STUDY PRESENTATION

SUSC Summer 2023

강화학습 기본 이론

동아대 AI학과 김지선

Introduce MySelf

01 동아대학교 AI학과 3학년 김지선

02 다수의 스터디 주최 및 진행
네트워크/웹해킹, 머신러닝

03 빅데이터 분석 대상(산자부 장관상) 수상
제 10회 BI 아이디어 공모전에 팀장으로 참여

04 강화학습, 로봇 AI에 관심 있음

05 깃허브 - https://github.com/Prcnsi
블로그 - https://perconsi.tistory.com/

스터디 주교재

- 아래 책 두 권을 기반으로 정리하여 PPT로 직접 제작해 강의함.

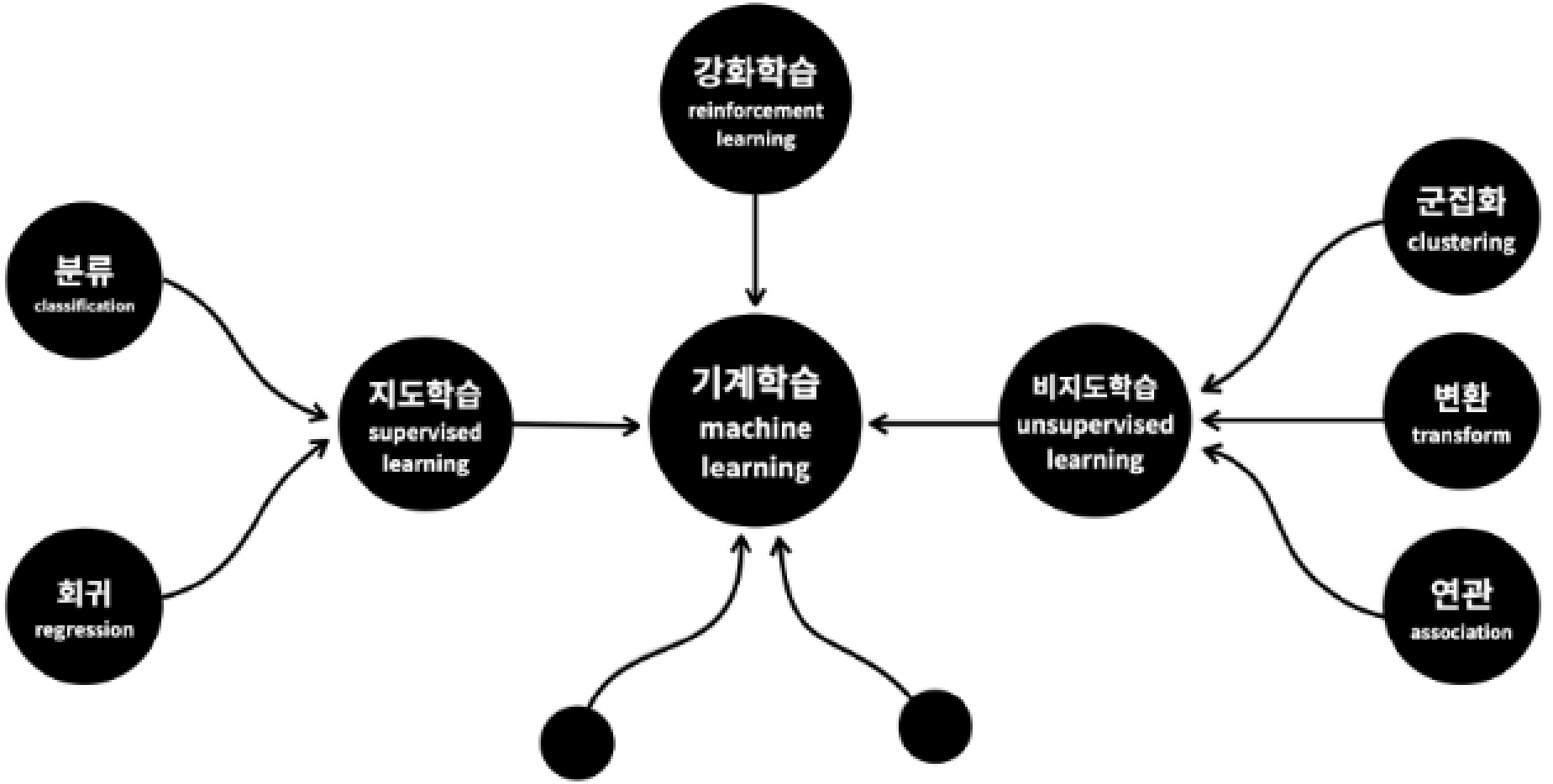
강화(Reinforce)의 의미

- 행동심리학에서 **“강화”의 개념은 동물이 “시행착오“를 통해 학습하는 방법** 중 하나이다.
- 즉, 강화라는 것은 동물이 이전에 배우지 않았지만, **직접 시도하면서 행동과 그 결과로 나타나는 좋은 보상 사이의 상관관계를 학습**하는 것이다. 좋은 보상을 얻게 해주는 행동을 점점 더 많이 하는 것을 말한다.
- 강화라는 개념이 강화학습의 모티브가 되어 강화학습은 컴퓨터가 스스로 학습하여 주어진 데이터를 토대로 스스로 성능을 높여가는 것.

What is Machine Learning?

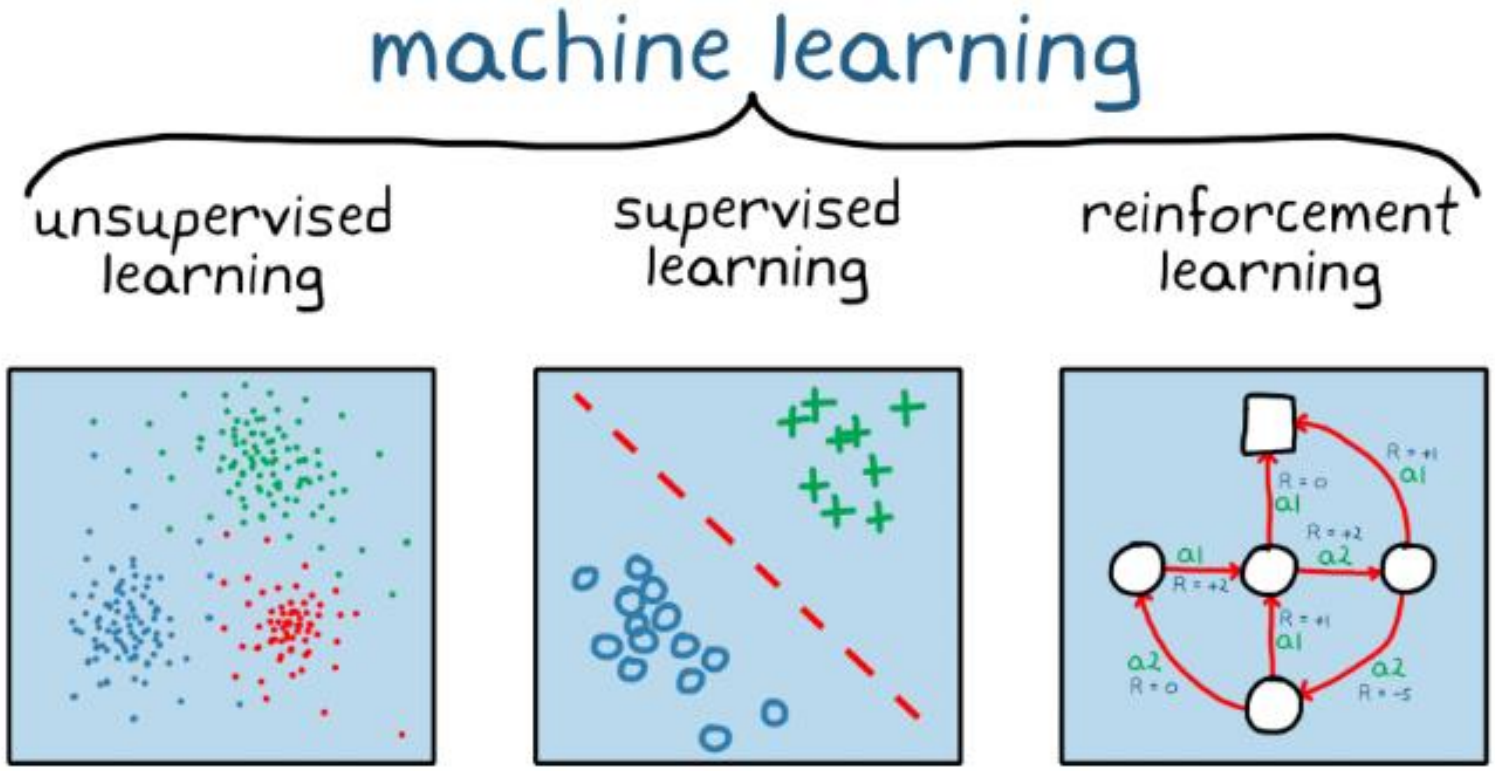
- 강화학습도 기계학습의 일종
- 머신러닝
  - 손실함수(비용 함수, 목적 함수)를 최소화하는 방향으로 모델을 학습
  - 지도학습(정답을 알고 있는 데이터를 활용해 컴퓨터를 학습)
  - 비지도학습(정답을 모르는 상태에서 주어진 데이터에 대해 학습)
  - 강화학습 (정답이 주어진 것은 아니지지만 행동에 대한 보상을 통해 학습함.
    (보상은 컴퓨터가 선택한 행동에 대한 환경의 반응)

What is Machine Learning?



https://opentutorials.org/module/4916/28934

What is Reinforcement Learning



https://kr.mathworks.com/discovery/reinforcement-learning.html

Reinforcement Learning (RL)

- 기계학습(머신러닝)의 한 영역으로, 행동 심리학에서 영감을 받음
- 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식
- 최종 누적 보상 합을 최대로 하는 것을 목표로함.
- 선택 가능한 행동들 중 보상을 최대화하는 순차적인 행동을 결정해야하는 문제.
- ChatGPT와 같은 LLM에서도 인간 피드백형 강화학습(RLHF)을 핵심으로 사용

강화학습 응용 분야

- 로봇 강화학습
- 주식투자에 강화학습 적용 (퀀트 투자)
- 멀티 에이전트 강화학습
- 자율주행
- ChatGPT

강화학습의 기본 용어

- Agent: 의사결정자, 행동하는 주체
- Reward: 행동에 대해 받는 보상
- (현재 타임스텝 t에서 a라는 행동을 했을 때 다음시점 t+1에서 받을 수 있는 보상)
- State: 환경의 변화를 표현하는 상태
- Action: Agent가 하는 행동 (대문자는 집합, 소문자는 한 상태 값)

$$A_t = a$$

[수식1] 시간 t에서의 행동

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

수식 2.7 보상함수의 정의

강화학습의 기본 용어

- Observation: State를 관찰하는 것
- Environment: 시스템을 Agent의 환경이라고 함.

- 상태 변환 확률 (상태 천이 변환 확률)

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

수식 2.12 상태 변환 확률

강화학습의 기본 용어

- Episode: 시작에서 끝까지 상태, 행동, 보상의 기록
  - 에피소드 (무한/유한): 행동-보상의 연속적인 집합
- Trajectory: 에이전트가 행동한 경로

$$\text{trajectory } \tau = (\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \dots)$$

강화학습에서는 어떤 상태에서, 어떤 행동을 했는지가 중요
=> 상태와 행동을 같이 기억하는게 좋음. (어떤 State에서 Action을 한다)

강화학습의 기본 용어

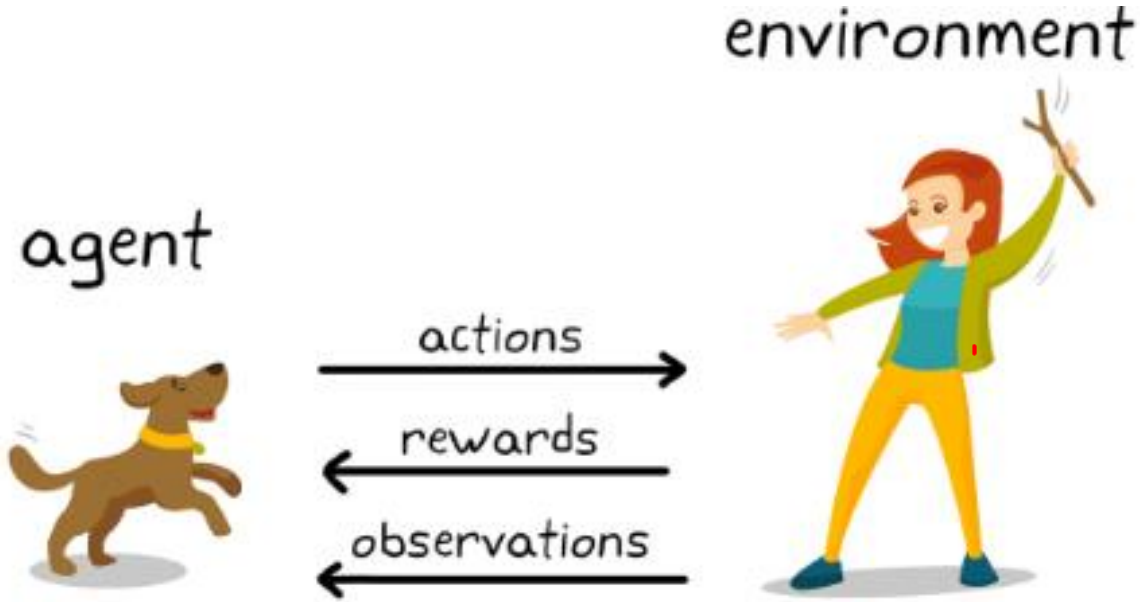
- Policy: 에이전트가 최선의 행동을 선택하기 위한 규칙/방법
  - 에이전트가 행동을 선택할 때 기준이 되는 것으로, **상태가 입력으로 들어오면 행동을** 출력으로 내보내는 일종의 함수
  - 각 상태마다 어떤 행동을 해야 할지 알려줌.
  - 정책은 각 상태에서 하나의 행동만을 나타낼 수도 있고, 확률적으로 각 행동에 대한 확률 값 a=10%, b=90%와 같이 나타낼 수도 있음
  - 에이전트가 강화학습을 통해 학습해야할 것은 최적 정책

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

수식 2.15 정책의 정의

강화학습의 목표

- Reward(보상)을 최대화하는 방향으로 정책을 학습 (Reward Function)
- RL 최종 목표 - 강화학습은 단기 보상을 최대화하는 것이 아닌
  최종 누적 합의 보상을 최대화하는 것을 목적으로 한다. (최종 합)
  => 환경으로부터 받는 누적 보상을 최대화하는 최적 정책을 구하는 것



https://kr.mathworks.com/discovery/reinforcement-learning.html

반환값(discounted return)

- 시간 스텝 t 이후 미래에 얻을 수 있는 보상의 총합
- **감가율(할인율, discount factor):**
  - 에이전트는 항상 현재에 판단을 내리기 때문에, 현재에 가까운 보상일수록 더 큰 가치를 지님. 그래서 에이전트는 그 보상이 얼마나 시간이 지나서 받는지를 고려해서 현재의 가치로 따짐
  - 보상은 나중에 받을수록 가치가 줄어든다는 의미.
- R는 감가율로 감가율에 따라, 가까운 미래/먼 미래 중 어디에 더 가중치를 줄지 정할 수 있다.

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

반환값(discounted return)

- R(disciount factor): 0<=x<=1
- R(감가율)이 클수록:
  - 가까운 미래의 보상에 더 큰 가중치를 둠

- R(감가율)의 역할:
  - 반환값이 무한대로 발산하는 것을 막는 수학적 장치 역할을 함.

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

수식 2.13 할인율의 정의

강화학습의 분류

- On-Policy
  - 정책 업데이트 시에 실제로 행동하고 있는 **최신 Policy로 수집된 데이터만 사용**하는 방식

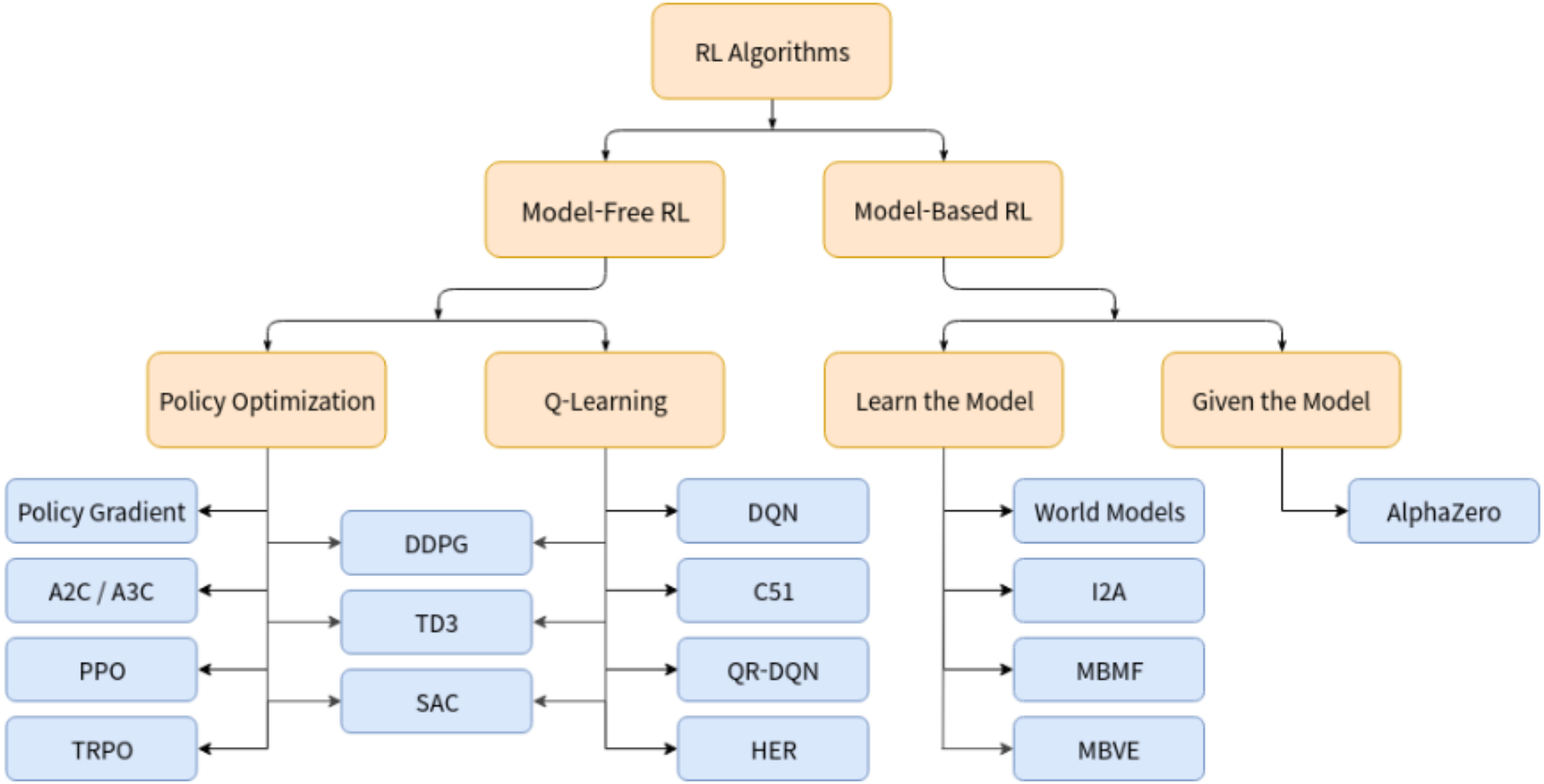
- Off-Policy
  - 정책 업데이트에 **어떤 데이터를 써도 관계 없고**, 최근에 업데이트한 정책에서 수집된 데이터 아니여도 상관 없음.

강화학습의 분류

- Model-Free RL
  - Environment의 Model은 모르지만 Interaction을 통해 문제를 푸는 것, **시행착오에 의한 학습** (Learning)

- Model-Based RL
  - Environment의 Model을 어느정도 **알거나 주어진 상태에서 문제를 푸는 것**. (Planning, 로봇 제어, Dynamic Programming)

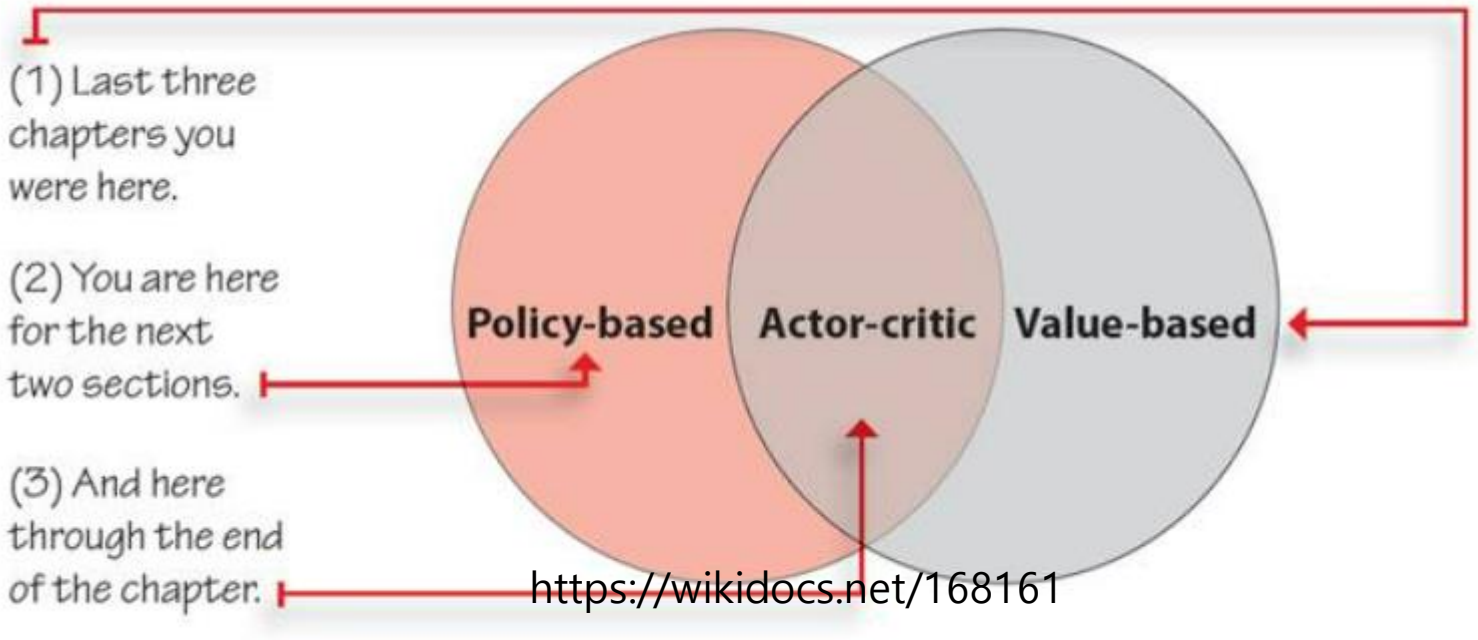
강화학습 알고리즘 분류



https://velog.io/@kjb0531/%EA%B0%95%ED%99%94%ED%95%99%EC%8A%B5-%EA%B0%9C%EB%85%90%EC%A0%95%EB%A6%AC3

Actor-Critic 알고리즘 (DeepMind AlphGo)

- Actor-Critic은 정책함수와 가치함수를 함께 학습하는 방법
- Actor는 상태가 주어졌을 때 행동을 결정하고,
- Critic은 상태의 가치를 평가한다.
- 이 알고리즘을 기반으로 A2C, A3C 등의 알고리즘이 나옴



https://wikidocs.net/168161

Markov Decision Process (MDP)

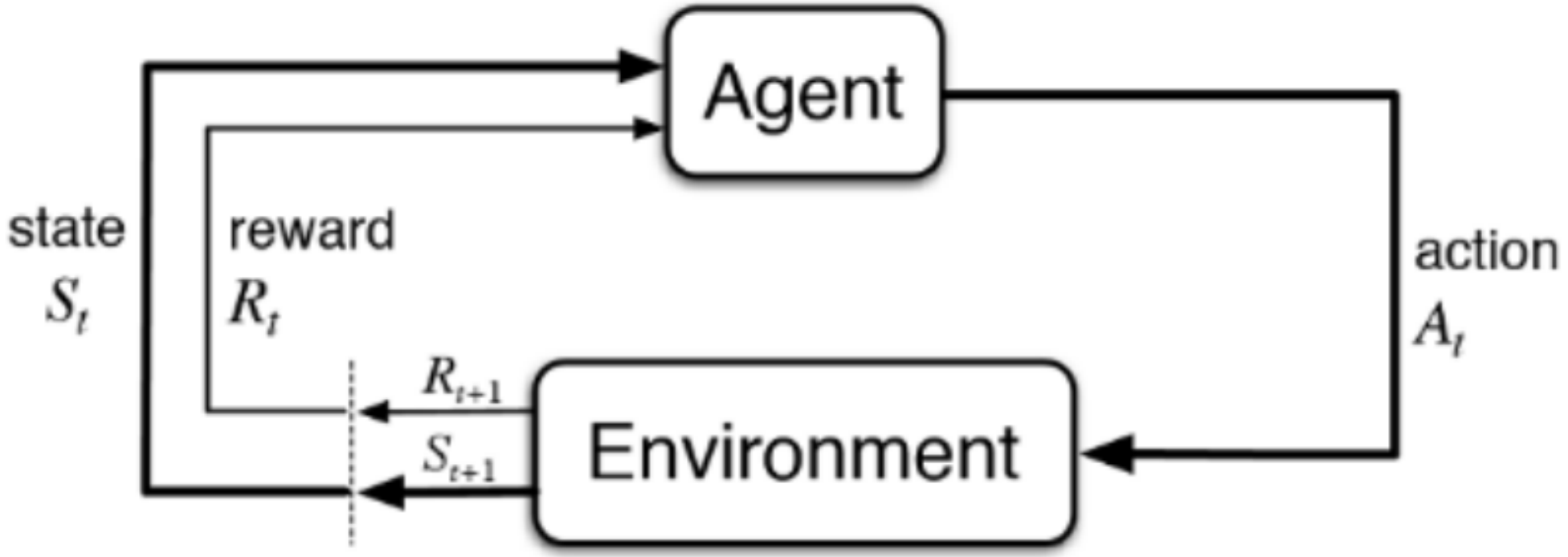
- 마르코프 결정 프로세스는 강화학습의 가장 기본적인 방법론
- 상태, 상태천이 확률밀도함수, 행동, 보상함수로 이루어짐
- **순차적으로 행동을 결정**해야하는 **문제를 풀기위한 수학 모델**

- 용어
  - 행동: a_t, U_t
  - 상태: S_t, X_t,
  - 정책: $\pi(u|x)$
  - 상태천이 확률 밀도 함수$P(x|x,u)$

S_t = s에 에이전트가 있을 때 가능한 행동 중 A_t = a를 할 확률 (정책, Policy)
$\pi(a \mid s) = \mathrm{P}[\mathrm{A_t} = a \mid \mathrm{S_t} = \mathrm{s}]$

강화학습의 프로세스

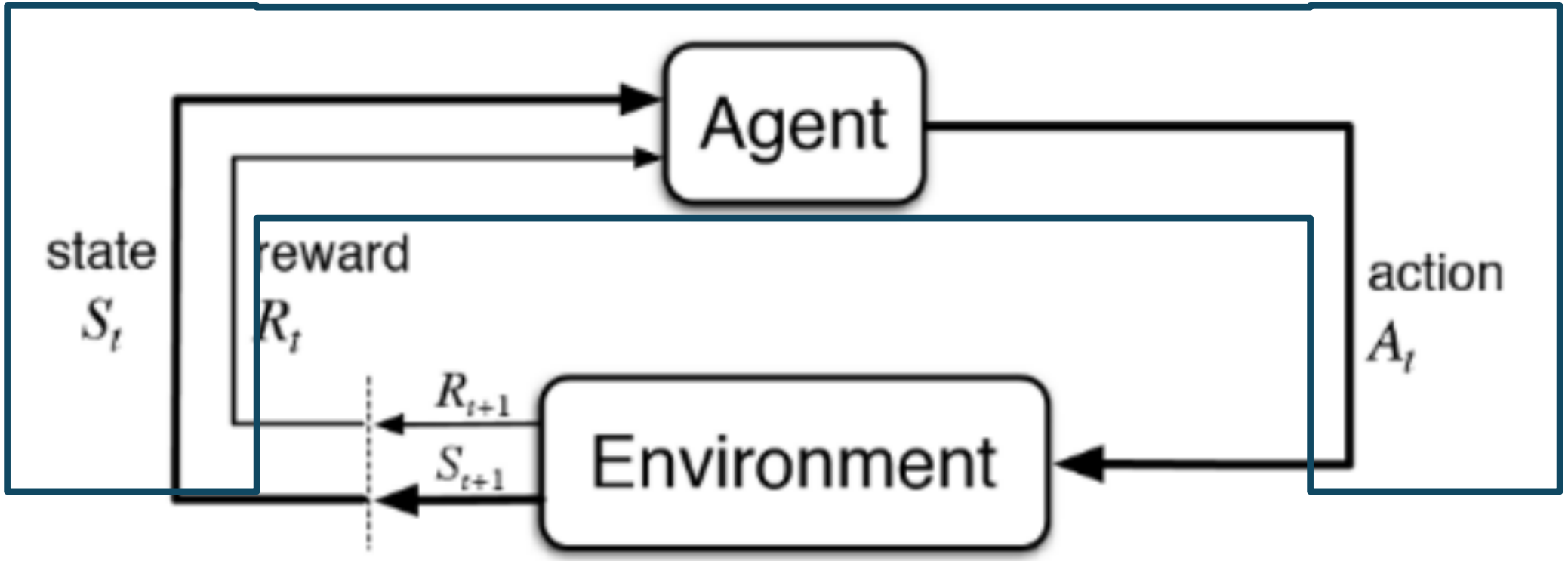
- 강화학습의 프로세스를 수학적으로 나타낸 것이 Markov Decision Process(MDP)



출처: https://velog.io/@jekim5418/Reinforcement-Learning%EA%B0%95%ED%99%94%ED%95%99%EC%8A%B5-%EA%B0%9C%EB%85%90

강화학습의 프로세스

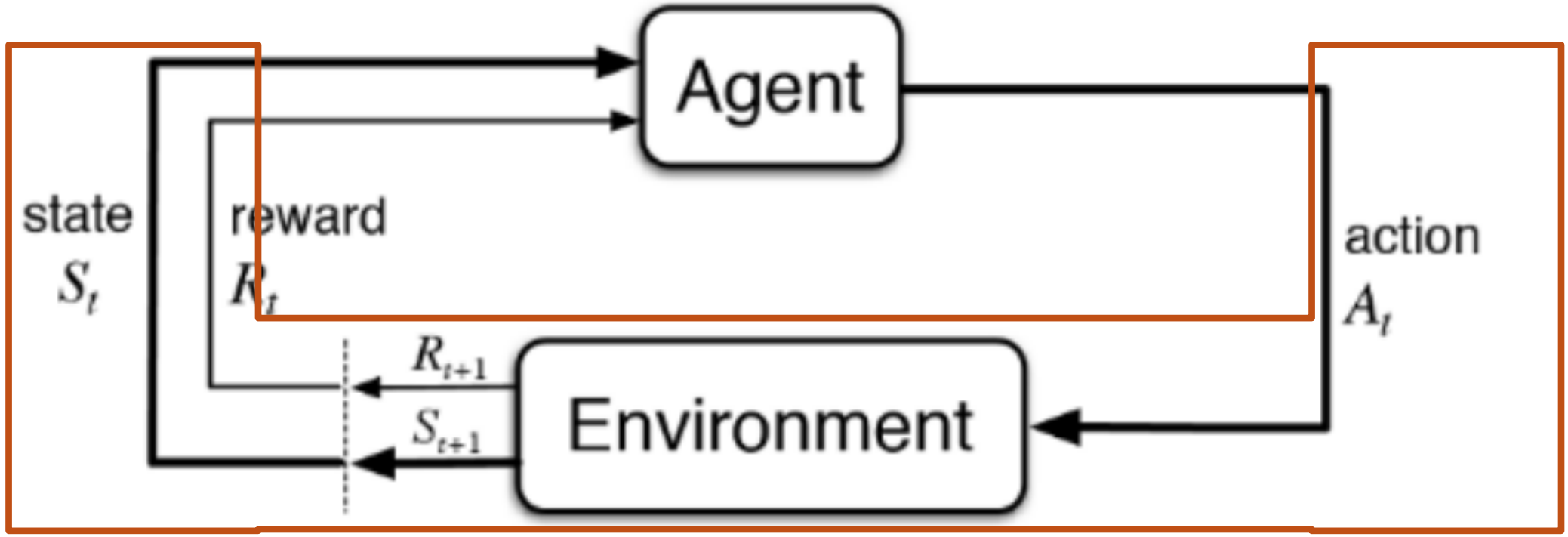
Agent가 한 상태(State)에서, 정책(규칙, Policy)에 의해 행동(Action)한다.



출처: https://velog.io/@jekim5418/Reinforcement-Learning%EA%B0%95%ED%99%94%ED%95%99%EC%8A%B5-%EA%B0%9C%EB%85%90

강화학습의 프로세스

행동에 의해 환경이 행동에 대해 보상을 주고 다음 상태(t+1)으로 업데이트



=> 이를 수학적으로 모델링한 것이 마르코프 결정 프로세스(MDP)

정리

- 어떤 상태에서 **정책(기준)에 의해 행동**을 한다.
- **정책에 의해 행동**이 선택되면 **상태가 업데이트** 된다.
- **상태가 업데이트** 되면, 환경에 의해 **보상이 주어진다.**

*정책: 측정한 상태를 바탕으로 최선의 행동을 선택하기 위한 에이전트의 규칙/방법

$r(x_t, u_t)$- 상태 x_t에서 행동 u_t를했을 때 받는 보상

⇒**어떤 상태에서 정책에 의해 행동하면**, 보상이 주어진다.

⇒어떤 State에서 Policy에 의해 Action하면, Reward가 주어진다.

(State-Action-Reward)

행동했을 때 기대할 수 있는 반환값을 상대가치라고 하고,
이를 구하는 것을 가치함수라고 한다(Q함수, 벨만 기대 방정식 등의 얘기가 전개됨)

STUDY PRESENTATION

THANK YOU

발표를 마치겠습니다

궁금한 점은 바로 질문해 주세요!