Введение в обучение с подкреплением

Тема 7: TD методы

Лектор: Кривошеин А.В.

Особенности TD методов

Методы ДП позволяют агенту заранее просчитать оптимальные действия, максимизирующие доход. Но для их реализации надо знать модель среды.

Методы МК оценивают совершённые действия в конце эпизода уже после взаимодействия агента со средой.

Оба этих метода не позволяют агенту **обучаться он-лайн**, то есть обучаться прямо в процессе взаимодействия агента со средой. Для обучения он-лайн, надо решить проблему мгновенной оценки действий и найти возможность искать лучшую стратегию действий по ходу взаимодействия.

Класс методов, решающих эту задачу, называют **методами обучения на основе временных различий** или **TD методами** обучения (англ. Temporal Difference Learning). Эти методы позволяют сделать один шаг по траектории и сразу же обновить оценку сделанного действия.

Напомним, в **методах ДП** для обновления оценки ценности текущего состояния использовалось уравнение Беллмана, которое просматривает возможные состояния и вознаграждения на следующем шаге и проводит усреднение вознаграждений и ценностей новых состояний по всем вероятностям перехода

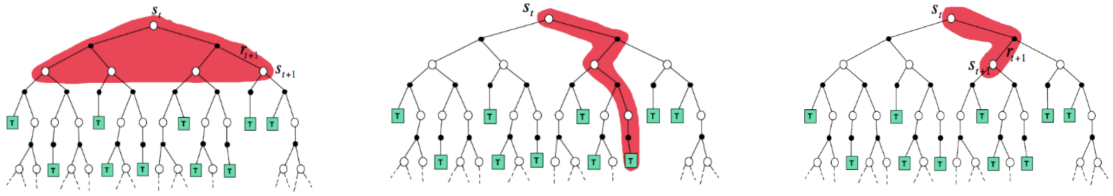
$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in S} \sum_{r \in \mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, v_\pi(s')).$$

Причём в итерационных алгоритмах обновление оценки для текущего состояния использует бутстреппинг, то есть обновление основано на текущих оценках следующих состояний.

Особенности TD методов

Методы МК основаны на генерации траекторий по выбранной стратегии. Оценка ценности состояния — это усреднение фактически полученных доходов при старте из этого состояния по многим сгенерированным траекториям.

Схематично, эти алгоритмы можно проиллюстрировать так.



Закрашенная область — это область, требуемая для работы алгоритма обучения, чтобы обновить ценности состояний, а значит и стратегию.

Для методов ДП — это все возможные состояния следующего временного шага

Для методов МК — это состояния по траектории от текущего до завершающего.

Для TD методов — это часть траектории от текущего состояния до следующего.

Как и методах МК для оценки функции ценности надо генерировать траектории взаимодействия агента со средой. Но в методах TD не надо ждать конца эпизода, чтобы обновлять оценки и не требуется знание модели среды.

TD: обновление оценок ценности

В методах МК мы ждём конца эпизода, считаем фактически полученный доход и используем это число как цель для обновления оценки $V(S_t)$:

$V^{\text{new}}(S_t) := V(S_t) + \alpha(G_t - V(S_t))$.

В TD методах мы не ждём конца эпизода, а обновляем оценку после каждого шага.
Цель для обновления формируется с использованием ближайшего вознаграждения R_{t+1} и оценки ценности нового состояния $V(S_{t+1})$ в виде $R_{t+1} + \gamma\, V(S_{t+1})$.

Правило обновления в этом случае имеет вид

$V^{\text{new}}(S_t) := V(S_t) + \alpha\,(R_{t+1} + \gamma\, V(S_{t+1}) - V(S_t))$.

Разность $\delta := R_{t+1} + \gamma\, V(S_{t+1}) - V(S_t)$ между целью обновления и текущей оценкой называют **TD ошибкой**.

Обновление в TD методах основано на ранее полученной оценке, то есть TD метод использует **бутстреппинг**, как и методы ДП.
Напомним, что функция ценности состояний имеет следующий вид

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

из уравнения Беллмана : $v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in S} \sum_{r\in\mathcal{R}} p(s', r \mid s, a)\,(r + \gamma\, v_\pi(s')) = \mathbb{E}_\pi[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s]$.

В методах МК целью обновления является фактически полученный доход G_t.
В TD методах целью обновления является сумма полученного вознаграждения и оценка ценности следующего состояния $R_{t+1} + \gamma\, V(S_{t+1})$.

TD: оценка стратегии, псевдокод

Приведём алгоритм работы TD метода для оценки функции ценности состояний v_π при фиксированной стратегии π и размере шага обучения α.

---

1. Инициализировать:

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

2. Оценка стратегии

Повторять для каждого эпизода:

Выбрать начальное состояние s

Повторять:

Выбрать действие a по стратегии π в состоянии s

Наблюдать r, s'

Обновить оценку $V(s) := V(s) + \alpha\left(r + \gamma V(s') - V(s)\right)$

$s := s'$

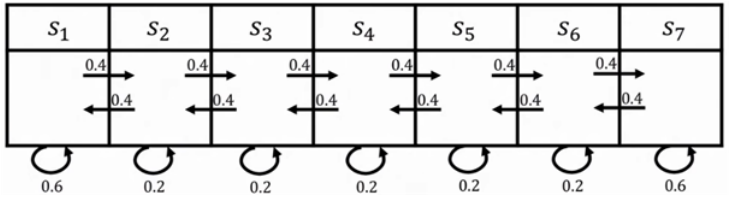
Если s заключительное состояние, то выйти из цикла.

Известно, что оценка ценности состояния $V_n(s)$ (n номер итерации) сходится к истинному значению $v_\pi(s)$, если каждое состояние посещается потенциально бесконечное число раз и размер шага обучения α_n убывает в соответствии с условиями Роббинса-Монро: $\sum_n \alpha_n = \infty, \quad \sum_n \alpha_n^2 < \infty.$

TD: особенность оценки стратегии на примере

Пример. Рассмотрим пример с 7-ю состояниями. Действия: $\{a_0, a_1, a_2\}$ = {влево, вправо, на месте}.

Пусть выбрана некоторая стратегия действий, $\gamma = 1$.



Вознаграждения зададим следующим образом: $r(s_2, a_0) = -1$, $r(s_6, a_1) = 10$, $r(s_i, a_j) = 0$ иначе.

Пусть начальные оценки ценностей состояний равны нулю. Рассмотрим некоторую траекторию:

($\boldsymbol{s_4}$, a_1, 0, $\boldsymbol{s_5}$, a_1, 0, $\boldsymbol{s_6}$, a_1, 10, $\boldsymbol{s_7}$).

Тогда по методу МК обновлённые оценки будут иметь вид $V(s_4) = 10$, $V(s_6) = 10$,

а по TD методу обновлённые оценки будут иметь вид $V(s_4) = 0$, а $V(s_6) = 10$.

То есть по методу МК информация о финальном вознаграждении сразу доступна для состояния из начала траектории. Для TD методов это не так. Для такой траектории только для состояния s_6 будет по существу обновлена оценка: $V(s_6) = 10$. Однако, для TD методов обновление оценок происходит гораздо чаще, чем для методов МК, что компенсирует эту особенность.

TD: метод SARSA

Метод поиска оптимальной стратегии основан на оценке Q-функции по схеме ОИС + ε-мягкие стратегии:

1. **обновляем** оценки Q-функции на основе TD метода, генерируя траектории по текущей стратегии;
2. **улучшаем** стратегию, используя ε-жадный выбор действий относительно текущей Q-функции.

Обновление оценки Q-функции имеет вид:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma\, Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\,].$$

Для заключительных состояний полагаем, что $Q(S_T, A_T) = 0$ для всех возможных действий. То есть оценка на последнем шаге будет иметь вид

$$Q(S_{T-1}, A_{T-1}) \leftarrow Q(S_{T-1}, A_{T-1}) + \alpha[R_T - Q(S_{T-1}, A_{T-1})\,]$$

Чтобы осуществить обновление оценки Q-функции необходимо знать пять значений: S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}. Причём новое действие A_{t+1} выбирается исходя из текущей стратегии π. Этот алгоритм называют **SARSA** по первым буквам элементов пятерки, определяющей переход из одной пары состояние-действие в другую.

TD: SARSA, псевдокод

Приведём псевдокод алгоритма SARSA для поиска оптимальной стратегии и оптимальной функции ценности действий q_*. Фиксируем размер шага обучения $\alpha > 0$ и $\varepsilon > 0$, а также $\gamma \in (0, 1)$.

---

1. Инициализировать:

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

стратегию π определить ε-жадно относительно Q-функции

2. SARSA

Повторять для каждого эпизода:

Выбрать начальное состояние s

Выбрать a следуя ε-жадной стратегии относительно Q-функции

Повторять:

По s, a наблюдать вознаграждение r и новое состояние s'

Выбрать a' следуя ε-жадной стратегии относительно Q-функции

Обновить оценку $Q(s, a) := Q(s, a) + \alpha\left[r + \gamma Q\left(s', a'\right) - Q(s, a)\right]$

$s := s'$, $a := a'$

Если s заключительное состояние, то выйти из цикла.

Уменьшить ε

Вернуть стратегию π на основе найденой Q-функции.

SARSA является методом обучения с единой стратегией: траектории генерируются по той стратегии, которую мы оцениваем. Значит нельзя обучаться на старых траекториях, так как стратегии изменяются со временем.

TD: SARSA

Теорема. SARSA для конечного МППР сходится к оптимальным значениям функции ценности действий $Q(s,\ a) \to q_*(s,\ a)$ при следующих условиях:

1. Итерация по ε-стратегиям является GLIE, то есть ε убывает к нулю с ростом числа итераций;
2. Шаг обучения α_n удовлетворяет условиям Роббинса-Монро: $\sum_n \alpha_n = \infty, \quad \sum_n \alpha_n^2 < \infty$.

Например, можно взять $\alpha_n = \frac{1}{n}$.

SARSA является методом обучения с единой стратегией: траектории генерируются по той стратегии, которую мы оцениваем. То есть при обучении формально нельзя использовать прошлый опыт взаимодействия со средой, так как стратегии изменяются со временем.

Сравнение МК и TD методов:

Нет однозначного ответа на вопрос какой метод лучше: МК или TD.

На практике, как правило, TD методы демонстрируют более быструю сходимость с постоянным шагом обновления оценки α.