강화학습 개요

Slides from

1. 이웅원 외, 파이썬과 케라스로 배우는 강화학습, 주교재
2. 이웅원, 가깝고도 먼 DeepRL, PPT
3. David Silver, Reinforcement Learning, PPT

References

1. Richard S. Sutton and Andrew G. Barto, Reinforcement Learning: An Introduction, MIT Press
2. 유튜브, 전민영, 노승은, 강화학습의 기초 이론, 팡요랩

Contents

머신러닝의 종류

- Machine Learning의 종류
  1. Supervised Learning : 정답이 있는 데이터 학습
  2. Unsupervised Learning : 데이터 자체의 특성 학습
  3. Reinforcement Learning : 보상으로부터 학습

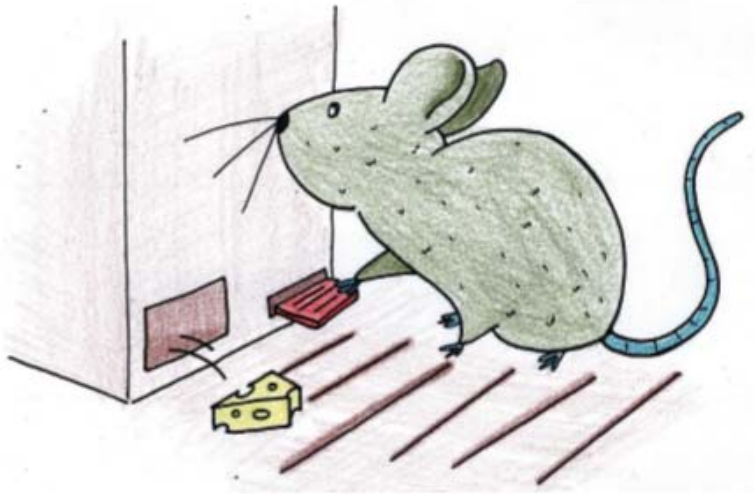
머신러닝의 분류

- 지도학습 (Supervised Learning)
  - 입력값과 함께 결과값(정답 레이블)을 같이 주고 학습을 시키는 방법
  - 분류(classification)
    - KNN, SVM, Decision Tree, ANN(Artificial Neural Network)
  - 예측(prediction)
    - 회귀(regression)
- 비지도 학습 (Unsupervised Learning)
  - 입력 데이터만을 이용한 학습
  - 군집(clustering)
    - K-means clustering, DBSCAN
  - 생성모델(Generative Model)
    - Auto-Encoder, GAN(Generative Adversarial Network)
- 강화 학습 (Reinforce Learning)
  - 결과값(label) 대신 보상(reward)로 주어짐
  - MDP, Q-learning, DQN, Actor-Critic

https://brunch.co.kr/@gdhan/10

Reinforcement란?

- Skinner : 행동심리학자
  - ➢강화(Reinforcement) : 동물이 시행착오(Trial and Error)를 통해 학습하는 방법
- Skinner의 쥐 실험 : 레버를 누르면 먹이가 나오는 상자 안에 굶긴 쥐를 넣고 실험



Reinforcement란?

1. 굶긴 쥐를 상자에 넣는다
2. 쥐는 돌아다니다가 우연히 상자 안에 있는 지렛대를 누르게 된다
3. 지렛대를 누르자 먹이가 나온다
4. 지렛대를 누르는 행동과 먹이와의 상관관계를 모르는 쥐는 다시 돌아다닌다
5. 그러다가 우연히 쥐가 다시 지렛대를 누르면 쥐는 이제 먹이와 지렛대 사이의 관계를 알게 되고 점점 지렛대를 자주 누르게 된다
6. 이 과정을 반복하면서 쥐는 지렛대를 누르면 먹이를 먹을 수 있다는 것을 학습한다[1]

https://namu.wiki/w/행동주의

- Reinforcement : 동물이 이전에 배우지 않았지만 직접 시도하면서 행동과 그 결과로 나타나는 보상 사이의 상관관계를 학습하는 것
  -> 보상을 많이 받는 행동을 점점 더 많이 하는 것

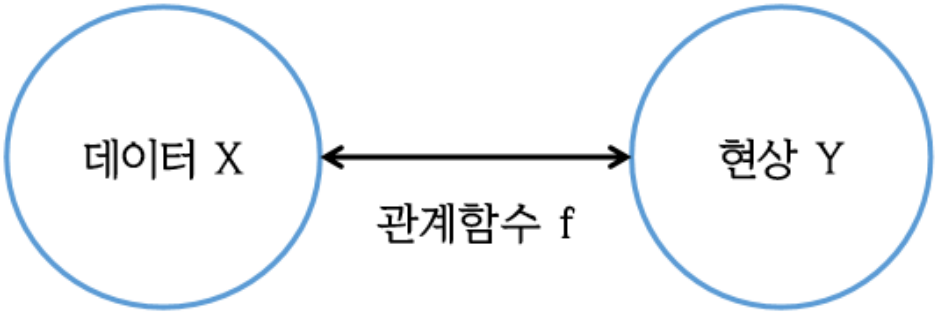
강화학습(Reinforcement learning)이란?

- 기계 학습의 한 영역이다.
- 행동심리학에서 영감을 받았으며,
- 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법이다.

https://ko.wikipedia.org/wiki/%EA%B0%95%ED%99%94_%ED%95%99%EC%8A%B5

강화학습이란?

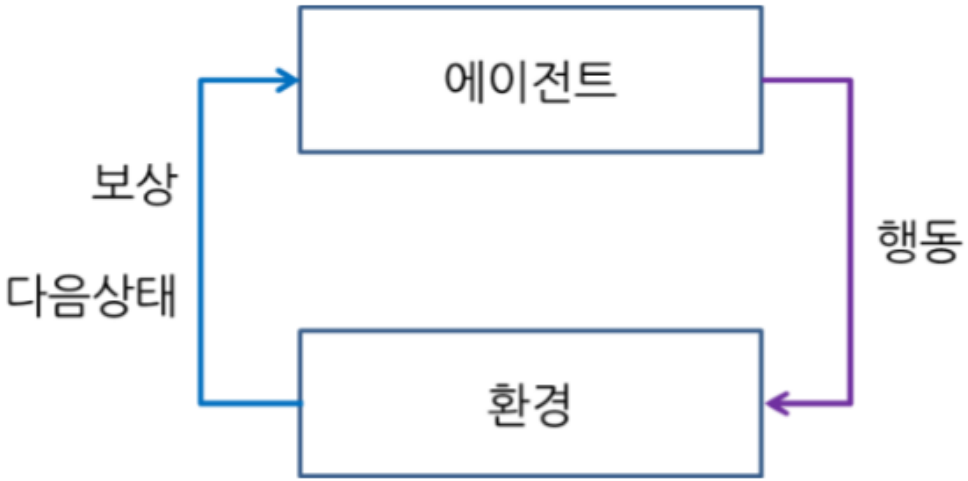
- Reinforcement Learning = Reinforcement + Machine Learning



- 데이터 X : 어떤 상황에서 어떤 행동을 했는지,
  현상 Y : 보상을 얼마나 받았는지
  -> 어떤 행동을 해야 보상을 많이 받는지를 데이터로부터 학습

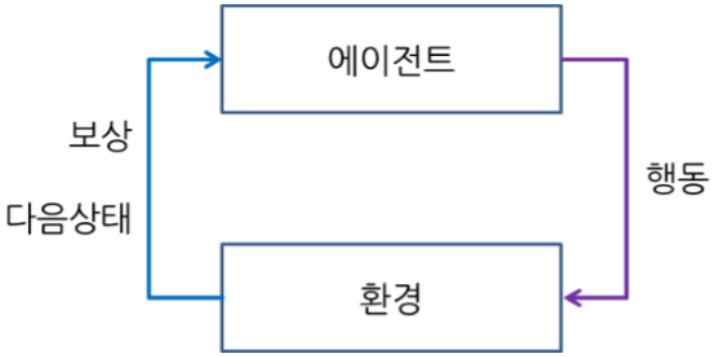
강화학습이란?

- 에이전트와 환경의 상호작용 -> 데이터 생성 (미리 모아 놓을 수 없다)
- 특정 상태에서 특정 행동을 선택 -> 보상 -> 학습



에이전트와 환경의 상호작용

- Agent : 강화학습을 통해 스스로 학습하는 컴퓨터
  - ➢환경에 대한 사전 지식이 없는 상태에서 학습
  - ➢자신이 놓인 상태에서 자신의 상태를 인식한 후 행동
- Environment(환경) : Agent에게 보상을 주고 다음 상태를 알려줌



- Reward(보상) : Agent가 선택한 행동에 대한 환경의 반응 -> Agent는 보상을 통해 학습
- 강화학습 : Agent가 마치 사람처럼 환경과 상호작용하면서 스스로 학습하는 방식 -> 목적 : Agent가 환경을 탐색하면서 얻는 보상들의 합을 최대화

Characteristics of Reinforcement Learning

- What makes reinforcement learning different from other machine learning paradigms?
  - There is no supervisor, only a reward signal
  - Feedback is delayed, not instantaneous
  - Time really matters (sequential, non i.i.d data)
  - Agent's actions affect the subsequent data it receives

David Silver, Reinforcement Learning, PPT

Rewards

- A reward R_t is a scalar feedback signal
- Indicates how well agent is doing at step t
- The agent's job is to maximise cumulative reward

Reinforcement learning is based on the reward hypothesis

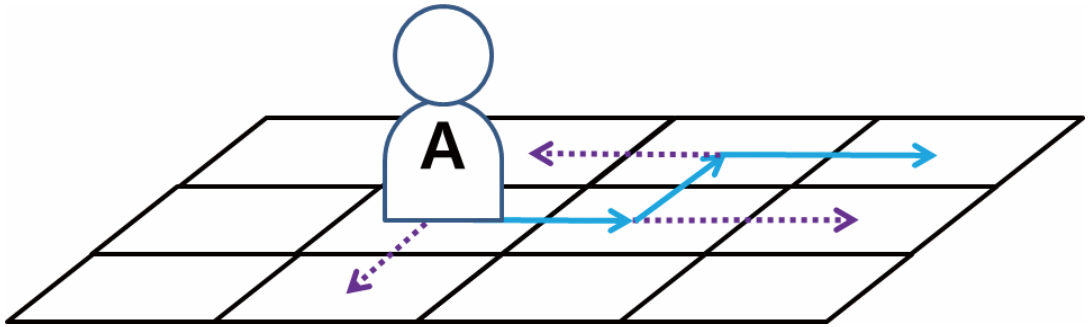
Definition (Reward Hypothesis)

All goals can be described by the maximisation of expected cumulative reward

David Silver, Reinforcement Learning, PPT

강화학습 문제

- 강화학습은 어떤 문제를 풀기 위한 것일까?
  -> 순차적 행동 결정 문제 (Sequential Decision Making)



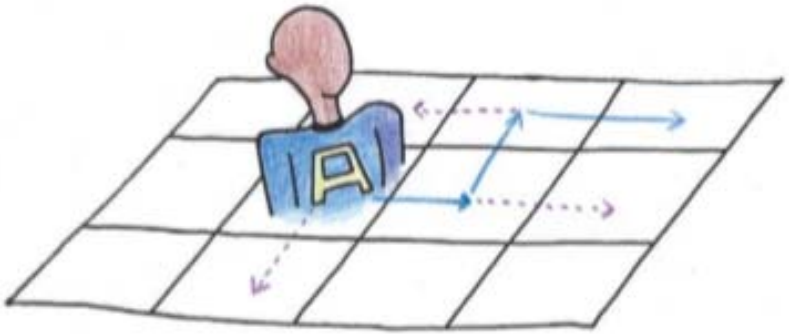
- 순차적 행동 결정 문제의 수학적 표현 -> Markov Decision Process (MDP)

Sequential Decision Problem

- 여러 번의 연속적 선택을 하는 문제 : Sequential Decision Problem
  - 주식 투자
  - 게임 (아타리, 바둑 등)
  - 집에서 직장까지 차를 운전하는 경우

Sequential Decision Problem

- 에이전트 : 환경을 관찰, 행동을 선택, 목표지향,
  -> “Decision Maker!”



- an autonomous, goal-directed entity which observes and acts upon an environment – 위키피디아
- 환경을 관찰하고 행동하는 자율적이고 목표 지향적인 주체 - 구글번역기

Sequential Decision Problem

- 환경 : 에이전트를 제외한 나머지

아이라는 주체를 빼고 길과 자전거 등이 환경이 된다

Sequential Decision Making

- Goal: *select actions to maximise total future reward*
- Actions may have long term consequences
- Reward may be delayed
- It may be better to sacrifice immediate reward to gain more long-term reward
- Examples:
  - A financial investment (may take months to mature)
  - Refuelling a helicopter (might prevent a crash in several hours)
  - Blocking opponent moves (might help winning chances many moves from now)

David Silver, Reinforcement Learning, PPT

순차적 행동 결정 문제(MDP)의 구성 요소

1. 상태(state)
   - ➢현재 에이전트의 정보
   - ➢환경으로부터의 관찰(observation)
   - ➢정적인 요소 + 동적인 요소

예) 탁구 에이전트의 상태
   - ➢탁구공의 위치, 속도, 가속도

순차적 행동 결정 문제(MDP)의 구성 요소

2. 행동(action)
   - 에이전트가 어떤 상태에서 취할 수 있는 행동
   - 예1: 상, 하, 좌, 우
   - 예2: 게임기의 입력
   - 에이전트는 학습을 하면서 보상을 최대화 할 행동을 할 확률을 높임

3. 보상(reward)
   - 에이전트가 학습할 수 있는 유일한 정보
   - 보상을 통해 에이전트는 자신이 했던 행동을 평가함
   - 보상은 에이전트에 속하지 않는 환경의 일부로서 행동을 하기 전에는 미리 알 수 없음

순차적 행동 결정 문제(MDP)의 구성 요소

4. 정책(policy)
   - ➢순차적 행동 결정 문제에서 구해야 할 답
   - ➢모든 상태에 대해 에이전트가 어떤 행동을 해야 하는지 정해 놓은 것
   - ➢최적 정책(optimal policy) : 보상의 합을 최대로 받을 수 있는 정책
   - ➢에이전트가 최적 정책을 얻었다는 것은 순차적 행동 결정 문제를 풀었다는 의미임

Major Components of an RL Agent

- An RL agent may include one or more of these components:
  - Policy: agent's behaviour function
  - Value function: how good is each state and/or action
  - Model: agent's representation of the environment

David Silver, Reinforcement Learning, PPT

Policy

- A policy is the agent's behaviour
- It is a map from state to action, e.g.
- Deterministic policy: $a = \pi(s)$
- Stochastic policy: $\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$

David Silver, Reinforcement Learning, PPT

Value Function

- Value function is a prediction of future reward
- Used to evaluate the goodness/badness of states
- And therefore to select between actions, e.g.

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s \right]$$

David Silver, Reinforcement Learning, PPT

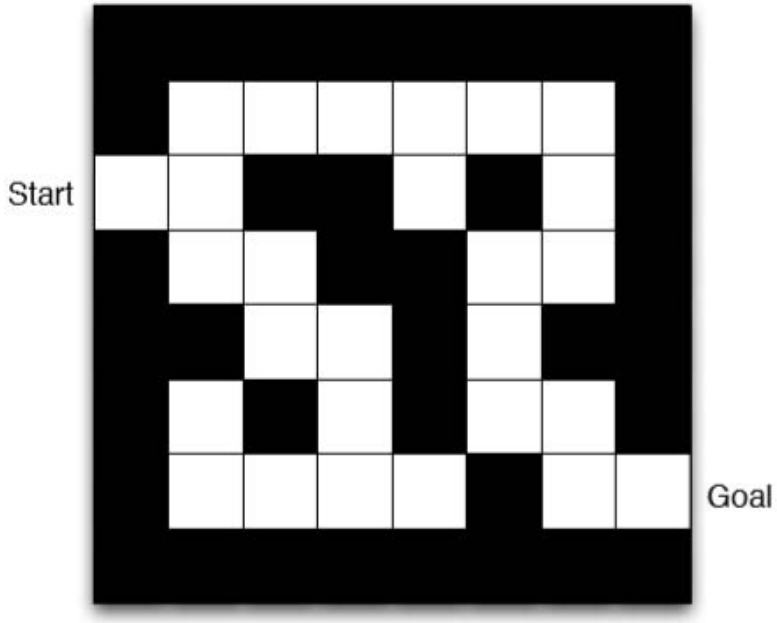
Model

- A model predicts what the environment will do next
- $\mathcal{P}$ predicts the next state
- $\mathcal{R}$ predicts the next (immediate) reward, e.g.

$$\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$$
$$\mathcal{R}^a_s = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$$

David Silver, Reinforcement Learning, PPT

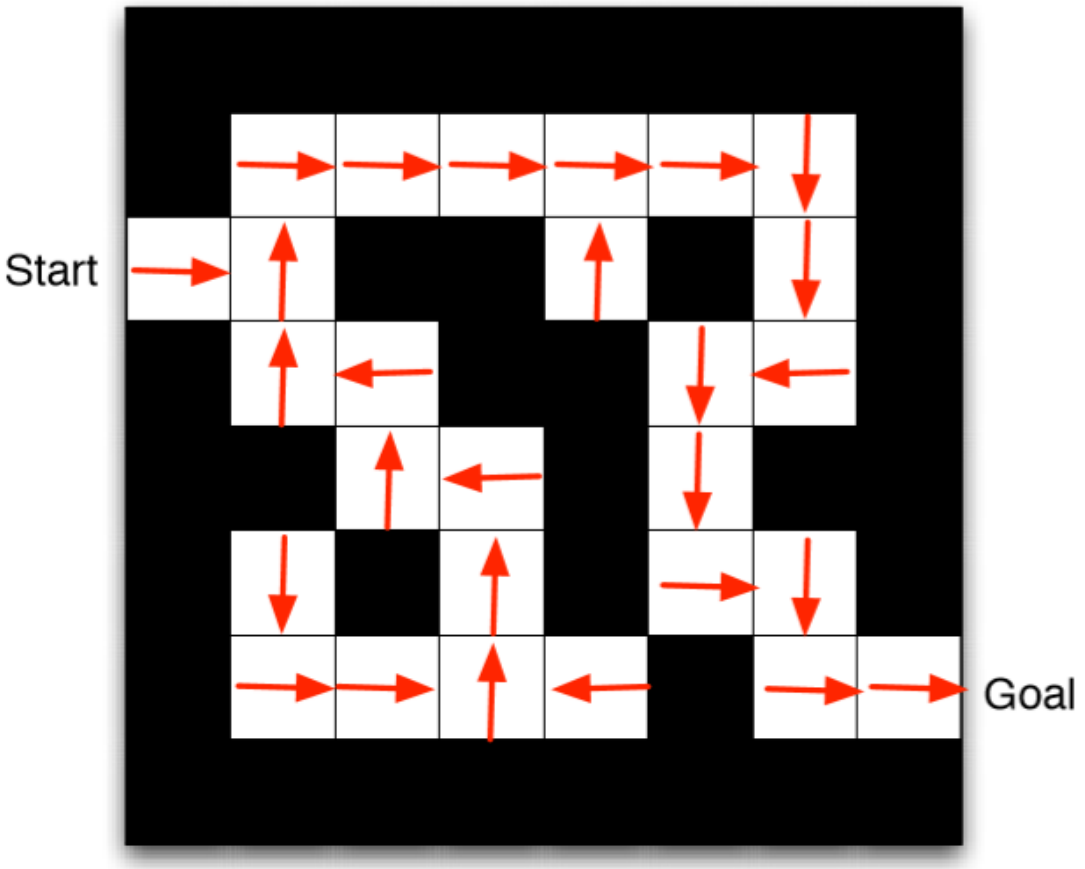
Maze Example



- Rewards: -1 per time-step
- Actions: N, E, S, W
- States: Agent's location

David Silver, Reinforcement Learning, PPT

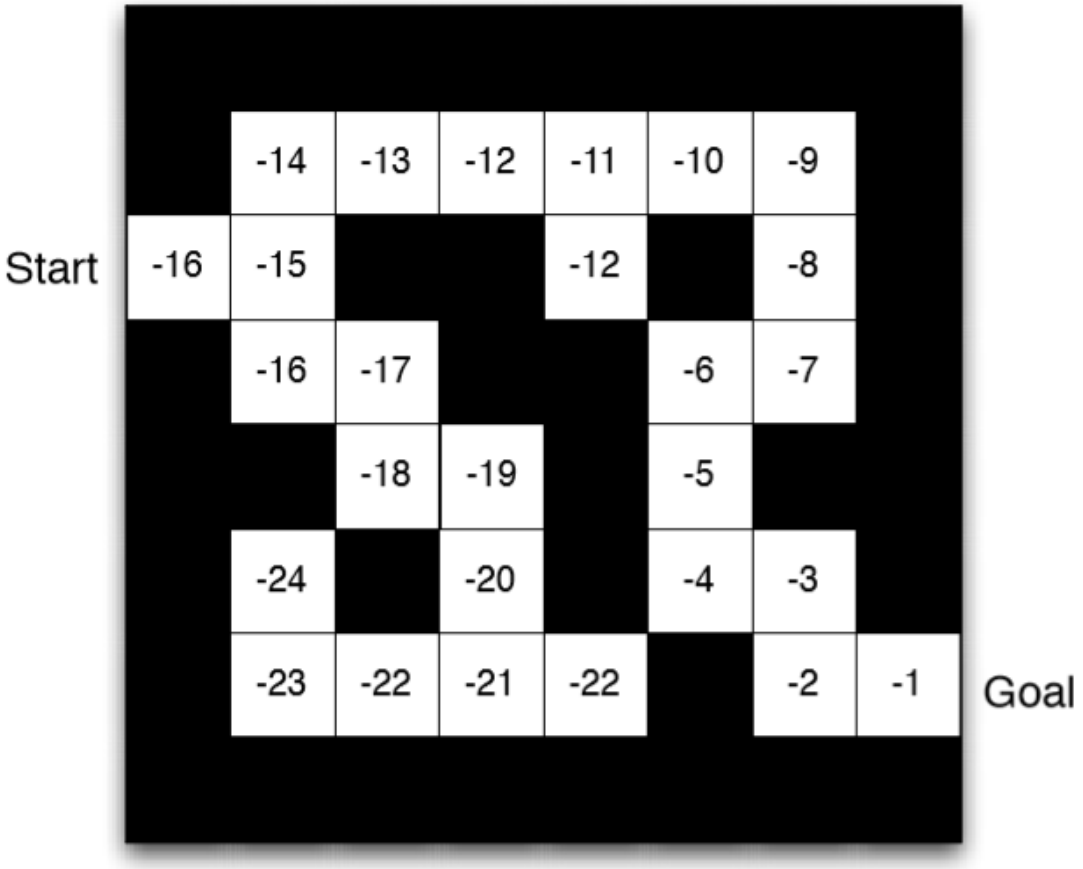
Maze Example: Policy



- Arrows represent policy $\pi(s)$ for each state s

David Silver, Reinforcement Learning, PPT

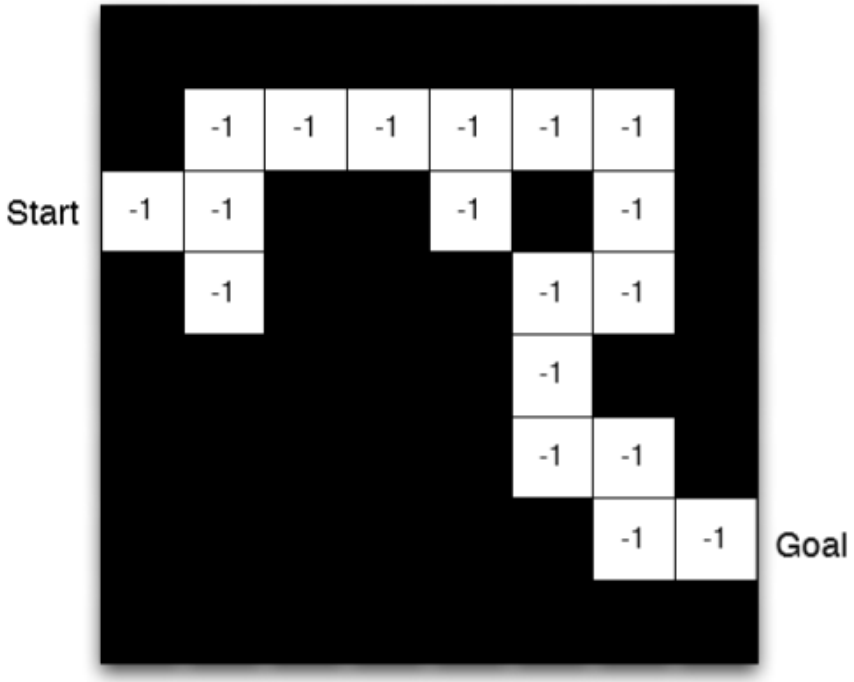
Maze Example: Value Function



- Numbers represent value $v_\pi(s)$ of each state s

David Silver, Reinforcement Learning, PPT

Maze Example: Model



- Agent may have an internal model of the environment
- Dynamics: how actions change the state
- Rewards: how much reward from each state
- The model may be imperfect

- Grid layout represents transition model $\mathcal{P}^a_{ss'}$
- Numbers represent immediate reward $\mathcal{R}^a_s$ from each state s (same for all a)

David Silver, Reinforcement Learning, PPT

Categorizing RL agents (1)

- Value Based
  - No Policy (Implicit)
  - Value Function
- Policy Based
  - Policy
  - No Value Function
- Actor Critic
  - Policy
  - Value Function

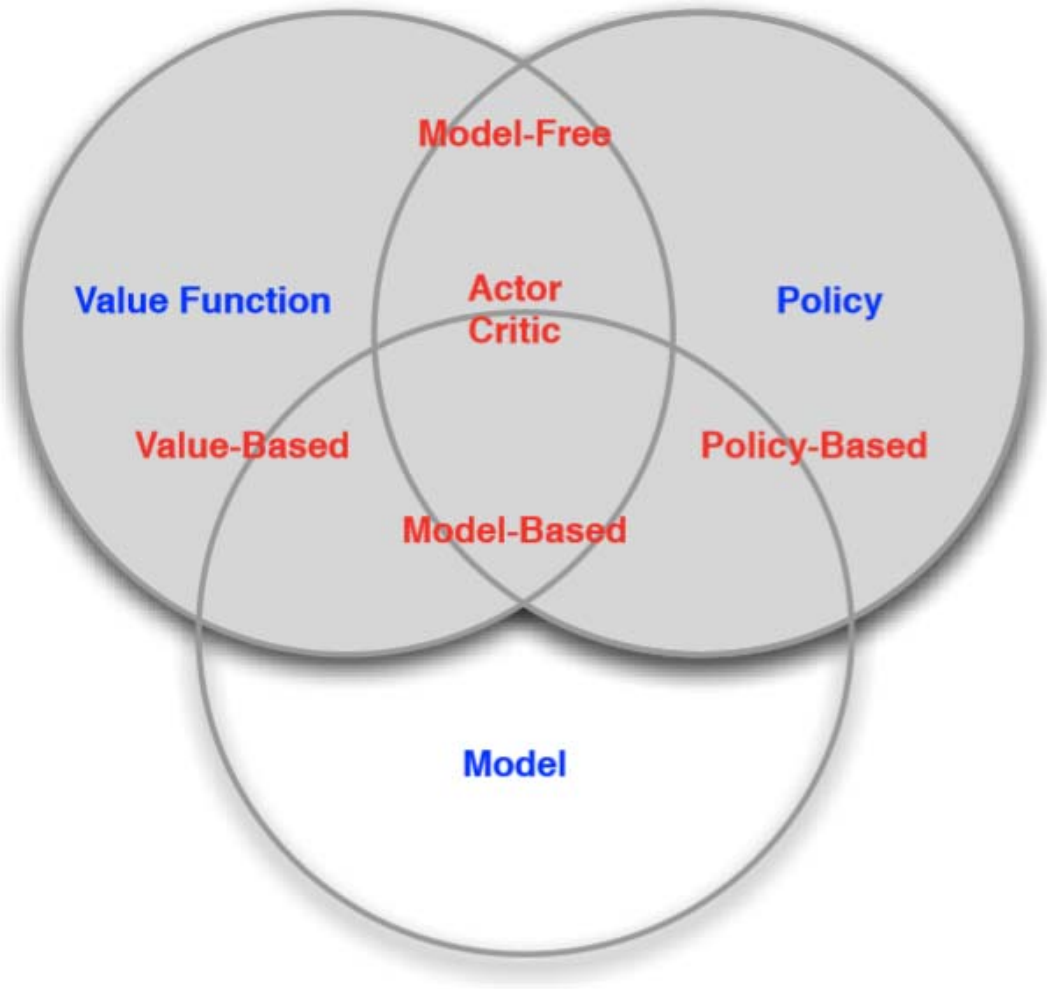
David Silver, Reinforcement Learning, PPT

Categorizing RL agents (2)

- Model Free
  - Policy and/or Value Function
  - No Model
- Model Based
  - Policy and/or Value Function
  - Model

David Silver, Reinforcement Learning, PPT

RL Agent Taxonomy



David Silver, Reinforcement Learning, PPT

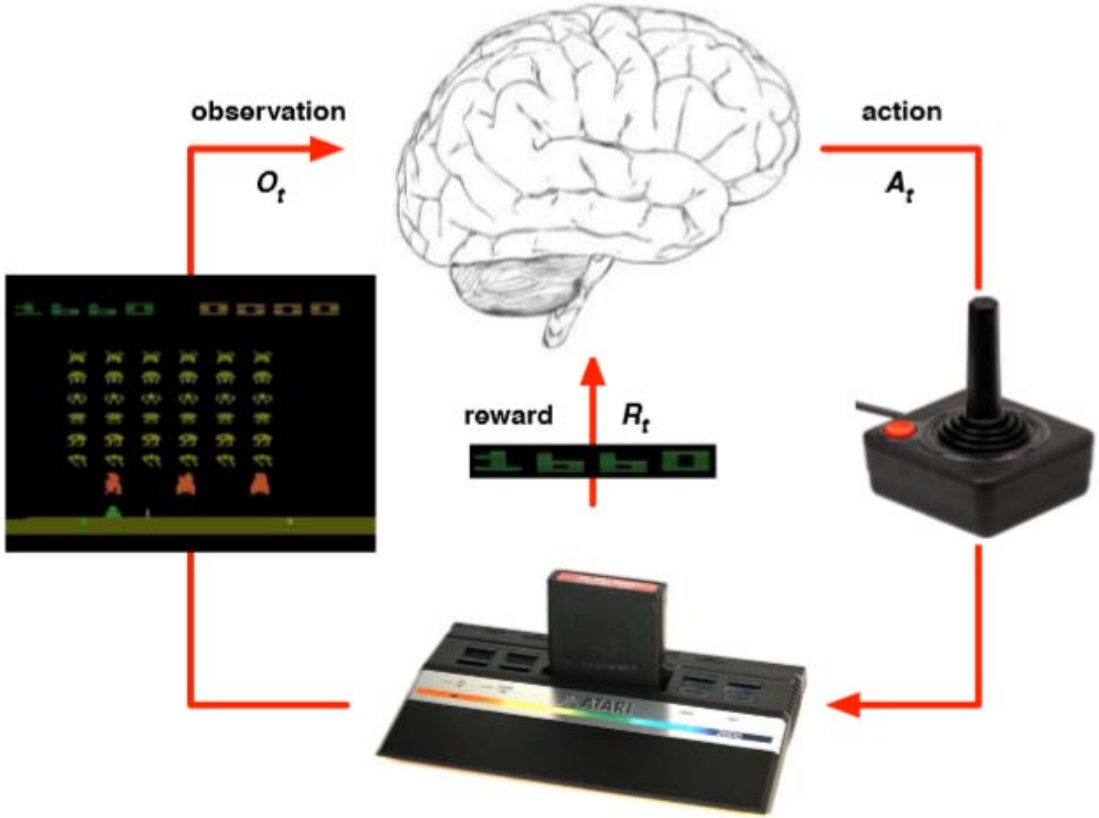
Learning and Planning

Two fundamental problems in sequential decision making

- Reinforcement Learning:
  - The environment is initially unknown
  - The agent interacts with the environment
  - The agent improves its policy
- Planning:
  - A model of the environment is known
  - The agent performs computations with its model (without any external interaction)
  - The agent improves its policy
  - a.k.a. deliberation, reasoning, introspection, pondering, thought, search

David Silver, Reinforcement Learning, PPT

Atari Example: Reinforcement Learning

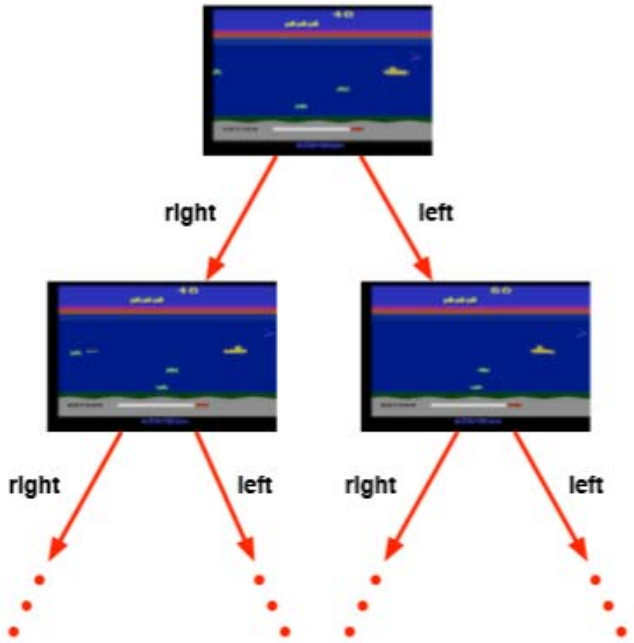


- Rules of the game are unknown
- Learn directly from interactive game-play
- Pick actions on joystick, see pixels and scores

David Silver, Reinforcement Learning, PPT

Atari Example: Planning

- Rules of the game are known
- Can query emulator
  - perfect model inside agent's brain
- If I take action a from state s:
  - what would the next state be?
  - what would the score be?
- Plan ahead to find optimal policy
  - e.g. tree search



David Silver, Reinforcement Learning, PPT

Exploration and Exploitation (1)

- Reinforcement learning is like trial-and-error learning
- The agent should discover a good policy
- From its experiences of the environment
- Without losing too much reward along the way

David Silver, Reinforcement Learning, PPT

Exploration and Exploitation (2)

- *Exploration* finds more information about the environment
- *Exploitation* exploits known information to maximise reward
- It is usually important to explore as well as exploit

David Silver, Reinforcement Learning, PPT

Examples

- Restaurant Selection
  - Exploitation Go to your favourite restaurant
  - Exploration Try a new restaurant
- Online Banner Advertisements
  - Exploitation Show the most successful advert
  - Exploration Show a different advert
- Oil Drilling
  - Exploitation Drill at the best known location
  - Exploration Drill at a new location
- Game Playing
  - Exploitation Play the move you believe is best
  - Exploration Play an experimental move

David Silver, Reinforcement Learning, PPT

Prediction and Control

- Prediction: evaluate the future
  - Given a policy
- Control: optimise the future
  - Find the best policy

David Silver, Reinforcement Learning, PPT

강화학습 응용 사례

- Atari game - Deepmind
- AlphaGo – Deepmind
- 로보트 – Arm 조작, Walking, Running, …
- 자율주행
- …

Deep Reinforce Learning

http://www.hankookilbo.com/News/Read/201603121722964715

- 알파고 무너뜨린 이세돌의 '신의 한 수' 78수 / YTN https://www.youtube.com/watch?v=5MEorOEsI1I
- [이세돌 신의한수 78수의 비밀-요약] 알파고VS이세돌 4국 김성룡 직접 본 이야기6 https://www.youtube.com/watch?v=LgDHz5xJodI

Question?