DS 598
Introduction to Reinforcement Learning

Xuezhou Zhang

Additional Information

- Office Hour: Tue/Thu 2:00-3:00 PM (right after class), CDS 14th Floor
- TF: Gaurav Koley
- Course Website: zhangxz1123.github.io/DS598.html
- Blackboard only used for HW turn-in.

Reading Materials

- Reinforcement Learning: Theory & Algorithms
- https://rltheorybook.github.io/
- This is an advanced RL textbook, so we will pick specific subsections for you to read.

This course introduces Reinforcement Learning (RL)

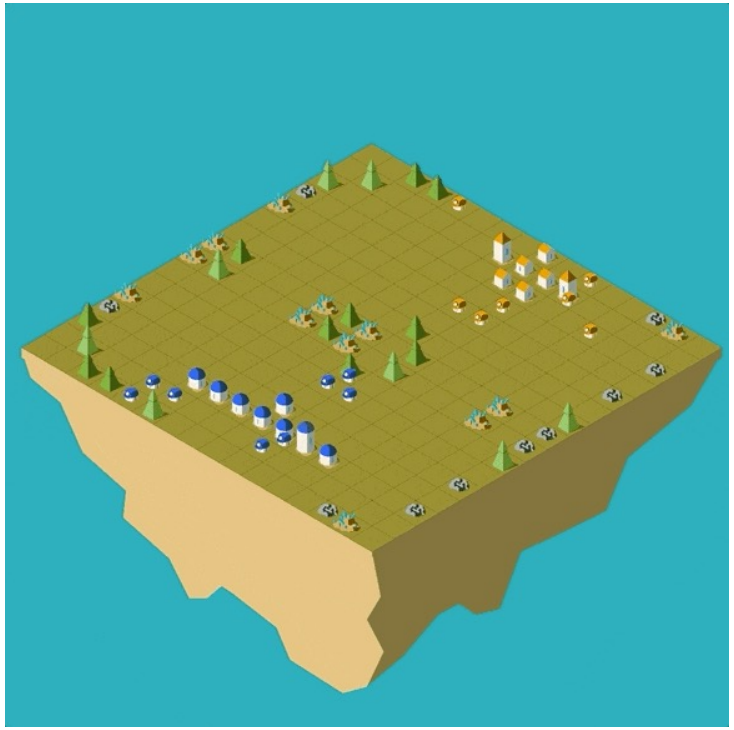
I. Markov Decision Process (MDP): Dynamic Programming & planning.

II. Model-based, value-based, policy based learning paradigms.

III. Modern challenges in RL.

However, the most fun part..

- Game AI Competition!
- Details will come in the following weeks.



Logistics

- Written Assignment: 40%
- Competition: 50%-105%

Written Assignment (40%)

- 4 assignments: 10% each

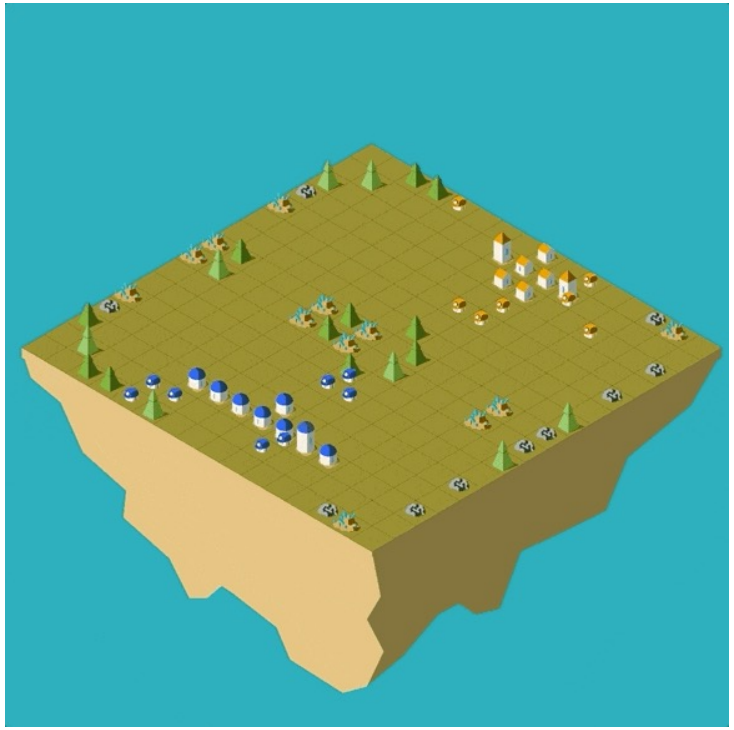
- Type your solution using LaTeX.

- LaTeX tutorial in week 1 Discussion Session.

Competition (50%-105%)

- Form a team of $\leq$ 3 people by Jan 27th [link].

- Design Sharing Presentation: 10%
- Beating the Midterm Champion: 10%
- Final Report: 30%
- Midterm tournament: 15%(1) /10%(2-3) /5%(4-8)
- Final tournament: 30%(1) /20%(2) /15%(3) /10%(4-8)
- (Bonus) Least domain knowledge: 10%(1)



Prerequisites

- Linear algebra & probability
- Programming in Python
- ML background*

What is machine learning?

- Given a dataset $\{x_1, x_2, \ldots, x_n\} \sim P$
- Find patterns in it that applies to future samples from P.

- Unsupervised Learning: pattern = $\hat{P}$.
- Supervised Learning: pattern = $p(y|\mathrm{x}_{-y})$.

ML vs. Reinforcement Learning

- ML:
  - Make predictions.
  - Rely on existing data.

vs.

- RL:
  - Perform actions.
  - Collect its own data.

ML vs. Reinforcement Learning

- Example: Trading in the Stock Market.

- SL: What are the stock prices tomorrow?
- RL: How many shares of each stock should I purchase?

What's different in RL?

1. Collect your own data.

2. Actions have consequences. Future observations are determined by past actions.

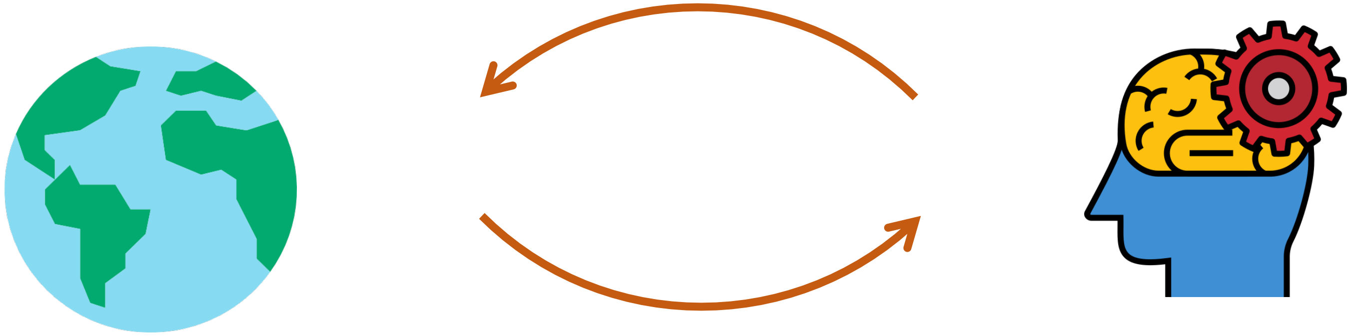
3. To solve a task, we often need to perform a sequence of actions.

What differentiate good vs. bad decisions?

- In SL, you fit to labels, e.g. cat vs. dog.
- In RL, you maximize a utility function, e.g. $ profit/day.

The Mathematical framework:
Markov Decision Process (MDP)

Perform action: $\boldsymbol{a_h \sim \pi(\cdot\, | s_h)}$



Environment

RL Agent

Receive Reward: $\boldsymbol{r_h \sim r(s_h, a_h)}$

Observe Next state: $\boldsymbol{s_{h+1} \sim P(\cdot\, | s_h, a_h)}$

- Markovian Transition: $\boldsymbol{s_{h+1}}$ only depends on $\boldsymbol{s_h, a_h}$.

Infinite Horizon Discounted MDP

- MDP $\mathcal{M} = \{S, A, P, r, \gamma\}$
  - S is the state space.
  - A is the action space.
  - $P: S \times A \to \Delta(S)$ is the transition probability function.
  - $r: S \times A \to [0,1]$ is the reward function.
  - $\gamma \in [0,1)$ is the discounting factor.

- A policy is defined as $\pi: S \to \Delta(A)$.

How good is a policy π?

- Value function

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) | s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot|s_h, a_h)\right]$$

- Q function

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) | (s_0, a_0) = (s, a), a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot|s_h, a_h)\right]$$

Bellman Equation

$$
\begin{aligned}
V^\pi(s,a) &= \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h,a_h) | s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot|s_h,a_h)\right] \\
&= \mathbb{E}[r(s,\pi(s))] + \mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^h r(s_h,a_h) | s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot|s_h,a_h)\right] \\
&= \mathbb{E}[r(s,\pi(s))] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h,a_h) | s_0 = s', a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot|s_h,a_h)\right] \\
&= \mathbb{E}[r(s,\pi(s))] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s')
\end{aligned}
$$

Bellman Equation: $V^\pi(s,a) = \mathbb{E}[r(s,\pi(s))] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s')$

Today we covered

- SL vs. RL

- Infinite horizon discounted MDP

- Value function and Q function

- Bellman Equation