Введение в обучение с подкреплением

Тема 4: Марковский процесс принятия решений

Лектор: Кривошеин А.В.

Напоминание о МПВ

Марковский процесс вознаграждений — это марковский процесс, в котором агент наблюдает переходы между состояниями среды и получает вознаграждения, не совершая при этом действий.

Траектория бесконечного МПВ является реализацией последовательности случайных величин

$$S_0, R_1, S_1, R_2, \ \dots$$

Функция ценности состояния $v(s)$ определяется как ожидаемый доход по всем возможным траекториям при старте из состояния s:

$$v(s) := \mathbb{E}[G_t \mid S_t = s] = \mathbb{E}[R_{t+1} + \gamma\, G_{t+1} \mid S_t = s]\,.$$

Уравнение для ценности состояний:

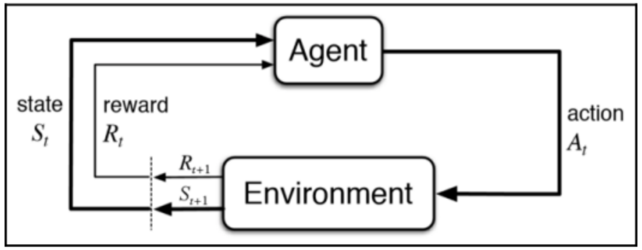
$$v(s) = r(s) + \gamma \sum_{s' \in S} p(s' \mid s)\, v(s'), \quad s \in \mathcal{S}.$$

Принцип сжимающих отображений обеспечивает сходимость итеративного метода решения:

$$V_{k+1} := R + \gamma\, P\, V_k, \quad \text{и} \quad V_k \to V \ \text{при}\ k \to \infty, \quad \text{где} \ \ V = (v(s_1), \ \dots, v(s_N))^T, \ \ R = (r(s_1), \ \dots, r(s_N))^T, \ \ P = \{p(s' \mid s)\}_{s, s' \in S}.$$

МППР

Марковский процесс вознаграждения с агентом, который может совершать действия, влияющие на будущие состояния и вознаграждения называют **марковским процессом принятия решений** (МППР, англ. Markov decision process, MDP).



Чтобы задать МППР, необходимо определить следующие составляющие.

1. Множество состояний среды $\mathcal{S}$.
2. Набор допустимых действий $\mathcal{A}(s)$ в каждом из состояний $s \in \mathcal{S}$.
3. Модель среды и способ формирования вознаграждения.

В детерминированном случае для любой допустимой пары состояние-действие $(S_t, A_t) = (s, a)$ в любой момент времени t определено новое состояние среды $S_{t+1} = s' \in \mathcal{S}$ и вознаграждение $R_{t+1} = r \in \mathcal{R}$, где $\mathcal{R}$ — это множество вознаграждений.

В стохастическом случае для любой допустимой пары состояние-действие $(S_t, A_t) = (s, a)$ в любой момент времени t определены вероятности перехода в новые состояния и случайная величина, определяющая получаемое вознаграждение.

Траектория взаимодействия агента со средой в ходе МППР является реализацией последовательности случайных величин:

$S_0, A_0, R_1, S_1, A_1, R_2, \ \dots$

Конечный МППР

В **конечном МППР** множества допустимых состояний, действий и вознаграждений ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$) имеют конечное число элементов.

Задание модели среды в этом случае сводится к заданию набора вероятностей перейти в состояние s' и получить вознаграждение r, находясь в состоянии s и совершив действие a:

$$p(s', r | s, a) := \mathbb{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$$

$$\text{для всех возможных } s, s' \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A}.$$

Независимость правой части от t значит, что эти вероятности со временем не меняются. Таким образом, для задания модели среды надо задать функцию $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} : \to [0, 1]$, такую что

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

Эта функция полностью определяет **динамику переходов среды** (англ. transition dynamic). Будем считать, что среда стационарна и со временем её динамика не меняется.

Конечный МППР

Для конечного МППР функция $p(s',\ r|\ s,\ a)$, задающая модель среды, позволяет найти вероятность перехода между состояниями s, s' при совершении действия a:

$$p(s'|\, s, a) \ := \ \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) \ = \ \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,.$$

Эта же функция позволяет найти ожидаемые вознаграждения для пары $(s,\ a)$:

$$r(s, a) \ := \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] \ = \ \sum_{r \in \mathcal{R}} r \left(\sum_{s' \in \mathcal{S}} p(s', r \mid s, a) \right),$$

здесь $\sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$ задаёт вероятность вознаграждения r

при текущем состоянии s и действии a.

Формализм МППР с одной стороны позволяет высказывать точные теоретические утверждения. С другой стороны, он применим к широкому кругу задач в RL.

Связь МППР и МПВ

Пусть агент в МППР действует по некоторой выбранной стратегии π.

В стохастическом случае стратегия значит, что для каждого s из $\mathcal{S}$ определены числа

$\pi(a \mid s)$, означающие вероятности выбора действия a в состоянии s для каждого действия $a \in \mathcal{A}$.

При фиксированной стратегии можно считать, что агент просто наблюдает смену состояний и получает вознаграждения.

При этом, легко вычислить вероятность смены состояния s на состояние s':

$$P^{\pi}(s' \mid s) := \sum_{a \in \mathcal{A}} \pi(a \mid s)\; p(s' \mid s, a) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{r \in \mathcal{R}} p(s', r \mid s, a).$$

Ожидаемые вознаграждения для пары (s, a) мы нашли на прошлом слайде. Чтобы получить ожидаемое вознаграждение в состоянии s, надо эти вознаграждения усреднить с учётом вероятности выбора действия a, то есть

$$R^{\pi}(s) := \mathbb{E}[R_{t+1} \mid S_t = s] = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(\sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a) \right).$$

Поскольку задача агента заключается в поиске “хорошей” стратегии, то может потребоваться вычислять величины вида $\mathbb{E}[R_{t+1} \mid S_t = s]$ для различных стратегий. Ясно, что для различных стратегий эти ожидания могут принимать различные значения. Чтобы отметить зависимость математических ожиданий от стратегии, будем использовать обозначение вида:

$$R^{\pi}(s) = \mathbb{E}_{\pi}[R_{t+1} \mid S_t = s].$$

МППР: функция ценности состояний

Функция ценности состояния (англ. state value function) или V-функция — это функция $v_\pi : \mathcal{S} \to \mathbb{R}$, выражающая насколько хорошо для агента нахождение в данном состоянии с точки зрения длительной перспективы при заданной стратегии поведения π.
Формально, это математическое ожидание дохода G_t при старте из состояния $S_t = s$ и выборе действий по стратегии π:

$$v_\pi(s) := \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma\, G_{t+1} \mid S_t = s].$$

Установленная связь МППР и МПВ позволяет применить изученные ранее методы вычисления функции ценности состояний для вычисления $v_\pi(s)$ для некоторой фиксированной стратегии π.

Обозначим $V_\pi := \{v_\pi(s)\}_{s\in\mathcal{S}}$, $R := \{R^\pi(s)\}_{s\in\mathcal{S}}$, $P := \{P^\pi(s' \mid s)\}_{s,s'\in\mathcal{S}}$.
Уравнение для функции ценности состояний ранее было установлено для МПВ, оно имеет вид:

$$V_\pi = R + \gamma\, P\, V_\pi \quad \text{или} \quad v_\pi(s) = R^\pi(s) + \gamma \sum_{s'\in\mathcal{S}} P^\pi(s' \mid s)\, v_\pi(s'), \quad s \in \mathcal{S}.$$

Подставив выражения для $R^\pi(s)$ и $P^\pi(s' \mid s)$, получим

$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \left(\sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r \mid s, a) \right) + \gamma \sum_{s'\in\mathcal{S}} \left(\sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{r\in\mathcal{R}} p(s', r \mid s, a) \right) v_\pi(s') =$$

$$\sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r \mid s, a)\, (r + \gamma\, v_\pi(s')).$$

Это уравнение относительно $v_\pi(s)$ называют **уравнением Беллмана** для функции ценности состояний.

МППР: оценка стратегии

Для вычисления функции ценностей состояний $v_\pi(s)$ используем итерационный метод.

Зафиксируем начальное приближение $V_0 = \{V_0(s)\}_{s \in \mathcal{S}}$. Итерация имеет вид:

$$V_k(s) := \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, V_{k-1}(s')).$$

Как установлено выше $V_k(s) \to v_\pi(s)$ при $k \to +\infty$ для каждого $s \in \mathcal{S}$ при $\gamma < 1$. Этот итерационный подход и является одним из методов **оценивания стратегии** (англ. policy evaluation).

На каждой итерации обновляются ценности по всем состояниям на основе значений ценности на прошлом шаге итерации. Для организации вычислений требуется два массива V_k и V_{k-1}.

In-place вычисления: есть один массив V, где хранятся текущие оценки ценностей $v_\pi(s)$.

Приближения обновляются в цикле, пробегающем по всем состояниям, по формуле

$$V(s) := \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, V(s')).$$

Для каждого состояния s новое значение $V(s)$ вычисляется с помощью массива V, после чего это значение сразу обновляется в массиве V.

Известно, что при in-place вычислениях сходимость сохраняется и она более быстрая, так как обновлённые приближения доступны сразу.

МППР: псевдокод для оценки стратегии

Приведём псевдокод для оценки стратегии с помощью итерационного метода и **in-place** вычислениях.

Инициализировать:

число $\theta > 0$ (порог для критерия остановки итераций)

значения $V(s)$ для $s \in \mathcal{S}$ произвольным образом

Повторять:

$\Delta := 0$

Повторять для каждого $s \in \mathcal{S}$:

$v := V(s)$

$$V(s) := \sum_{a\in\mathcal{A}} \pi(a\,|\,s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r\,|\,s, a)\,(r + \gamma\, V(s'))$$

$\Delta := \max\{\Delta, |v - V(s)|\}$

пока не окажется, что $\Delta < \theta$.

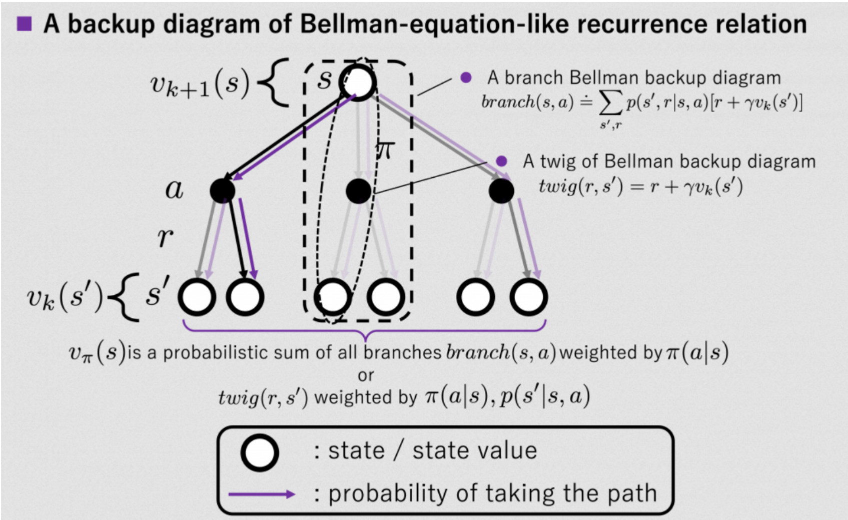
Функция ценности состояний $v_\pi(s)$ полезна, чтобы оценить конкретную стратегию π.

Для построения оптимальной стратегии более полезна функция ценности действий.

МППР: диаграмма для уравнения Беллмана

$$V_{k+1}(s) \ := \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in S} \sum_{r\in\mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, V_k(s'))$$

Для нахождения новой оценки ценности $V_k(s)$ для состояния s мы заглядываем в будущие возможные состояния на один шаг и проводим усреднение по возможным значениям $r + \gamma\, V_k(s')$.



Источник: https://data-science-blog.com/blog/2022/03/01/four-propositions-to-dynamic-dynamic-programming-dynamic-programming-and-the-bellman-equation-part-two

В итерационном методе обновление оценки для текущего состояния основаны на текущих оценках следующих состояний. Этот процесс построения новых оценок на основе других оценок называется **бутстреппингом** (англ. bootstrapping).

Бутстрэп — «ремешки на ботинках», **бутстреппинг** происходит от выражения «потянуть самого себя за ремешки на ботинках и так перелезть через ограду».

МППР: функция ценности действий

Функция ценности действий (англ. action value function) или Q-функция — это функция $q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, её значения $q_\pi(s, a)$ выражают насколько хорошо для агента в данном состоянии s сделать действие a с точки зрения длительной перспективы при заданной стратегии поведения π.

Формально, это математическое ожидание дохода G_t при старте из состояния $S_t = s$ и действии $A_t = a$ и следовании затем стратегии π:

$$q_\pi(s, a) := \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi[R_{t+1} + \gamma\, G_{t+1} \mid S_t = s, A_t = a].$$

Установим взаимосвязь между $v_\pi(s)$ и $q_\pi(s, a)$:

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, q_\pi(s, a).$$

Если известна модель среды, то можно выразить $q_\pi(s, a)$ через $v_\pi(s)$ (см. следующий слайд).

Если же модели среды нет в наличии, то для агента строить приближения функции ценности действий $q_\pi(s, a)$ полезнее, чем $v_\pi(s)$.
Функция ценности действий даёт прямой ответ на вопрос о том какие действия лучше совершать:

в состоянии s надо делать такое действие a, что значение $q_\pi(s, a)$ максимально, такой выбор ведёт к большему ожидаемому доходу.

В реальных задачах агента сразу обучают функции $q_\pi(s, a)$.

МППР: функция ценности действий

Выразим $q_\pi(s, a)$ через $v_\pi(s)$.

$$q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma\, G_{t+1} \,|\, S_t = s, A_t = a] =$$
$$\mathbb{E}_\pi[R_{t+1} \,|\, S_t = s, A_t = a] + \gamma\, \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s, A_t = a] =$$
$$\sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r \,|\, s, a) + \gamma \sum_{s'\in\mathcal{S}} p(s' \,|\, s, a)\, \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] =$$
$$\sum_{r\in\mathcal{R}} \sum_{s'\in\mathcal{S}} p(s', r \,|\, s, a)\,(r + \gamma\, v_\pi(s')) =$$
$$\sum_{r\in\mathcal{R}} \sum_{s'\in\mathcal{S}} p(s', r \,|\, s, a) \left(r + \gamma \sum_{a'\in\mathcal{A}} \pi(a' \,|\, s')\, q_\pi(s', a') \right).$$

Это **уравнение Беллмана** для функции ценности действий.

Для итеративного решения надо задать начальные значения $Q_0(s, a)$ для всех $s \in \mathcal{S}$ и $a \in \mathcal{A}$ и использовать формулу

$$Q_{k+1}(s, a) = \sum_{r\in\mathcal{R}} \sum_{s'\in\mathcal{S}} p(s', r \,|\, s, a) \left(r + \gamma \sum_{a'\in\mathcal{A}} \pi(a' \,|\, s')\, Q_k(s', a') \right).$$

Принцип сжимающих отображений гарантирует сходимость $Q_k(s, a) \to q_\pi(s, a)$ при $\gamma < 1$.

Процесс вычислений также можно проводить **in-place**.