CMPUT 365: Markov Decision Processes

Rupam Mahmood

Jan 17, 2022

Admin

Due dates C1M2:

- Practice quizzes: Tues Jan 18
- Peer-graded assignment: Thu Jan 20
- Peer-review: Sat Jan 22

Assignment 1:

- Will be released soon (This Sat?)
- One week time
- Two worksheet(-like) questions

Midterm:

- Based on worksheet questions, book reading, and lectures

Bandit review

- Policy $\pi(a) \doteq Pr(A=a)$.

$$\pi(a) \doteq Pr\left(A=a \mid \pi\right) \qquad Pr\{\ \}$$
$$\pi_1(\cdot),\ \pi_1 = Pr\left(A=a \mid \pi_1\right)$$
$$\pi_2 = Pr\left(A=a \mid \pi_2\right)$$

- Action value $q_*(a) = E[R \mid A=a] = \sum_r r\, P(R=r \mid A=a)$

  The pro of observing a reward of choosing an action.

  - Does not depend on the policy in bandits
- Value ⇒ goal: maximize the expect rewards $q_*(a)$.
  - Depends on the policy even in bandits

$$E_\pi[R] = \sum_r r\, P(R=r \mid \pi)$$
$$= \sum_r r \sum_a P(R=r \mid A=a)\, P(A=a \mid \pi)$$

→ under the control of the agent.

$P(R=r, A=a)$

$P(R=r) = \sum_a P(R=r, A=a)$.

$$= \sum_a \underbrace{P(A=a \mid \pi)}_{\pi(a)} \underbrace{\sum_r r\, P(R=r \mid A=a)}_{q_*(a)}$$
$$= \sum_a \pi(a) \cdot q_*(a)$$

(∴可以说 $E_{\pi_1}[R] < E_{\pi_2}[R]$)

Aaron Skiba to Everyone 1:23 PM

A Would that have to be E[R|A=a] ?

Wouldn't E[R] require us to iterate through all actions? ⇒ Yes !!

- Contextual bandit

$\pi(a \mid s) = P(A=a \mid \underbrace{X=x}_{\text{the particular context.}})$ (under this particular context, what's the pro of the agent to chose action A?)

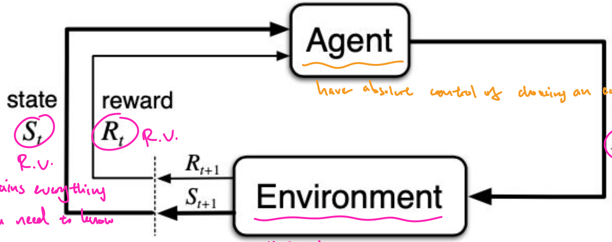
(其实就是 state).

$$E[R \mid A=a, X=x]$$
$$= q_*(x, a)$$

Coursera: video 1 - Intro to MDPs

- The dynamics of an MDP

The agent interacts with the environment.



have absolute control of choosing an action. It can't control the distribution of reward and state even if the agent knows about it.

S_t R.V. contains everything you need to know

R_t R.V.

A_t R.V. has a sequence to change the state of the environment.

Environment: maintain states that can be influence by your actions!!

The boundary represents the limit of the agent's absolute control

not of its knowledge

time steps, $t = 0, 1, 2, 3, \ldots$

Sequence or a trajectory

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$

Dynamics of the MDP

The next state and reward depends on the previous state and action.

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

dynamics function $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

David Wang to Everyone 1:48 PM

If I may. In layman's terms, what the dynamics means is the probability of moving to a specific following state and getting a specific reward, given the state you were in previously and the action you chose. Yes?

Yes!!

Intro to MDPs (cont’d)

probability of setting s' as the next state given 当前状态s与当前action a.

state-transition probabilities a three-argument function $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$)

The boundary between the agent and the environment是可变的.

$$p(s' \mid s, a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r \mid s, a)$$

<- according to which law? => marginal probabilities $P(R=r) = \sum_a P(R=r, A=a)$

doesn't depend on the policy.

only for the next state base on the current state and action.

different from $P(s' \mid s, a)$, 这个是 the likelihood of observing next state s' and reward given the current state and the action.

Parham Golestaneh Talaei to Everyone 1:51 PM

is the first line the expected reward of choosing an action leading to state s'

expected rewards for state–action pairs

define a new object

$$r(s, a) \doteq \mathbb{E}[R_t \mid \underbrace{S_{t-1} = s}_{\text{previous state}}, \underbrace{A_{t-1} = a}_{\text{previous action}}] = \boxed{\sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)}$$

<- according to which law?

=: mathematically equality

≐: this two expression equality

A∩B: A与B的intersection

$= \sum_r r P(R_t = r \mid S_{t-1} = s, A_{t-1} = a)$ 大写

$= \sum_r r \sum_{s'} p(s', r \mid s, a)$ 小写

expected rewards for state–action–next-state triples

$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

<- according to which law?

inputs (are given)

$= \sum_r r P(R_t = r \mid S_{t-1} = s, A_{t-1} = a, S_t = s') =$

$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$ The joint probability of A and B (A B同时发生)

probability of A given B.

Coursera: video 2 - Examples of an MDP

- Task: the goal of the robot is to pick and place object
- State: latest readings of joint angles and velocities
- Action: the amount of voltage applied to each vector
- Reward: -c every time step, where c > 0 => every timestep you get a negative reward.

  when the reward is all negative, we force the robot to terminate the process as soon as possible.
- Termination: when an object is placed successfully

Coursera: video 3 - The goal of RL
Video 5: Continuing tasks

Return in an episodic task, where an episode ends in a terminal state and T is the terminal step at that episode

$$\underbrace{G_t}_{\text{reward}} \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

=> goal: maximize the sum of future rewards.
(everything matters).

Return in a continuing task 要有 discount rate γ

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1},$$

goes to infinite.

where γ is a parameter, $0 \leq \gamma \leq 1$, called the *discount rate*.

Return expressed recursively

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots\right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

- Short-sighted agent vs. Far-sighted agent

γ 接近 0 | γ 接近 1.

Cameron Jen to Everyone 1:49 PM

CJ So it's best for gamma to be 1 for continuing tasks to ensure we are optimizing all steps ?

David Wang to Everyone 1:49 PM

DW No. If it is one, then the thing doesn't converge. ...

Coursera: video 4 - The reward hypothesis

That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).

Coursera: video 6 - Examples of Episodic and Continuing Tasks

- Examples?

Notations

r.v. X, Y, A, R, S

outcomes x, y, a, r, s

大写 P → probability distribution

$P(\text{event})$ output $\in [0,1]$ output一个0~1之间的数。(包括0, 1)

take an event as input

大写 E → expectation

input: r.v. output $\in \mathbb{R}$ output一个real number

take an random variable as input

$P(A=a \mid S=s)$ conditional probability: take 2 events as input 并 output a given s的发生的概率.

P和q是完全相同的一个东西，代表MDP，但是，是2个不同的MDP

$MDP_1 \rightarrow$ 小写 $p(s' \mid s, a) = P(S_{t+1}=s' \mid S_t=s, A_t=a)$ the probability of next state given the current state and current action.

take outcomes as input.

$MDP_2 \rightarrow q(s' \mid s, a)$

For action policy: $\pi(a \mid s)$ π代表一种policy.

$\mu(a \mid s)$ μ代表另一种policy.

Notations (2)

$p(r \mid s, a)$ ⇒ probability of this current reward given current state and action

$= \sum_{s'} p(s', r \mid s, a)$ by law of total probabilities.

$p(s', r \mid s, a)$ ⇒ probability of next state and this current reward given current state and action

Expected reward for s, a state-action pair

$$r(s,a) = \sum_r r \, p(r \mid s, a)$$

⇒ by law of total probabilities.

s与a not indicated, ∴它们用作为 r(s,a)的 input.

在不同的 policy 下，这个值会有所不同.

$p_\pi(s' \mid s) = \sum_a p_\pi(s', a \mid s)$

probability of next state given the current state.

$= \sum_a p(s' \mid s, a) \, p(a \mid s)$ an action probability, which comes up to be policy.

usually denote by π. (policy)

↳ depend on the policy!! if the policy is different, then this value will be different!!

所以要以π作为下标，写为:

$$\begin{cases} p_\pi(s' \mid s) = \sum_a p_\pi(s', a \mid s) \\ \quad = \sum_a p(s' \mid s, a) \, \pi(a \mid s) \end{cases}$$

$p(r \mid s)$ is this independent of the policy π? ⇒ Yes.

probability of taking a particular reward r given current state.

bellman-eq.mp4

Bellman equation for v_π

Bellman function: express the value function of the given state in terms of the value of next states.

the current return — the current state

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

depend on policy π.

LOTE
LOTUS
MP — Markov property: everything happen just depends on the state given.
LE

use law of total expectations

$$= E_\pi\big[\,E_\pi[G_t \mid S_t = s, A_t] \mid S_t = s\,\big]$$

⇒ by law $E[E[X|Y]] = E[X]$.

the probability of this R.V. takes value.

需要 next state reward 作为 condition.
∴ 要 bring in new R.V.

LOTUS

$$= \sum_a E_\pi[G_t \mid S_t = s, A_t = a]\, P(A_t = a \mid S_t = s)$$

def of action value. ∴ 可被写为 $q_\pi(s,a)$

也写为 π, 因为 $P(A_t=a \mid S_t=s)$ 是 the policy probability. $\pi(a|s)$

$$= \sum_a \pi(a|s)\, q_\pi(s,a)$$

the relation between state value and action value.

引入的新变量

$$q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$$

LOTE

$$= E_\pi\big[\,E_\pi[G_t \mid S_t = s, A_t = a, R_{t+1}, S_{t+1}] \mid S_t = s, A_t = a\,\big]$$

by law $E[E[X|Y]] = E[X]$.

LOTUS

$$= \sum_{s',r} E_\pi[G_t \mid S_t = s, A_t = a, R_{t+1} = r, S_{t+1} = s']\; p(s', r \mid s, a)$$

state-action pair.

write the return recursively

probability of this R.V. takes value.

next state — next reward

probability of $S_{t+1} = s'$ and $R_{t+1} = r'$

$$= \sum_{s',r} p(s', r \mid s, a)\, E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a, R_{t+1} = r, S_{t+1} = s']$$

∴ R_{t+1} already given.
可直接写为 r.

1:23 / 19:37

bellman-eq.mp4

8:13 PM Sun Jan 31

Bandits Wed

Vf

use linearity of expectations.

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, \boxed{\mathbb{E}_{\pi}[G_{t+1} \mid S_{t+1} = s']}\right]$$

→ state value of the next state 即 $V_\pi(s')$

by Markov property. 所有发生的事情都只与 state 有关!!
∴只写 state 即可.

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, v_\pi(s')\right]$$

注!! don't write policy here!! 因为 the transition dynamic of the environment don't depend on policy!!
所以不需写 p_π!!

take s as an input, ∴ the whole expression should be the function of s (右式除了s都被消掉!!)

所以: $$v_\pi(s) = \sum_a \pi(a|s) \boxed{\sum_{s',r} p(s',r|s,a)}\left[r + \gamma\, v_\pi(s')\right]$$

the action depends on the policy

the next state value also depends on the policy.

注: 当该 value function is depend on policy π, 那么 policy π 也 should appear at the R.H.S. as well!!

S为唯一变量在左边!!
(即 s is given!!)

右边除了π和s以外的所有 extra variables 都 should be bounded by the summation!! 即 $\sum_a$ 与 $\sum_{s',r}$!!

- 如在 $\sum_a \pi(a|s)$ 中, a (variable) 没有作为 input 在该等式 L.H.S 出现, 所以必须 bounded by $\sum_a$!!
- 在 $\sum_{s',r} p(s',r|s,a)$ 中, s', r (variables) 没有作为 input 在该等式 L.H.S 出现. 所以必须 bounded by $\sum_{s',r}$!!

14:26 / 19:37