Введение в обучение с подкреплением

Тема 6: Методы Монте-Карло

Лектор: Кривошеин А.В.

Особенности методов Монте-Карло

Для работы методов динамического программирования требуется модель среды, то есть функция $p(s', r \mid s, a)$. В реальных задачах **модели среды, как правило, нет**.

Узнать информацию о поведении среды агент может лишь **методом проб и ошибок**, генерируя траектории в ходе взаимодействия со средой.

Методы Монте-Карло (МК) относятся к классу методов, позволяющих решать задачи обучения с подкреплением без модели среды. Особенность методов МК:

агент взаимодействует со средой **по эпизодам.**

Сначала решим задачу оценки стратегии с помощью методов МК. Для оценки стратегии π надо оценить функцию ценности состояний или V-функцию:

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

Идея метода МК в том, что **оценка** $v_\pi(s)$ **формируется усреднением наблюдавшихся доходов** после посещения состояния s в ходе взаимодействия агента со средой.

То есть из каждой траектории взаимодействия извлекаются вознаграждения и вычисляются фактически полученные доходы для каждого встреченного состояния. Затем улучшаются оценки V-функции.

Увеличение числа эпизодов взаимодействия ведёт к повышению точности оценки величины $v_\pi(s)$.

МК: оценка стратегий

Возможно два варианта того, как реализовать усреднение наблюдаемого дохода.

1. **Метод МК первого посещения** (англ. first visit MC) усредняет доходы, полученные после первого в каждом эпизоде посещения состояния s.

2. **Метод МК всех посещений** (англ. every visit MC) усредняет доходы, полученные после каждого посещения состояния s.

Пусть массив Gains(s) содержит фактически полученные доходы после первого (или после каждого) посещения состояния s.
Тогда оценка ценности состояния s равна:

$$V(s) := \frac{1}{|\text{Gains}(s)|} \sum_{G \in \text{Gains}(s)} G, \quad \text{где } |\text{Gains}(s)| - \text{это количество элементов в множестве } \text{Gains}(s).$$

Если стратегия агента π гарантирует посещение всех состояний, то имеет место сходимость $V(s) \to v_\pi(s)$ с ростом числа эпизодов.

Инкрементная реализация обновления оценки V-функции имеет вид:

$$V_n(s) := V_{n-1}(s) + \alpha_n (G - V_{n-1}(s)),$$

где индекс n у V_n означает число первых в эпизоде (или всех) посещений состояния s к текущему моменту,
$V_{n-1}(s)$ — это текущая оценка, G — это доход, полученный после посещения состояния s в n-ый раз.
Если $\alpha_n = \frac{1}{n}$, то $V_n(s)$ — это средний доход, полученный после n посещений состояния s.

МК: оценка стратегий, псевдокод

Приведём псевдокод метода **МК первого посещения** для оценки V-функции:

---

1. Инициализировать:

значения $V(s) = 0$ для всех $s \in S$ (оценки ценностей состояний)

значения $N(s) = 0$ для всех $s \in S$ (число посещений состояния s)

2. Оценка стратегии:

Повторять:

Генерация эпизода по заданной стратегии:

$S_0, A_0, R_1, S_1, A_1, \ldots, R_T, S_T$.

$G = 0$

Цикл по $t = T - 1, T - 2, \ldots, 1, 0$

$G = \gamma G + R_{t+1}$

Если состояние $s := S_t$ не встречается ранее в $S_0, \ldots, S_{t-1}$, то

$N(s) = N(s) + 1$

$V(s) = V(s) + \frac{1}{N(s)}(G - V(s))$ (обновить оценку)

Алгоритм действий для метода МК всех посещений можно записать аналогично, исключив проверку о встрече состояния $s = S_t$ ранее в траектории эпизода.

Методы МК удобны, если надо оценить ценность только одного или нескольких состояний, поскольку вычислительная сложность формально не зависит от общего числа состояний. Однако, для хорошей оценки стратегии может потребоваться значительное число эпизодов.

МК: как улучшать стратегию

В методах динамического программирования модель среды известна и это позволяло по V-функции найти Q-функцию, с помощью которой улучшалась стратегия. В рамках метода итерации по стратегиям новая улучшенная стратегия формировалась жадно относительно текущей оценки Q-функции:

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

Когда модели среды нет, то найти Q-функцию по V-функции невозможно. Поэтому **надо сразу обучать агента Q-функции** $q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$. Общая идея оценки Q-функции та же, что и для оценки V-функции.

Формула обновления оценки Q-функции имеет вид:

$$Q_n(s, a) = Q_{n-1}(s, a) + \alpha_n(G - Q_{n-1}(s, a)).$$

Методы МК первого посещения и всех посещений, сформулированные выше, можно адаптировать и для оценки функции ценности действий $q_\pi(s, a)$. Опять же, если стратегия агента π гарантирует посещение всех состояний, то имеет место сходимость $Q_n(s, a) \to q_\pi(s, a)$ с ростом числа посещений каждой пары (s, a).

МК: как улучшать стратегию

Для поиска оптимальной стратегии можно использовать алгоритм **итерации по стратегии**:

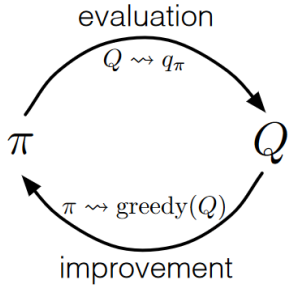
1. для текущей стратегии π сформировать хорошую оценку Q-функции $q_\pi(s, a)$;
2. сформировать новую стратегию жадно относительно $q_\pi(s, a)$.

Хорошая оценка для Q-функции может потребовать очень большого числа эпизодов. Однако, можно использовать метод **обобщённой итерации по стратегиям (ОИС):**

1. формируем траекторию по текущей стратегии и обновляем оценки функции ценности, чтобы быть ближе к истинной функции ценности текущей стратегии.
2. улучшаем стратегию относительно текущей оценки Q-функции и возвращаемся к пункту 1.

Эти два шага совместно приближают и функцию ценности и стратегию к оптимальным.

Out[•]=



МК: как улучшать стратегию

Проблема: при генерации эпизодов агент следует некоторой стратегии и при этом может оказаться так, что некоторые пары состояние-действие (s, a) могут быть вообще никогда не посещены. Тогда для этих пар нет оценок ценности, что может препятствовать нахождению оптимальной стратегии.

Решений этой проблемы две.
1. **Исследовательские старты**: будем принудительно помещать агента в каждую пару состояния-действия с ненулевой вероятностью.
2. Будем проводить итерацию только по **стохастическим стратегиям**.

МК: исследовательские старты

Исследовательские старты для метода МК (англ. MC-ES, exploring starts) заключаются в том, что начальная пара (s, a) выбирается случайно с ненулевой вероятностью. Приведём алгоритм ОИС для метода МК с исследовательскими стартами.

---

1. Инициализировать:

значения $Q(s, a) = 0$ для всех состояний $s \in \mathcal{S}$ и действий $a \in \mathcal{A}$

значения $N(s, a) = 0$ для всех состояний $s \in \mathcal{S}$ и действий $a \in \mathcal{A}$ для хранения числа посещений пары (s, a).

2. ОИС:

Повторять:

Выбор начальной пары (S_0, A_0) из набора всех пар с ненулевой вероятностью

Генерация эпизода по заданной стратегии π:

$S_0, A_0, R_1, S_1, A_1, \ldots, R_T, S_T.$

$G = 0$

Цикл по $t = T - 1, T - 2, \ldots, 1, 0$

$G = \gamma G + R_{t+1}$

Если пара S_t, A_t не встречается ранее в парах $(S_0, A_0), \ldots, (S_{t-1}, A_{t-1})$, то

$N(S_t, A_t) = N(S_t, A_t) + 1$

$Q(S_t, A_t) = Q(S_t, A_t) + \frac{1}{N(s,a)} (G - Q(S_t, A_t))$

$\pi(S_t) = \arg\max_a Q(S_t, a)$

Метод МК первого посещения, легко изменить на метод МК всех посещений, убрав проверку соответствующего условия.

Недостатком алгоритма MC-ES является то, что в некоторых задачах нет возможности стартовать из произвольной выбранной пары состояние-действие.

МК: eps-мягкие стратегии

Стратегию π называют **мягкой**, если $\pi(a \mid s) > 0$ для всех s и a (то есть вероятность выбора каждого действия положительна).
Стратегию π называют **ε-мягкой**, если $\pi(a \mid s) > \frac{\varepsilon}{|\mathcal{A}(s)|}$ для всех s и a.

Будем проводить итерации ε-мягким стратегиям и искать оптимальные стратегии среди них. Аналогом жадных стратегий для ε-мягких стратегий будут ε-**жадные стратегии**:

с вероятностью $1 - \varepsilon$ выбирается жадное действие относительно текущей оценки $Q(s, a)$.

с вероятностью ε случайно выбирается действие из $\mathcal{A}(s)$ с вероятностью $\varepsilon / |\mathcal{A}(s)|$.

Иными словами, если есть одно действие, равное $\underset{a}{\operatorname{argmax}}\, Q(s, a)$, то

$$\pi(a \mid s) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|} & \text{при } a = \underset{a}{\operatorname{argmax}}\, Q(s, a) \\ \frac{\varepsilon}{|\mathcal{A}(s)|} & \text{иначе.} \end{cases}$$

Рассмотрим алгоритм **итерации по ε-мягким стратегиям**:

0. Фиксируем ε-мягкую стратегию π

1. Вычисляем функцию ценности действий текущей стратегии $q_\pi(s, a)$.

2. Формируем ε-жадную относительно $q_\pi(s, a)$ стратегию π', полагаем $\pi := \pi'$ и к пункту 1.

Далее, покажем, что стратегия, выбранная ε-жадно относительно Q-функции текущей стратегии действительно даёт более лучшую стратегию.

МК: eps-мягкие стратегии

Теорема. Пусть π является некоторой ε-мягкой стратегией. Пусть π' является ε-жадной стратегией относительно $q_\pi(s,\ a)$. Тогда $\pi' \geq \pi$, то есть $v_{\pi'}(s) \geq v_\pi(s)$.

Доказательство. Рассмотрим цепочку из равенств и неравенств

$$\sum_{a \in \mathcal{A}} \pi'(a \mid s)\, q_\pi(s, a)$$

$$= \frac{\varepsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} q_\pi(s, a) + (1 - \varepsilon) \max_a q_\pi(s, a)$$

$$= \frac{\varepsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} q_\pi(s, a) + \sum_{a \in \mathcal{A}} \left(\pi(a \mid s) - \frac{\varepsilon}{|\mathcal{A}|} \right) \max_a q_\pi(s, a)$$

$$\geq \frac{\varepsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} q_\pi(s, a) + \sum_{a \in \mathcal{A}} \pi(a \mid s)\, q_\pi(s, a) - \sum_{a \in \mathcal{A}} \frac{\varepsilon}{|\mathcal{A}|} q_\pi(s, a)$$

$$= \sum_{a \in \mathcal{A}} \pi(a \mid s)\, q_\pi(s, a) = v_\pi(s).$$

Тогда по общей теореме об улучшении стратегий $\pi' \geq \pi$. •

Можно показать, что алгоритм итерации по ε-мягким стратегиям действительно позволяет найти оптимальную ε-мягкую стратегию, то есть ту, которая лучше или равна любой другой ε-мягкой стратегии.

МК: ОИС + eps-мягкие стратегии

Далее, можно также завести счётчик для сгенерированных эпизодов $k = 1,\ 2,\ \ldots$ и уменьшать ε по некоторому правилу, так что $\varepsilon \to 0$ при $k \to \infty$.

Такой метод МК называют greedy in the limit with infinite exploration (GLIE), так как при $\varepsilon \to 0$ стратегия становится жадной, но все пары (s, a) потенциально могут быть выбраны бесконечное число раз.

Известно, что при GLIE оценки ценностей $Q(s, a)$ сходятся к оптимальным $q_*(s, a)$.

Оценки ценностей $Q(s, a)$ формируются как выборочные средние или с использованием шагов обучения α_n по условиям Роббинса-Монро

$$\sum_n \alpha_n = \infty, \quad \sum_n \alpha_n^2 < \infty.$$

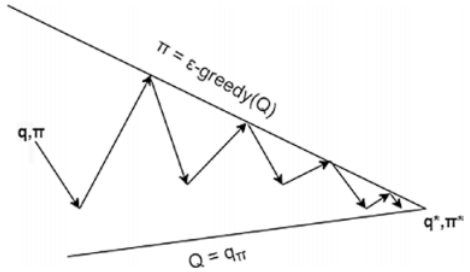
Объединим в один алгоритм эти две идеи: ОИС + ε-мягкие стратегии:

0. Фиксируем ε-мягкую стратегию π.

1. Формируем траекторию по текущей стратегии и обновляем оценки функции ценности, чтобы быть ближе к истинной функции ценности текущей стратегии.

2. Формируем ε-жадную стратегию, относительно текущей оценки ценности и к пункту 1.

Схема работы алгоритма проиллюстрирована ниже



МК: ОИС + eps-мягкие стратегии, псевдокод

1. Инициализировать:

значения $Q(s, a)$ произвольным образом (или нулю) для всех состояний $s \in \mathcal{S}$ и действий $a \in \mathcal{A}$,

значения $N(s, a) = 0$ для всех состояний $s \in \mathcal{S}$ и действий $a \in \mathcal{A}$ для хранения числа посещений пары (s, a),

некоторую ε-жадную стратегию π для некоторого $\varepsilon \in (0, 1)$.

2. ОИС с ε-мягкими стратегиями:

Повторять:

Генерация эпизода по заданной стратегии π:

$S_0, A_0, R_1, S_1, A_1, \ldots, S_T, R_T$.

$G = 0$

Цикл по $t = T - 1, T - 2, \ldots, 1, 0$

$G = \gamma G + R_{t+1}$

$N(S_t, A_t) = N(S_t, A_t) + 1$

$Q(S_t, A_t) = Q(S_t, A_t) + \frac{1}{N(S_t, A_t)} (G - Q(S_t, A_t))$,

$A^* = \arg\max_a Q(S_t, a)$

Для всех $a \in \mathcal{A}(S_t)$ обновить стратегию:

$$\pi(a \mid S_t) = \begin{cases} 1 - \varepsilon + \varepsilon / \left|\mathcal{A}(S_t)\right| & \text{при } a \in A^* \\ \varepsilon / \left|\mathcal{A}(S_t)\right| & \text{иначе} \end{cases}$$

В алгоритме оценки формируются как выборочное среднее доходов, то есть шаг обучения равен $\alpha_n = \frac{1}{n}$, где n число выборов пары (S_t, A_t), но можно использовать иные способа выбора шага обучения.

МК: дополнительные особенности

В представленных псевдокодах нет условия выхода из итераций, как это было в методах ДП. Обучаться можно до тех пор, пока нас не устроит результат обучения.

Важно отметить, что траектории, полученные на прошлых итерациях уже нельзя использовать для обучения на текущем шаге, так как ранее траектория генерировалась с другой стратегией. Эта особенность позволяет ввести следующую классификацию методов обучения.

Метод обучения называется **методом с единой стратегией** (англ. on-policy method), если он при обучении использует данные, сгенерированные только по текущей стратегии. Если при обучении у нас меняется стратегия, то на каждой итерации надо использовать свой набор данных, сгенерированных по этой стратегии.

Метод обучения называется **методом с разделённой стратегией** (англ. off-policy method), если любые сгенерированные данные можно использовать в обучении.

При обучении методом с разделённой стратегией, та стратегия, по которой генерируются траектории называется **стратегией поведения** (она более исследовательская). А вторая называется **целевой стратегией** (она менее исследовательская и может быть даже детерминистической).

Такие методы более эффективны с точки зрения использования полученного ранее опыта (англ. sample efficiency).

Методы МК можно можно модифицировать, так чтобы проводить обучение по методу с разделённой стратегией. Однако, мы рассмотрим методы обучения с разделённой стратегией в контексте TD методов.