07

REINFORCE 알고리즘

1. 기본개념

RFINFOCE 기본개념

인공신경망 다시 보기

이 데이터를 표현하는 함수가 뭐지?
잘 모르겠는데, 데이터가 너무 복잡한 걸

→ 그럼 인공신경망을 사용해 볼까?

인공신경망은 함수야?

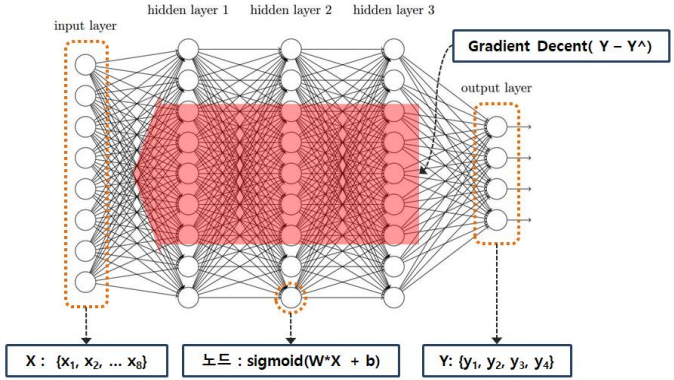
→ 응 인공신경망은 가중치(W: Weight)와 편향(B: Bias)으로 표현되는 함수야

그럼 가중치와 편향은 어떻게 알아내?

→ 데이터를 학습해서 알아내지

그럼 인공신경망으로 모든 데이터를 표현할 수 있겠네?

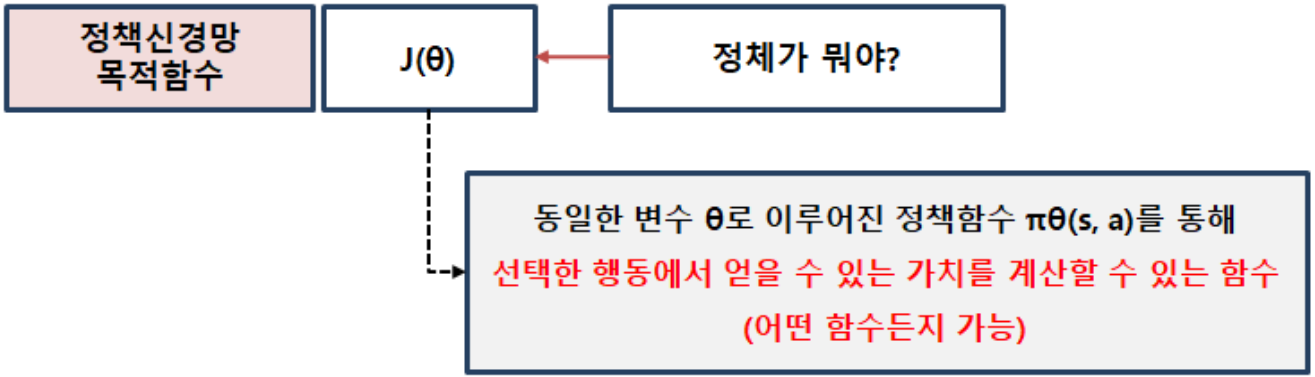
→ 맞아. 다른 말로 인공신경망은 모든 함수를 표현할 수 있어



RFINFOCE 기본개념 정책 그래디언트

정책 목적함수

- 정책 목적함수(Policy Object Function), J(θ) : θ 로 이루어진 정책을 평가하기 위한 함수



RFINFOCE 기본개념 정책 그래디언트

가치신경망과 정책신경망 평가함수

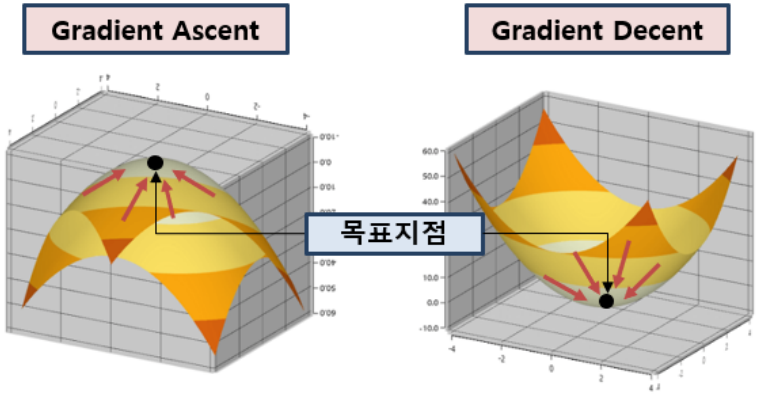
- J(w)는 평균제곱오차를 의미하므로 경사하강법
- J(θ)는 정책의 가치를 의미하기 때문에 경사상승법

가치신경망 평가함수	J(w)	MSE(참가치함수 – 가치신경망 값)	최소화
정책신경망 평가함수	J(θ)	정책이 가지는 가치	최대화

RFINFOCE 기본개념 정책 그래디언트

경사상승법

- 경사상승법은 경사하강법과 반대 방향, 서로 부호만 바꿔주면 동일하다.



$\Delta\theta = \propto \nabla_\theta J(\theta)$

$\Delta\theta = -\propto \nabla_\theta J(\theta)$

RFINFOCE 기본개념 정책 그래디언트

MDP에서 가치함수

- MDP에서의 가치함수를 응용해서 J(θ)를 정의할 수 있다.

MDP

$$v_\pi(s) = \mathbf{E}_\pi[G_t \mid S_t = s] \quad ①$$

$$= \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] \quad ②$$

$$= \sum_{a \in A} \pi(a|s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')\right) \quad ③$$

$$= \sum_{a \in A} \pi(a|s)\, R_s^a + \gamma \sum_{a \in A} \pi(a|s) \sum_{s' \in S} P_{ss'}^a v_\pi(s') \quad ④$$

RFINFOCE 기본개념 정책 그래디언트

정책 목적 함수

MDP $v_\pi(s) = \sum_{a \in A} \pi(a|s)\, R_s^a + \gamma \sum_{a \in A} \pi(a|s) \sum_{s' \in S} P_{ss'}^a v_\pi(s')$ ①

One Step MDP $v_\pi(s) = \sum_{a \in A} \pi(a|s)\, R_s^a$ ②

Policy Object Function $J(\theta) = \sum_{a \in A} \pi_\theta(a|s)\, R_s^a$ ③

- 원 스텝 MDP의 가치함수를 정책 목적 함수로 사용할 수 있다.

J(θ)는 아무 함수나 가능해

↓

하지만, 선택한 행동에서 얻을 수 있는 가치를 계산할 수 있는 함수여야 해

↓

그럼 One Step MDP에서 상태가치함수를 J(θ)로 사용해 볼까?

RFINFOCE 기본개념 정책 그래디언트

정책 그래디언트

Policy Gradient

$$\nabla_\theta J(\theta) = \sum_{a \in A} \nabla_\theta \pi_\theta(a|s)\ R_s^a \quad ①$$

$$= \underbrace{\sum_{a \in A} \pi_\theta(a|s) \nabla_\theta log \pi_\theta(a|s)}_{②\text{-}1}\ R_s^a \quad ②$$

$$= E_{\pi\theta}[\nabla_\theta log \pi_\theta(a|s)\ R_s^a] \quad ③$$

Policy Gradient with SGD

$$\fallingdotseq \nabla_\theta log \pi_\theta(a|s)\ r \quad ④$$

Likelihood Ratio

$$\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s)\ \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}$$

$$= \pi_\theta(a|s) \nabla_\theta log \pi_\theta(a|s)$$

- 정책 목적함수의 결과를 가장 크게 얻을 수 있는 θ 를 찾는 것이 목적이기 때문에 경사 상승법을 이용해서 θ 를 갱신하는 수식을 만들어야 한다

$$J(w) = E_\pi\ [(v_\pi(s) - \hat{v}(s, w))^2]$$

$$\Delta w = -\frac{1}{2} \propto \nabla_w J(w) \quad ②\text{-}1$$

$$= \propto E_\pi\ [(v_\pi(s) - \hat{v}(s, w))\ \nabla_w \hat{v}(s, w)]$$

$$\Delta w = \propto (R_{t+1} + \gamma \hat{q}(S_{t+1}, A_{t+1}, w) - \hat{q}(S_t, A_t, w))\ \nabla_w \hat{q}(S_t, At, w)$$

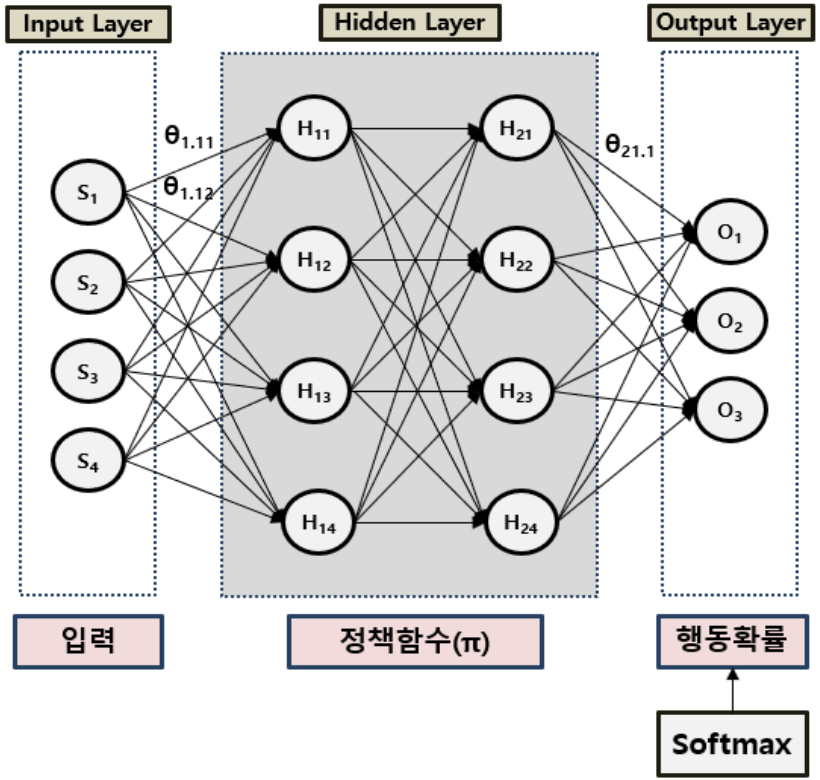
RFINFOCE 기본개념 정책 그래디언트

다양한 형태의 비용함수

One Step MDP	$\nabla_\theta J(\theta) = E_{\pi\theta}[\nabla_\theta log\pi_\theta(a\|s) R_s^a]$
Multi Step MDP	$\nabla_\theta J(\theta) = E_{\pi\theta}[\nabla_\theta log\pi_\theta(a\|s) Q^{\pi\theta}(s,a)]$
MC (REINFORCE)	$\nabla_\theta J(\theta) = E_{\pi\theta}[\nabla_\theta log\pi_\theta(a\|s) G_t]$

- REINFORCE는 가치를 계산하기 위해 G_t를 사용한다.

RFINFOCE 기본개념 정책 그래디언트



인공신경망을 활용한 정책 그래디언트

RFINFOCE 기본개념 알고리즘 동작방식