강화학습 원리 스터디

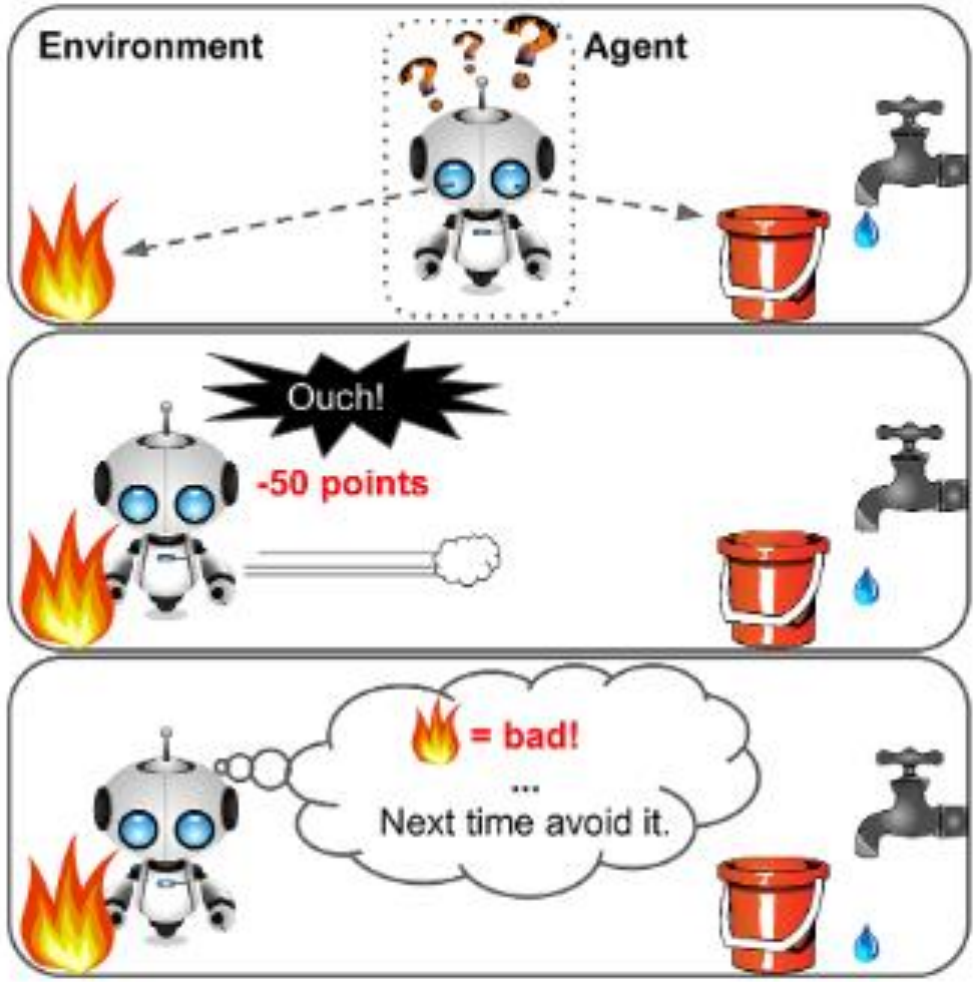
1장 – 밴디트 학습

강경민

목차

- 강화 학습이 무엇인가
- 밴디트 문제
- 밴디트 알고리즘
- 비정상 문제

강화 학습?

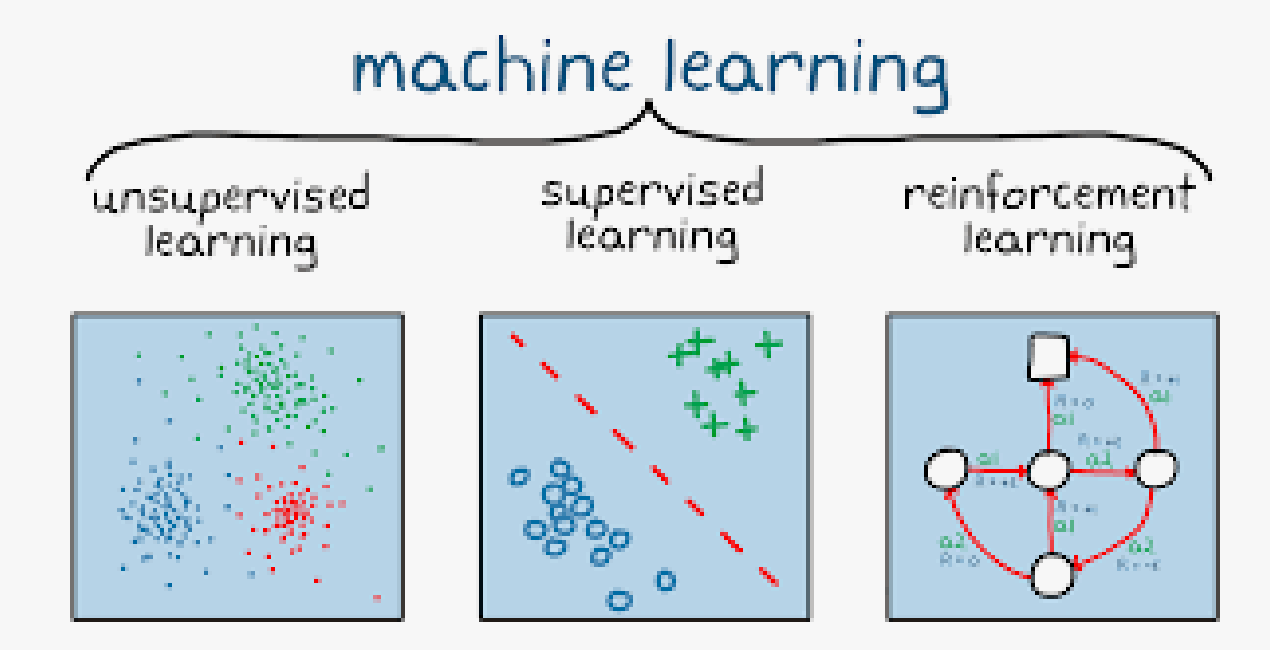


강화 학습의 구성 요소

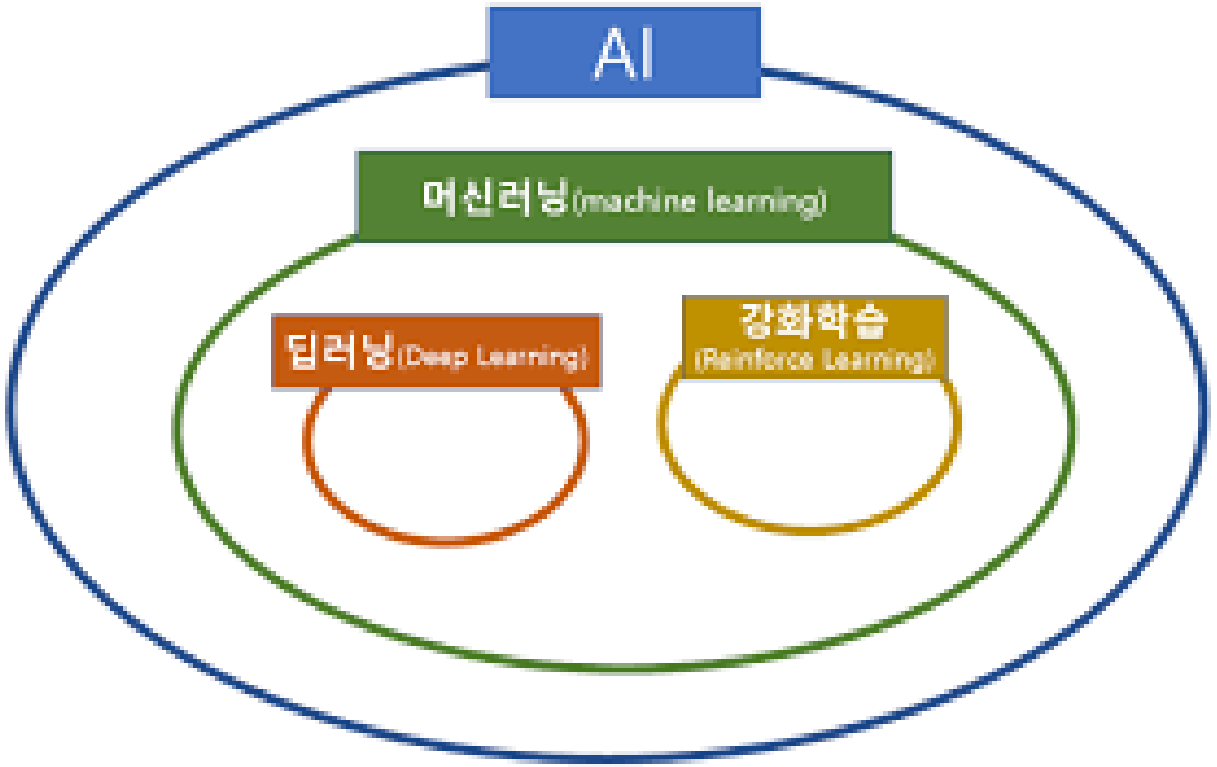
- 에이전트: 행동 주체
- 환경: 에이전트의 행동에 따라 보상 제공
- 목표: 보상을 극대화하는 행동 패턴 익히기

머신러닝 분류

- 비지도 학습: 데이터에 있는 구조 학습, 정답 X
- 지도 학습: 입력 -> 출력으로 변환하는 방법 학습, 정답 O
- 강화 학습: 에이전트-환경 상호작용, 더 많은 보상 얻는 방법 학습

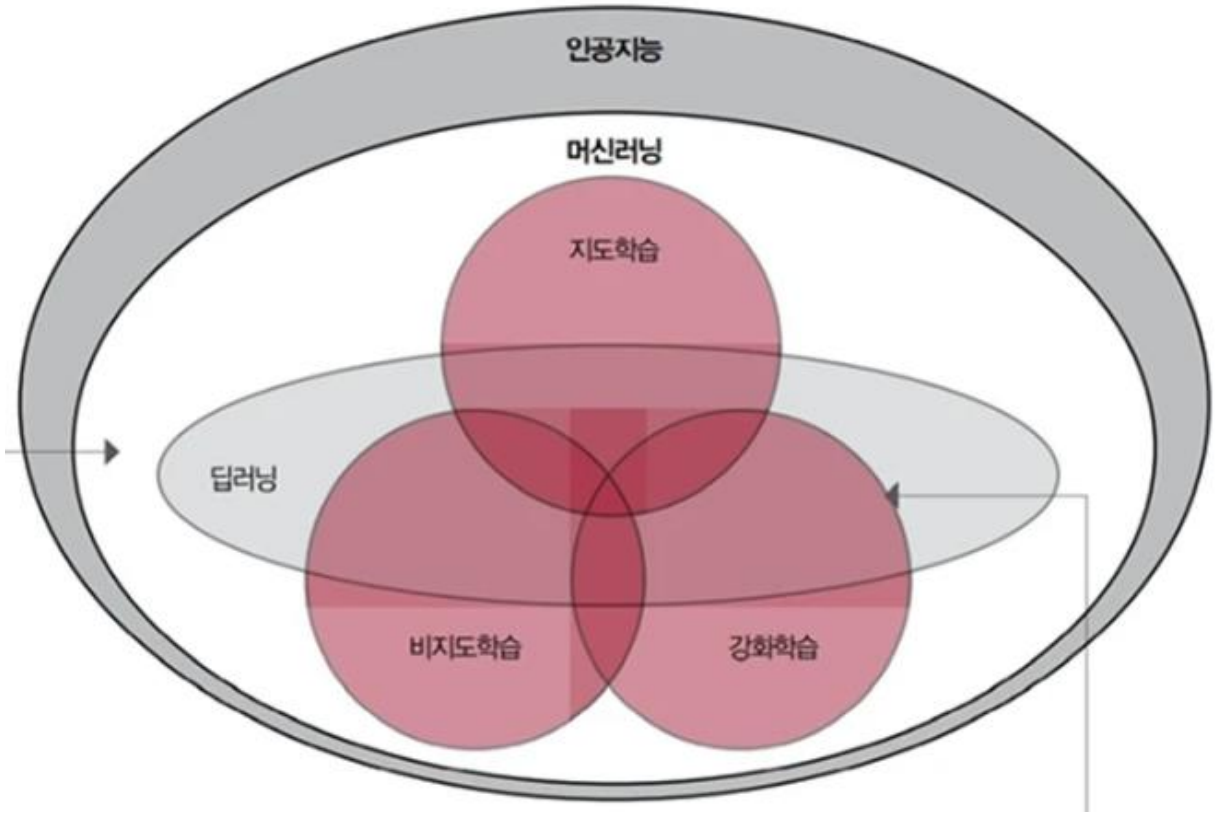


강화학습 vs 딥러닝?


AI
머신러닝(machine learning)
딥러닝(Deep Learning)
강화학습
(Reinforce Learning)

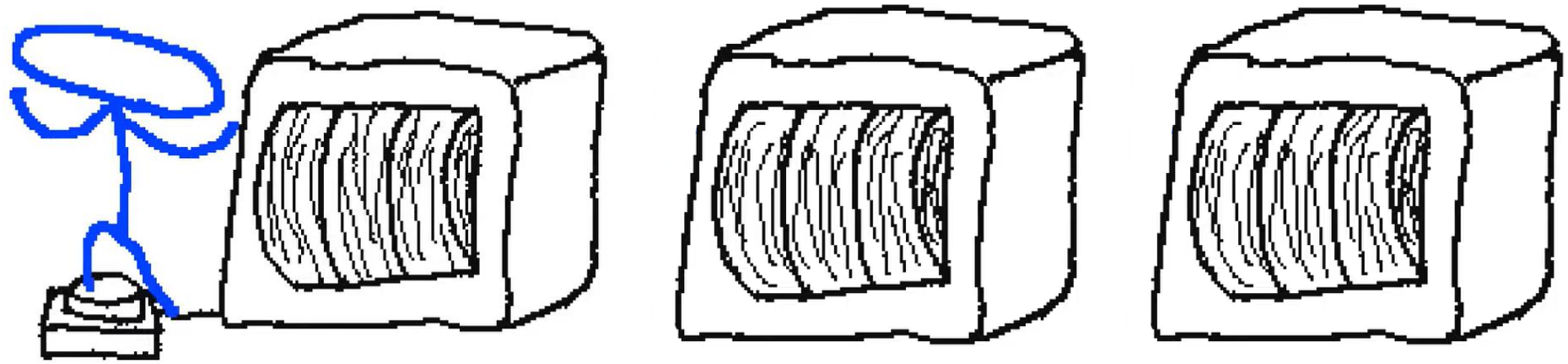
강화학습 vs 딥러닝?

- 심층 강화학습: 강화학습 + 딥러닝



밴디트 문제

- 여러 슬롯머신이 있는데, 각 슬롯의 보상 분포는 다름
- 목표: 최대한 많은 코인(보상) 얻기
- 그런데 플레이어는 각 슬롯의 가치를 모름



밴디트 알고리즘

- 몇 번 슬롯을 돌려보고 나온 보상의 평균으로 가치 추정

$$Q_n = \frac{R_1 + R_2 + \ldots + R_n}{n}$$

Q - 추정치, R - 나온 보상

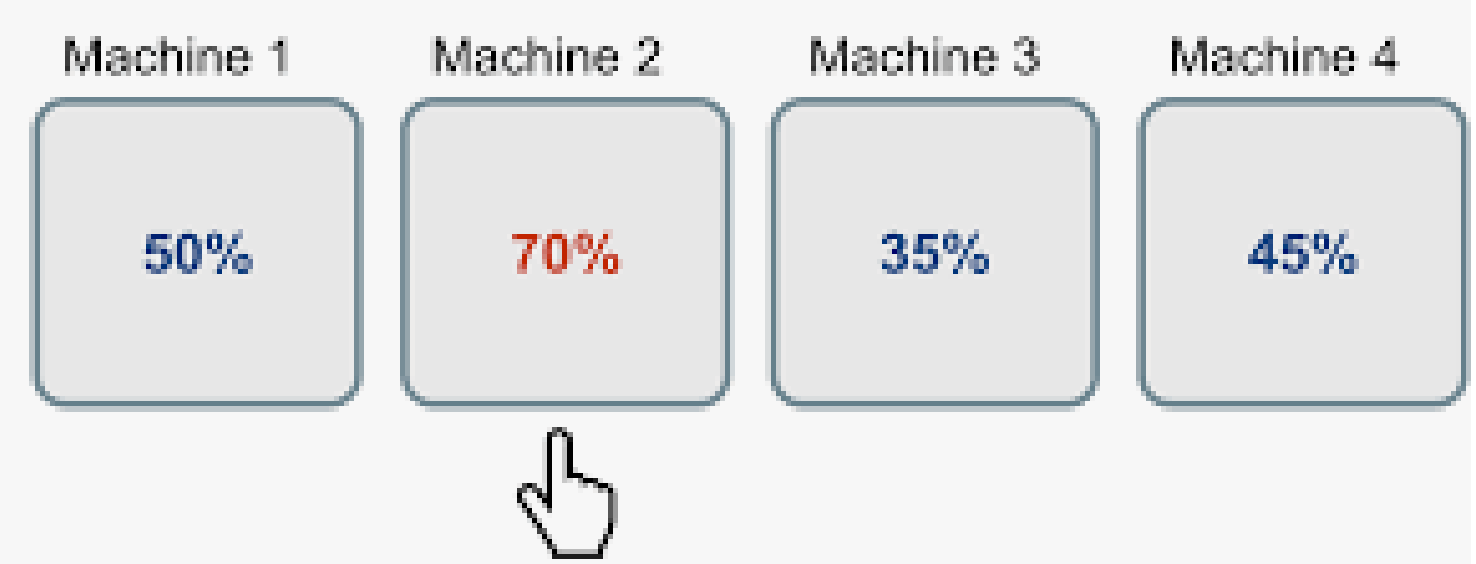
밴디트 알고리즘

- 추정치를 구하는 코드의 최적화

$$Q_n = Q_{n-1} + \frac{1}{n}(R_n - Q_{n-1})$$

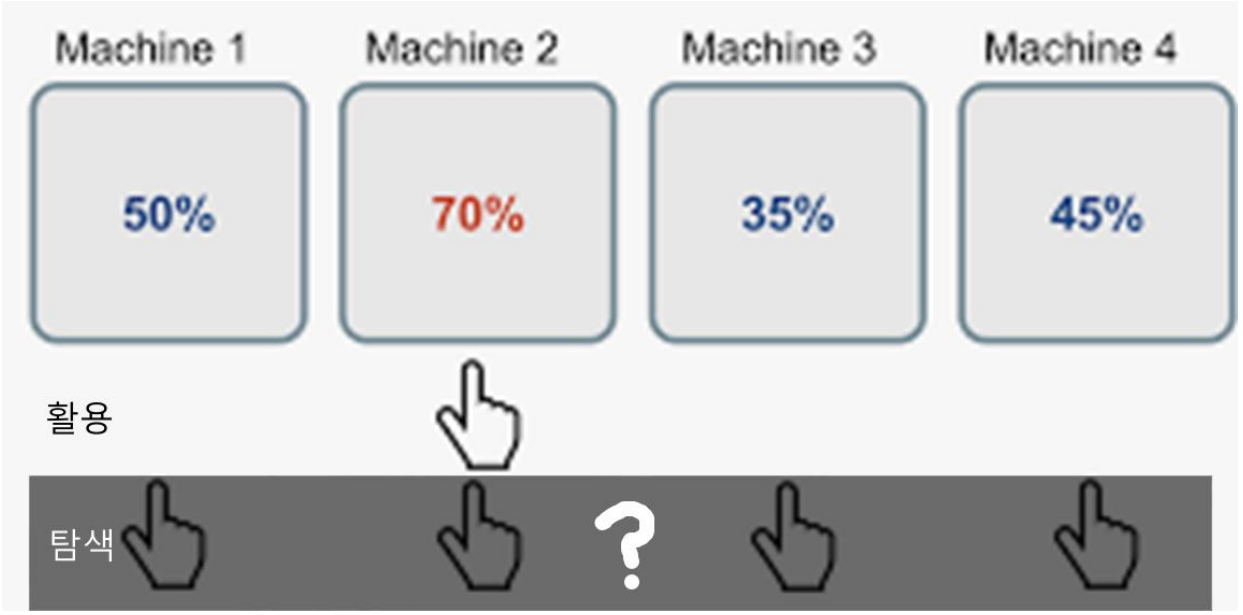
플레이어의 정책

- 탐욕 정책: 가장 가치 추정치가 높은 슬롯머신 선택
- 문제: 가치 추정치는 부정확하므로 더 좋은 슬롯을 놓칠 수 있음



플레이어의 정책

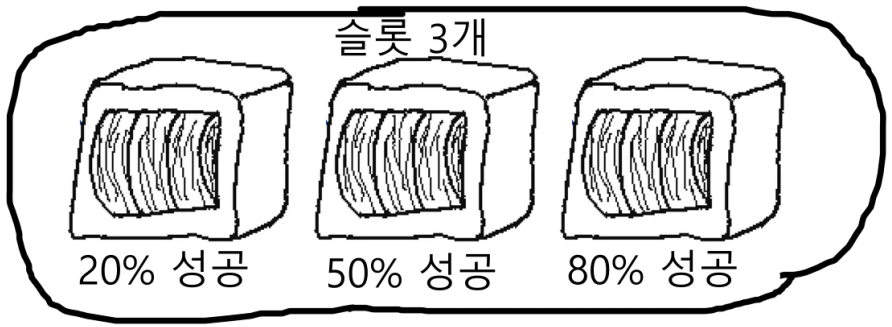
- 활용 (Exploitation): 가장 가치 추정치가 높은 슬롯 선택
- 탐색 (Exploration): 가치 추정을 위해 아무 슬롯이나 시도
- ε-탐욕 정책: ε의 확률로 탐색, 나머지는 활용



밴디트 알고리즘 구현

- Bandit 클래스: 슬롯의 개수 & 각각의 성공 확률
- Agent 클래스: 탐색 선택 확률 (ε),
  각 슬롯의 가치 추정치 Qs, 돌린 횟수 ns

Bandit



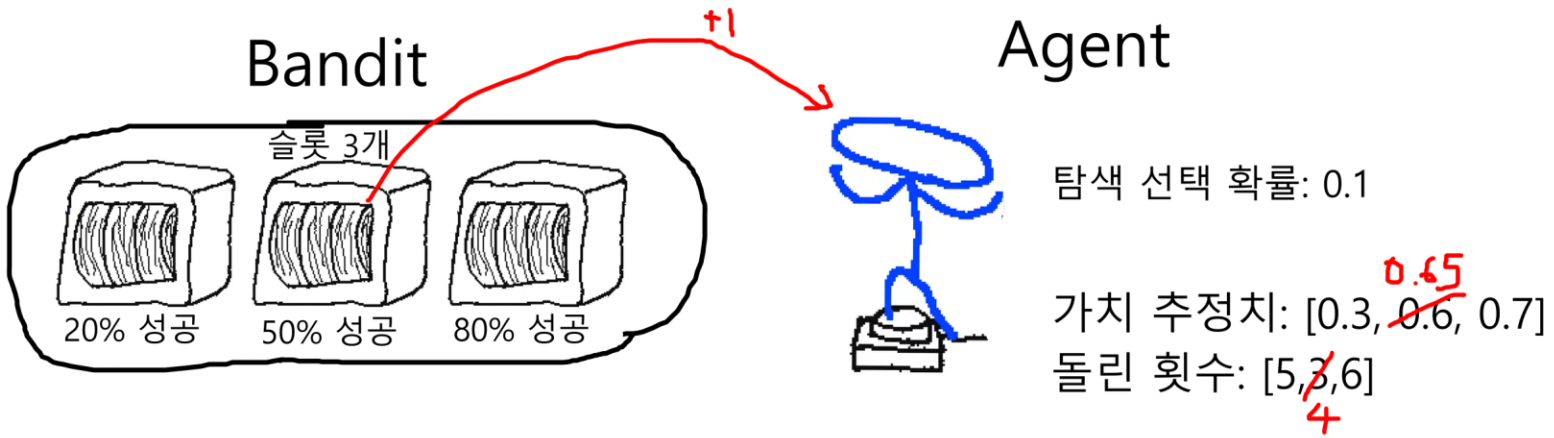
Agent

탐색 선택 확률: 0.1

가치 추정치: [0.3, 0.6, 0.7]
돌린 횟수: [5,3,6]

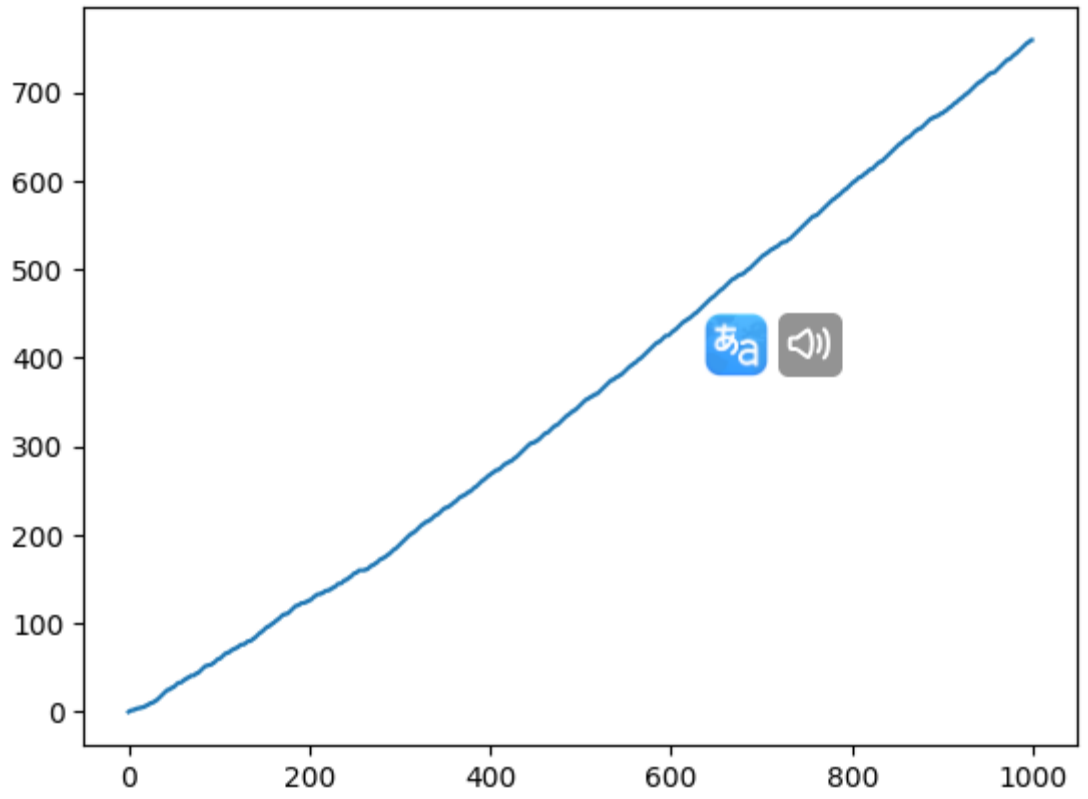
밴디트 알고리즘 구현

- Bandit 클래스: play() – 성공 확률에 기반해 보상 제공
- Agent 클래스: update() – 보상에 따라 Q, n 갱신



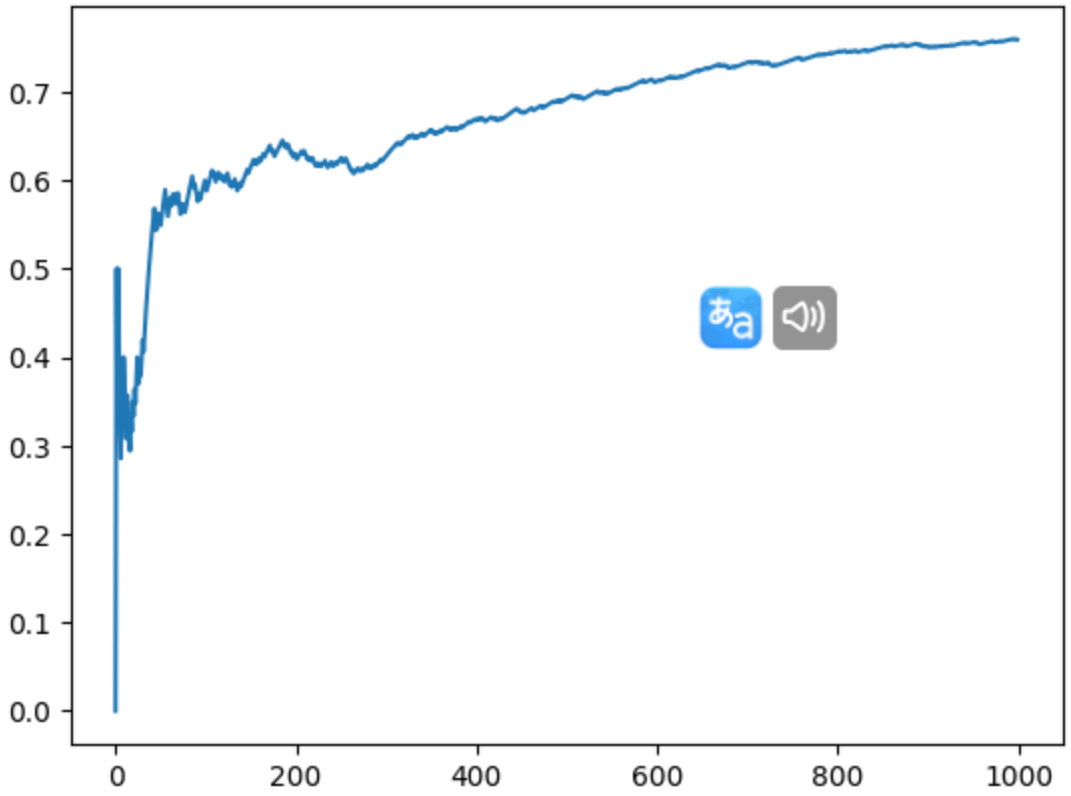
밴디트 알고리즘 구현

- X축: 시도 횟수
- Y축: 얻은 보상



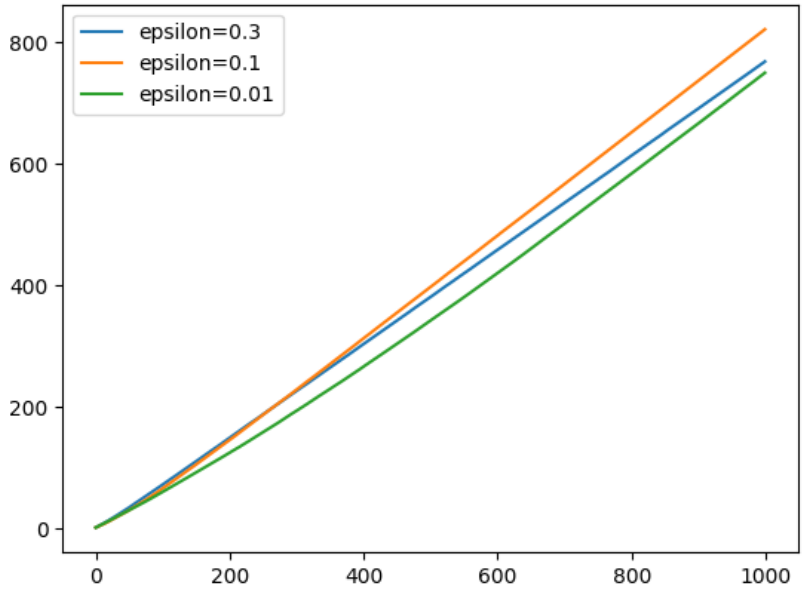
밴디트 알고리즘 구현

- X축: 시도 횟수
- Y축: 승률



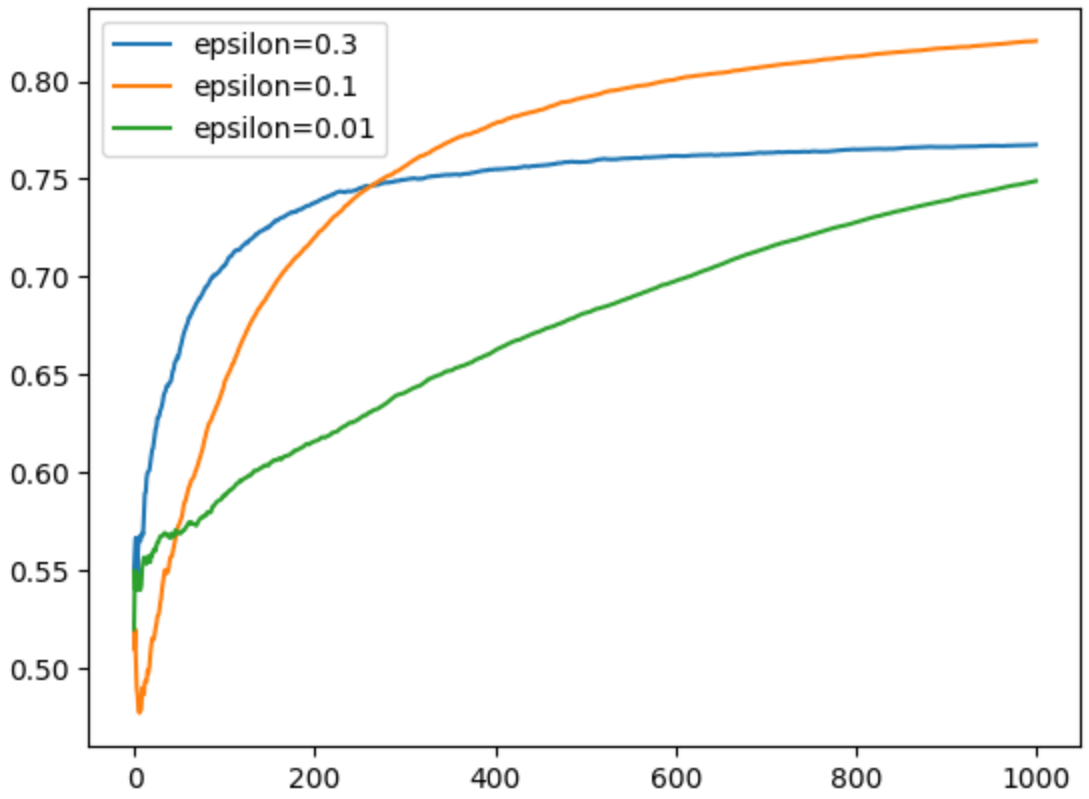
밴디트 알고리즘 구현

- 코드에 무작위성이 있으므로, 한 번의 실험이 아닌 여러 번의 평균을 이용해야 알고리즘의 우수성을 판단할 수 있음
- X축: 시도 횟수
- Y축: 얻은 보상



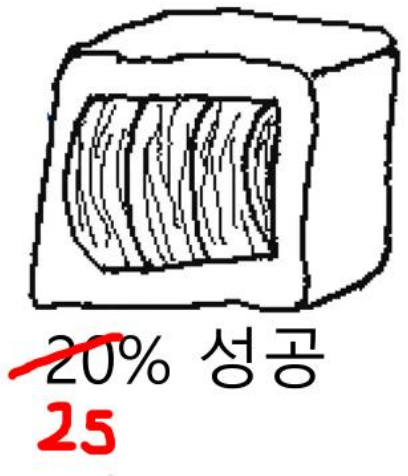
밴디트 알고리즘 구현

- X축: 시도 횟수
- Y축: 승률



비정상 문제

- 정상 문제: 보상의 확률 분포가 변하지 않는 문제
- 비정상 문제: 보상의 확률 분포가 플레이 도중 변하는 문제



비정상 문제

- 기존에는 모든 보상에 똑같은 가중치를 부여했었음

$$Q_n = \frac{1}{n}R_1 + \frac{1}{n}R_2 + \ldots + \frac{1}{n}R_n$$

- 그러나 비정상 문제에는 최신 보상에 더 높은 가중치 부여

$$Q_n = Q_{n-1} + a(R_n - Q_{n-1}) \quad (0 < a < 1)$$

$$Q_n = \ldots + a(1-a)^2 R_{n-2} + a(1-a)^1 R_{n-1} + aR_n$$

비정상 문제

- 표본 평균 (정상 문제) – 균일한 가중치

$$Q_n = \frac{1}{n}R_1 + \frac{1}{n}R_2 + \ldots + \frac{1}{n}R_n$$

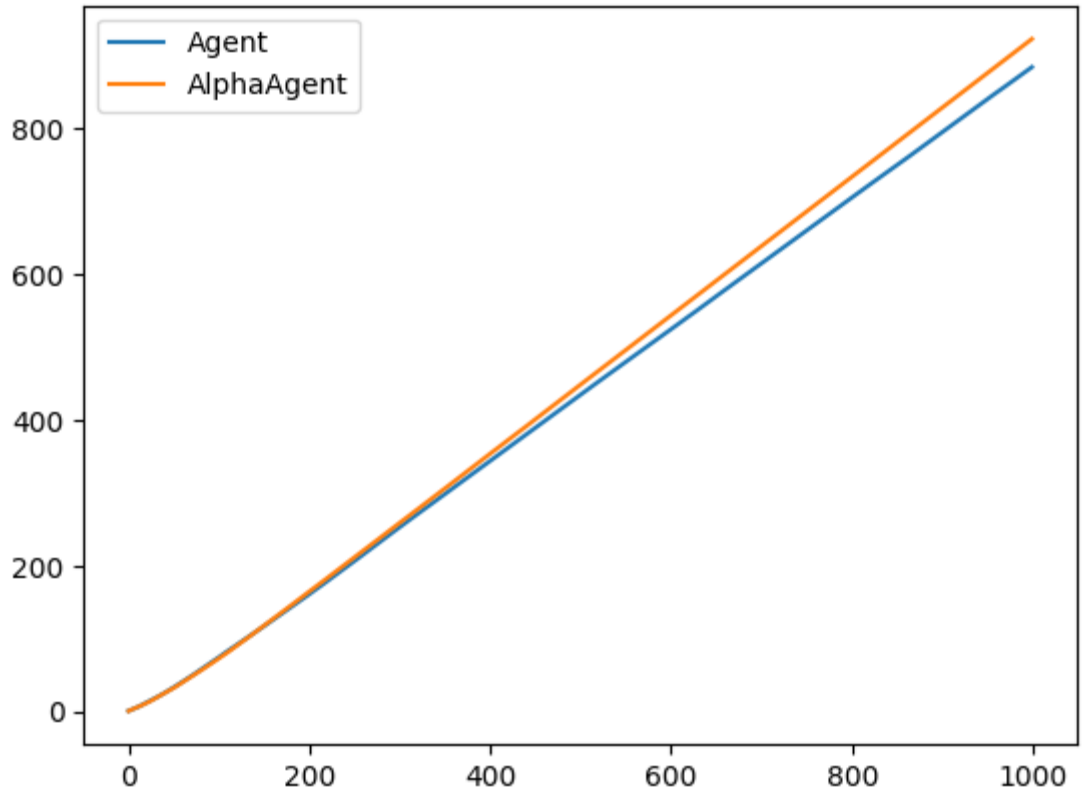
- 지수 이동 평균 (비정상 문제) – 최신 데이터에 더 큰 가중치

$$Q_n = Q_{n-1} + a(R_n - Q_{n-1}) \quad (0 < a < 1)$$

$$Q_n = \ldots + a(1-a)^2 R_{n-2} + a(1-a)^1 R_{n-1} + aR_n$$

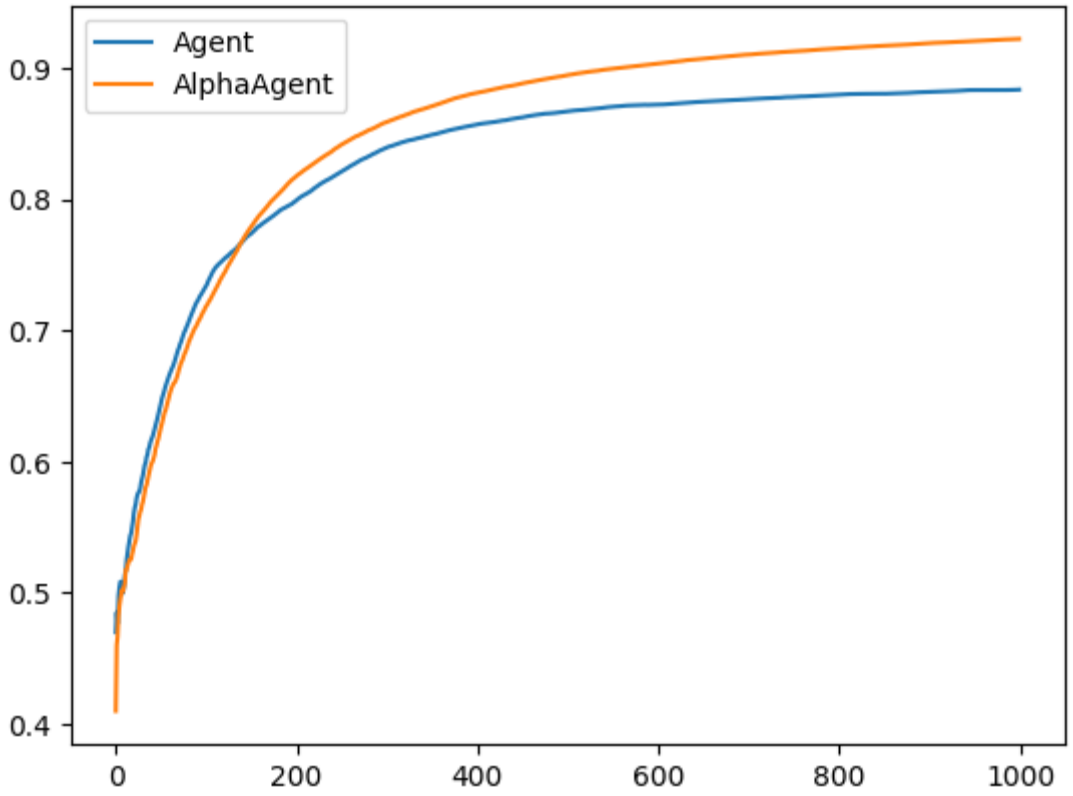
비정상 문제 – 에이전트 성능

- X축 – 시도 횟수
- Y축 – 얻은 보상



비정상 문제 – 에이전트 성능

- X축 – 시도 횟수
- Y축 – 승률



끝

- 감사합니다