

Q1.

에이전트(agent), 환경(environment), 상태(state), 행동(action), 보상(reward)

### 1. 에이전트 (Agent)

에이전트는 의사 결정을 내리는 주체로, 환경에서 행동을 선택하고 그 결과를 학습하는 역할을 합니다.

### 2. 환경 (Environment)

환경은 에이전트가 상호 작용하는 세계로, 에이전트의 행동에 반응하여 상태를 변경하고 보상을 제공합니다.

### 3. 상태 (State)

상태는 환경의 현재 상황을 기술합니다. 에이전트는 각 상태에서 어떤 행동을 취할지 결정해야 합니다.

### 4. 행동 (Action)

행동은 에이전트가 선택할 수 있는 옵션입니다. 에이전트의 선택에 따라 환경은 다른 상태로 전이하고, 이에 따라 다른 결과가 발생합니다.

### 5. 보상 (Reward)

보상은 에이전트의 행동이 얼마나 좋았는지를 나타내는 신호입니다. 강화학습의 목표는 보상의 총합을 최대화하는 것입니다.



Anonymous

강화학습에 대해서 배우고 싶어 기초개념부터 자세하게 알려줘!

위와 같이 질문한 이유는 Q1에서 요구하는 것이 강화학습의 기본적이고 기초적인 개념들이라는 것을 인지하여 기본적인 개념을 알려달라고 하였고 Q1에서 요구한 사항을 다 알려준 ChatGpt를 볼 수 있습니다.

Q2.

이익(return), 가치 함수(value function)

**이익(return)** : 취한 행동 이후, 에이전트는 새로운 상태로 이동합니다. 이 새로운 상태에서 에이전트는 다시 행동을 취하고, 이러한 과정이 반복됩니다. 각 행동은 미래에 추가적인 보상을 가져올 수 있으며, 이러한 미래 보상들의 총합을 고려해야 합니다. **할인된 미래 보상** 먼 미래의 보상은 가까운 미래의 보상보다 덜 중요할 수 있기 때문에, 미래 보상에는 종종 할인율(discount factor)  $\gamma$ 가 적용됩니다. 할인율은 0과 1 사이의 값으로, 이 값에 따라 미래 보상의 현재 가치가 결정됩니다.

**가치 함수(value function)** :

가치 함수는 특정 상태(또는 상태와 행동의 조합)의 중요성을 수치적으로 평가하는 함수입니다. 즉, 어떤 상태(또는 상태-행동 쌍)가 미래에 얻을 수 있는 보상의 총합을 예측합니다. 가치 함수는 두 가지 주요 형태로 나타납니다.

**상태 가치 함수(State Value Function)** :

특정 상태의 가치를 평가합니다. 상태 가치 함수  $V(s)$ 는 정책  $\pi$ 를 따랐을 때, 상태  $s$ 에서 시작하여 얻을 수 있는 기대되는 보상의 합계를 나타냅니다.

$$V_{\pi}(s) = E[R_t \mid s_t = s, \pi]$$

여기서  $R_t$ 는 시간  $t$ 에서의 보상이고,  $E$ 는 기대값을 의미합니다.

**행동 가치 함수(Action Value Function)** :

특정 상태에서 특정 행동을 취했을 때의 가치를 평가합니다. 행동 가치 함수  $Q(s, a)$ 는 정책  $\pi$ 를 따랐을 때, 상태  $s$ 에서 행동  $a$ 를 선택하고 그 이후의 행동들을 취함으로써 얻을 수 있는 기대되는 보상의 합계를 나타냅니다.

$$Q_{\pi}(s, a) = E[R_t \mid s_t = s, a_t = a, \pi]$$



**Anonymous**

설명해줘서 고마워 함수에 대해서 말해줬는데 더 자세하게 알려줘

ChatGpt가 기본개념에서 함수에 대해서 알려주었기에 조금 더 자세하게 알려달라고 하였더니 Q2에서 요구하는 가치함수, 상태 가치함수, 행동 가치함수에 대해서 설명해주었다.

**Anonymous**

그 이후의 행동들을 취함으로써 얻을 수 있는 기대되는 보상의 합계를 나타냅니다.

이게 무슨 의미인가요?

질문을 하였을 때, Return에 대한 설명이 부족한 것 같아서 가치함수에서 있는 설명에 대해서 더 자세하게 알려달라고 프롬프트하였습니다.

Q3..

**가치 기반 알고리즘(value-based algorithm) :**

가치 기반 알고리즘은 가치 함수를 최대화하는 행동을 선택합니다. 이 방법은 주로 상태 또는 상태-행동 쌍의 가치를 추정하고, 이 추정값을 기반으로 최적의 정책을 도출합니다.

- **Q-러닝(Q-Learning):** 가장 널리 알려진 가치 기반 강화학습 알고리즘 중 하나로, 상태-행동 쌍의 가치(행동 가치 함수  $Q$ )를 학습합니다. Q-러닝은 오프-폴리시(off-policy) 학습 방법으로, 현재 정책과 다른 정책을 탐색하면서 학습할 수 있습니다.
- **SARSA(State-Action-Reward-State-Action):** Q-러닝과 유사하지만, SARSA는 온-폴리시(on-policy) 학습 방법입니다. 즉, 현재 정책을 기반으로 다음 행동을 선택하고, 이를 통해 학습합니다.

**정책 기반 알고리즘(policy-based algorithm) :**

정책 기반 알고리즘은 가치 함수를 명시적으로 학습하지 않고, 직접적으로 정책(행동을 선택할 확률)을 최적화합니다. 이 방법은 특히 높은 차원의 행동 공간이나 연속적인 행동을 다룰 때 유용합니다.

- **REINFORCE:** 이 알고리즘은 정책의 파라미터를 조정함으로써 정책 자체를 직접적으로 최적화합니다. 에피소드가 끝난 후에 얻은 보상을 기반으로 정책을 업데이트합니다.
- **TRPO(Trust Region Policy Optimization)와 PPO(Proximal Policy Optimization):** 이 두 알고리즘은 정책의 업데이트가 너무 커져서 성능이 저하되는 것을 방지하기 위해, 안정적인 학습을 위한 특정 조치를 취합니다. 특히 PPO는 실용성과 성능으로 인해 널리 사용됩니다.



**Anonymous**

**고마워 그러면 강화학습의 알고리즘은 어떤 것들이 있을까?**

Q3에서 요구하는 것이 어떠한 기반에 대한 알고리즘이기 때문에 알고리즘의 종류에 대해서 질문하였습니다. 이 질문을 통해서 가치 기반 알고리즘, 정책 기반 알고리즘에 대한 대답이 나왔습니다.

#### Q4. 정책 경사 알고리즘(policy gradient algorithm)

##### 4-1) 정책 기반 알고리즘

경사 상승법(gradient ascent)을 사용하여 정책의 파라미터를 조정합니다. 여기서 목표는 보상의 기대값을 최대화하는 방향으로 정책 파라미터를 업데이트하는 것입니다. 각 업데이트는 정책이 더 높은 보상을 얻을 확률을 증가시키는 방향으로 이루어집니다.

##### 4-2) 정책 경사 계산과정

정책 기반 강화학습 알고리즘에서 경사 상승법을 사용하는 과정은 주로 정책 파라미터를 조정하여 보상의 기대값을 최대화하는 것을 목표로 합니다. 이 과정은 수학적으로 다음과 같은 단계를 포함합니다:

###### 1. 정책 함수의 파라미터화

정책  $\pi$ 는 상태  $s$ 에 대해 어떤 행동  $a$ 를 취할 확률을 정의합니다. 이 정책은 파라미터  $\theta$ 에 의해 파라미터화됩니다 (예: 신경망의 가중치). 따라서,  $(\cdot;)\pi(a|s;\theta)$ 는 상태  $s$ 에서 행동  $a$ 를 선택할 확률을 나타내는 함수입니다.

###### 2. 목적 함수(J)의 정의

목적 함수  $J(\theta)$ 는 파라미터화된 정책  $\pi$ 에 따른 보상의 기대값을 나타냅니다. 강화학습의 목표는 이 목적 함수를 최대화하는 것입니다.  $J(\theta)$ 는 보통 다음과 같이 표현됩니다:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)]$$

여기서  $R(\tau)$ 는 경로  $\tau$ (상태-행동 시퀀스)에 따른 총 보상입니다. 경로  $\tau$ 는 정책  $\pi$ 에 의해 생성됩니다.

###### 3. 목적 함수의 그래디언트 계산

목적 함수  $J(\theta)$ 를 최대화하기 위해 파라미터  $\theta$ 에 대한 그래디언트를 계산합니다. 그래디언트는 다음과 같은 형태로 주어집니다:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[ \sum_{t=0}^{\infty} \gamma^t \nabla \log \pi(a_t | s_t; \theta) R_t \right]$$

여기서  $R_t$ 는 시간  $t$ 에서 받은 보상과 그 이후 받을 보상의 할인된 합계입니다.

###### 4. 경사 상승법에 의한 파라미터 업데이트

계산된 그래디언트를 사용하여 파라미터  $\theta$ 를 업데이트합니다. 이는 일반적으로 다음과 같은 경사 상승 스텝을 통해 수행됩니다:

$$\theta \leftarrow \theta + \alpha \nabla J(\theta)$$

여기서  $\alpha$ 는 학습률입니다.

연속적인 행동 공간을 가진 문제에서 강화학습 모델이 어떻게 행동의 확률을 조정하며 최적의 전략을 찾아가나요?

Q4에서 요구하는 정책 기반 알고리즘, 정책 경사 계산과정을 유도하기 위해서

정책 기반 알고리즘을 사용할 만한 상황 즉, 연속적인 행동 공간을 가진 문제에 대한 질문을 하여서 답을 유도하였습니다. 이 질문을 통하여 정책 기반 알고리즘에 대한 답이 나왔고



Anonymous

알고리즘들은 행동의 확률을 조정하며 최적의 전략을 찾아가는 방법에 있어 몇 가지 핵심적인 접근 방식을 취합니다. 이때 최적의 전략을 찾을 때 어떤 수학적 방법을 사용하?

정책 경사 계산과정을 알아내기 위해 다음 질문에서 위에서 보이듯이 수리적인 방법을 물어봐서 계산과정에 대한 답을 유도하였습니다.

위 프롬프트들로 강화학습에 대한 공부를 ChatGpt를 통하여 Q1,2,3,4에서 요구하는 강화학습에 개념에 대해서 공부를 마칠 수 있었습니다.