(s) = \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$$

수식 2.40 벨만 최적 방정식

최적의 큐함수는 식 2.39처럼 표현되고, 큐함수를 가치함수로 표현하면 식 2.400이 되며 이를 벨만 최적 방정식이라고 한다.

큐함수에 대한 벨만 최적 방정식은 식 2.41과 같다.

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a]$$

수식 2.41 큐함수에 대한 벨만 최적 방정식

회의 수고하셨습니다.