به نام خدا

یادگیری ژرف

کوییز شماره ۶

سروش گوران

نیمسال اول ۱۴۰۰ – ۱۴۰۱

سوال ۱

تعریف‌ها:

فرآیند تصمیم گیری مارکوف

فرآیند تصمیم‌گیری مارکوف توسعه‌ای بر فرآیند پاداش مارکوف است و شامل تصمیماتی است که یک agent باید اتخاذ کند. تمام حالات موجود در محیط، مارکوف هستند.

Markov Decision Process is a tuple $\langle S, A, P, R, \gamma \rangle$
: S is a finite set of states
: A is a finite set of actions
: P is a state transition probability matrix,
$P^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
: R is a reward function, $R^a_s = [R_{t+1} | S_t = s, A_t = a]$
: γ is a discount factor, $\gamma \in [0, 1]$

سیاست‌ها

یک سیاست[۱] π، توزیعی بر روی action های حالات داده شده است که به طور کامل رفتار یک عامل را تعریف می‌کند. سیاست های MDP به وضعیت فعلی بستگی دارد نه تاریخچه. سیاست، نگاشت از یک حالت به حالت بعدی را معین می‌مند. اگر در حالت s باشیم، احتمال انجام هر عمل را از آن حالت مشخص می‌کند.

$$\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$$

تابع Policy

تابع ارزش برای MDP

ما action هایی را انجام می‌دهیم و بسته به اینکه چگونه عمل می‌کنیم، انتظارات متفاوتی وجود دارد.

تابع حالت-مقدار $v_\pi(s)$ یک MDP، مقداربازگشتی مورد انتظاری است که از حالت s شروع می شود و از خط مشی π پیروی می‌کند.

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$

تابع حالت-مقدار به ما می‌گوید که با پیروی از خط مشی π، بودن در حالت s چقدر خوب است.

تابع Action-Value (مقدار-عمل) $q_\pi(s, a)$ مقدار بازگشتی مورد انتظاری است که از حالت s شروع می‌شود، عمل a را انجام می‌دهد و خط مشی π را دنبال می‌کند.

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

---

[۱] policy

تابع Action-Value به ما می‌گوید که انجام یک عمل خاص از یک حالت خاص چقدر خوب است و به ما ایده می‌دهد که چه action هایی باید در حالت‌ها انجام دهیم.

معادله انتظار بلمن[۲]

توابع مقدار را می‌توان در قالب یک معادله انتظار بلمن به صورت زیر نوشت:

تابع ارزش-حالت:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] \\ &= \sum_{a\in A} \pi(a|s) q_\pi(s,a) \\ &= \sum_{a\in A} \pi(a|s)(R_s^a + \gamma \sum_{s'\in S} P_{ss'}^a v_\pi(s')) \end{aligned}$$

تابع ارزش عمل:

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \\ &= R_s^a + \gamma \sum_{s'\in S} P_{ss'}^a v_\pi(s') \\ &= R_s^a + \gamma \sum_{s'\in S} P_{ss'}^a \sum_{a\in A} \pi(a'|s') q_\pi(s',a') \end{aligned}$$

تابع ارزش بهینه

هدف اصلی در یادگیری تقویتی، یافتن خط مشی بهینه است که بازدهی ما را به حداکثر برساند.

تابع حالت-مقدار بهینه $v^*(s)$ تابع مقدار حداکثر در همه سیاست ها است و حداکثر پاداش ممکنی را که می‌توان از سیستم دریافت کرد به ما می‌گوید.

$$v_*(s) = \max_\pi v_\pi(s)$$

تابع مقدار عمل بهینه $q^*(s,a)$ حداکثر تابع مقدار عمل در همه سیاست ها است. این به ما می‌گوید حداکثر پاداش ممکنی که می‌توانید از حالت s و با انجام عمل a از سیستم دریافت کنید، چقدر است.

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

اگر $q*$ را بلد باشید، می‌دانید که چه عمل درستی را باید انجام دهید. $q*(s,a)$ نشان می دهد که برای رفتار بهینه چه اعمالی باید انجام شود.

پیدا کردن یک خط مشی بهینه

---

[۲] Bellman Expectation Equation

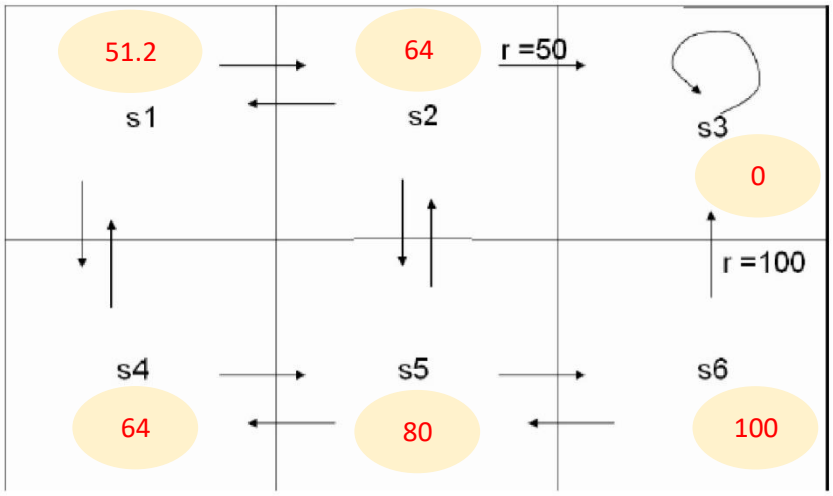
یک خط مشی بهینه را می توان با بیشینه کردن روی $q*(s,a)$ پیدا کرد:

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \text{argmax}_{a \in A} q_*(s,a) \\ 0 & \text{otherwise} \end{cases}$$

الف)

حالت S3 حالت نهایی است و از این حالت به جای دیگری نمی رویم. بنابراین $V^*(s3)$ صفر است.

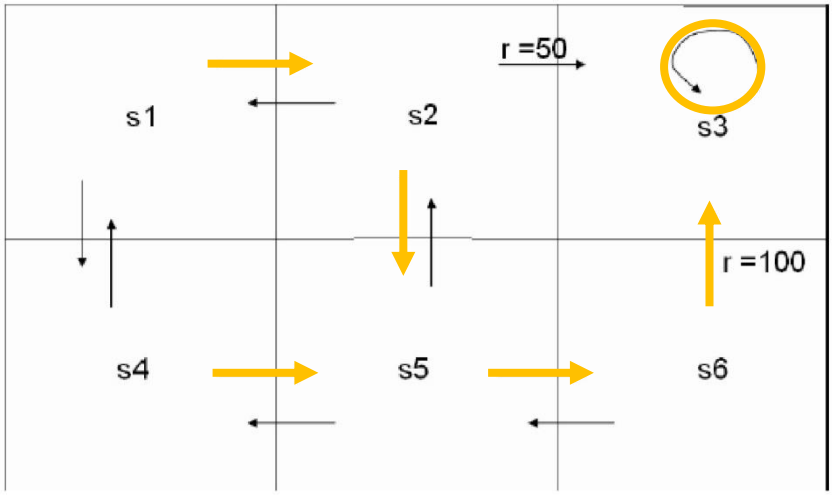
$$V^*(s3) = 0$$
$$V^*(s6) = 100 + 0.8 * V^*(s3) = 100$$
$$V^*(s5) = 0 + 0.8 * V^*(s6) = 80$$
$$V^*(s4) = 0 + 0.8 * V^*(s5) = 64$$
$$V^*(s2) = \max(0 + 0.8 * V^*(s5), 50 + 0.8 * V^*(s3)) = 64$$
$$V^*(s1) = \max(0 + 0.8 * V^*(s4), 0 + 0.8 * V^*(s2)) = 51.2$$



ب)

به این ترتیب می‌توانیم مسیرهای سیاست بهینه[۳] را روی شکل مشخص کنیم.

[۳] optimal policy



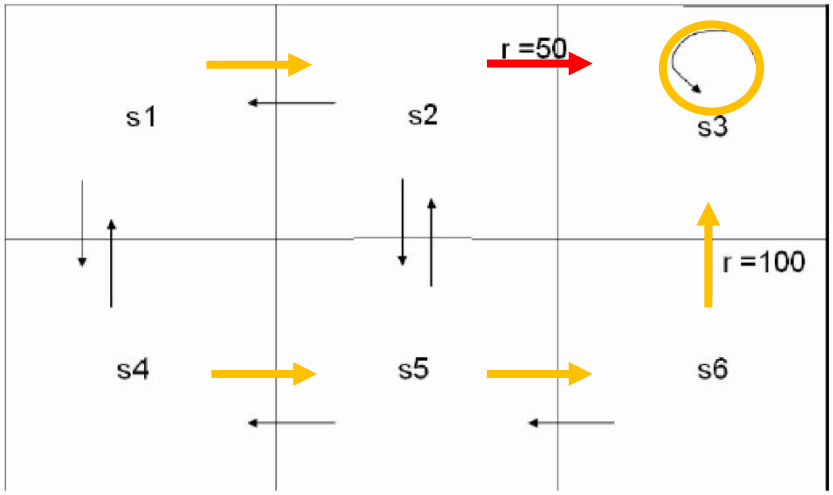
ج)

مقدار $\gamma = 0.7$ در نظر می‌گیریم:

$$V^*(s3) = 0$$
$$V^*(s6) = 100 + 0.7 * V^*(s3) = 100$$
$$V^*(s5) = 0 + 0.7 * V^*(s6) = 70$$
$$V^*(s4) = 0 + 0.7 * V^*(s5) = 49$$
$$V^*(s2) = \max(0 + 0.7 * V^*(s5), 50 + 0.7 * V^*(s3)) = 50$$
$$V^*(s1) = \max(0 + 0.7 * V^*(s4), 0 + 0.7 * V^*(s2)) = 35$$

در این حالت فقط یک *action* تغییر می کند. یعنی $\pi(S2) = S3$

اگر $\gamma \leq 0.6$ باشد، دو *action* تغییر خواهد کرد و اگر $\gamma > 0.8$ آنگاه هیچ *action* ای تغییر نخواهد کرد.



د)

اگر هر یک از *reward* ها را در یک ضریب ثابت مثل k ضرب کنیم، V^* نیز k برابر شده ولی π^* بدون تغییر باقی می‌ماند.