University of Paris-Saclay - UFR Sciences

Lab 1 - Reinforcement Learning

Diyun LU

17 November 2022

1 Warming Up Questions

1.1

What is the goal of a reinforcement learning algorithm? In this framework what is the goal of an agent? How can you formalize it mathematically?

The goal of a reinforcement learning algorithm is to take actions based on the environment to maximize the cumulative reward.

The goal of an agent is to learn an optimal policy that can maximize the reward function or other reinforcement signals provided by users and accumulated from immediate rewards.

We can formalize it mathematically as follows:

$$\mathbb{E}_{s_0,\pi}[\sum_{t=1}^{\infty}(\gamma^t r(s_t))] \tag{1}$$

1.2

What make reinforcement learning different from supervised learning ? What makes reinforcement learning hard?

In supervised learning, the training data has the labels with it so the model is trained with the ground truths. However, in reinforcement learning, there is no answer, the reinforcement agent needs to decide what to do to perform the tasks.

1.3

Give the formal definition of a Markov Decision Process (MDP). Give the meaning of it components.

MDP is discrete time stochastic control process. It provides a mathematical framework for modeling decisions where the outcomes are partly random and partly under the control of the decision maker.

$$MDP : p(s', r|a, s) = \{T, \pi, A, S, R\} \tag{2}$$

1. A set of states space S.
2. A set of actions space A.
3. A reward function $R(s)$.
4. A policy the solution of Markov Decision Process π :S $\rightarrow A$.
5. A transition model $p(s^{'}|a, s)$

1.4

Why do we call it Markov?

Because we use Markov decision process and the transition defines the Markov chain.

1.5

What does the discount factor γ mean or represent?

The discount factor can be used to decide how much the reinforcement learning agent cares about the distant future relative to the near reward.

1.6

Can you give examples of setup with non deterministic transitions?

For example, the machines in the casino. Anything that are not certain with randomness or partially uncertain can be non deterministic transitions.

1.7

What is the difference of an episodic setup and a continuous one ? Give an example of each.

Episodic setup are tasks that come to an end. For example: tic-tac-toe or pac-man; while continuous ones are tasks that never end. For example, tuning a heating system.

1.8

What does V function means or represents ?

The V function means what the expected overall value of a certain state under a policy is.

2 A simple Maze

2.1

What is the space of states?

All white spaces in the maze.

2.2

What is the space of actions?

Up, down, left, right.

2.3

Why do we give a negative reward when the agent does not reach the goal ?

It is a punishment, so we can reduce the chances of specific behavior occurring again. Thus, we can reach the goal faster.

2.4

Give the Bellman equation followed by V

$$V_\pi(s) = \mathbb{E}[r(s)] + \gamma \sum_{s'} p(s, \pi(s), s^{'}) V(s') \tag{3}$$

2.5

Deduce what is V for each state for the optimal policy (the one where the agent takes the best decision at each step)

$$\begin{aligned} V(26) &= 0 \\ V(25) &= -1 \\ V(24) &= -1 - \gamma^1 \\ V(23) &= -1 - \gamma^1 - \gamma^2 \\ &... \\ V(S) &= \sum_{i=0}^{m-1} -(\gamma)^i \end{aligned}$$

3 Crossing a river

3.1

Compute the value function V for a constant policy.

$$\begin{aligned} V(100) &= 100 \\ V(99) &= 100\gamma \\ V(98) &= 100\gamma\gamma \\ &... \\ V(S_i) &= 100\gamma^{N-i} \end{aligned}$$

3.2

Same question with p(si , right, si+1) = .9 ; p(s, right, s) = .1.

$$\begin{aligned} V(S_i) &= \gamma[0.9(V(S_{i+1}) + 0.1[V(S_{i+1}))] \\ V(S_i) &= \frac{0.9\gamma}{1 - 0.1\gamma} V(S_{i+1}) \end{aligned}$$

It is similar with what we did above, we just need to replace γ to $\frac{0.9\gamma}{1 - 0.1\gamma}$:

$$V(S_i) = 100(\frac{0.9\gamma}{1 - 0.1\gamma})^{N-i}$$

4 Computer Store Management

4.1

Is it an episodic or a continuous problem ? Give a justification?

It is a continuous problem. Because it is a continuous task for every week that does not come to an end.

4.2

What would be the state space ? Is it discrete or continuous ?

The current storage can be the state space, it is continuous.

4.3

What would be the action space ? Is it discrete or continuous ? Is it stochastic or deterministic ?

The actions would be how many units we buy on the Monday morning each week, it is continuous and it is stochastic.

4.4

Hard question As the owner, you want to maximise your balance each week. Write the reward as a function of the actions and the state for each week t. rt(at, st)

The reward will be the sales profit minus the maintenance cost and ordering cost. Also, we need to make sure that the ordering units are less than the maximal capacity of the storehouse.

$$r_t(a_t, s_t) = p \times min(D_t, a_t) - h \times S(t) - a_t \times c - min(c_0, c_0 \times a_t)$$
$$a_t < M$$

4.5

Write the transition model to finish the design of the MDP.

The transition will be the from last state to current state, which is the former storage to current storage. So we used the last storage minus the sold units plus the ordering ones.

$$S_t + 1 = S_t - D_t + a_t$$

5 Bonus Question - Pole balancing

5.1

Suggest a reward structure which would likely induce the desired behavior.

Our goal is to maintain the stick vertically as long as possible. The longer we keep the pole standing, the more rewards we get. Our actions would be moving left or right. We can design the state to be the angle between the stick and the vertical line, and we can set the threshold to be 15 degrees. Thus, if the new angle after the action is smaller than 15 degrees, which means it maintains standing, then we give it one point as reward; if it is larger than 15 degrees, then the reward will be 0.