Введение в обучение с подкреплением

Тема 12: Метод REINFORCE

Лектор: Кривошеин А.В.

Методы на основе приближения стратегии

Изученные ранее методы обучения с подкреплением относятся к двум классам:

1. **Методы на основе модели** (англ. model-based methods). В этом классе мы рассмотрели методы обучения на основе динамического программирования;
2. **Методы на основе значений функций ценности** (англ. value-based methods). В этом классе мы рассмотрели такие методы обучения, как методы МК, TD методы, в частности SARSA и модификации, а также Q-learning.

Обсудим ещё один класс методов обучения, **основанный на приближении стратегий** (англ. policy-based methods).

В этих методах агент занимается непосредственным улучшением своей стратегии. Хорошая стратегия должна генерировать хорошие действия, то есть траектории сгенерированные по этой стратегии должны максимизировать доход агента с начала эпизода:

$$\max_{\pi} \mathbb{E}_{\pi}[G_0], \quad \text{где} \quad G_0 = \sum_{t=0}^{T-1} \gamma^t R_{t+1}, \quad \text{где } T \text{ момент завершения эпизода.}$$

REINFORCE

Классический policy-based метод — это метод REINFORCE (1992). Суть метода REINFORCE заключается в построении приближения для **стохастических стратегий с помощью параметрической модели π_θ.**

Стохастическая стратегия π – это отображение из множества состояний $\mathcal{S}$ в набор вероятностей выбора действий из множества $\mathcal{A}$. Для вероятности выбора действия $a \in \mathcal{A}$ в состоянии $s \in \mathcal{S}$ будем использовать обозначение $\pi(a|\, s)$, причём

$$\pi(a|\, s) \in [0,\, 1] \quad \text{и} \quad \sum_{a \in \mathcal{A}} \pi(a \mid s) = 1.$$

Рассмотрим модель π_θ, которая на вход принимает состояния, а на выходе выдаёт распределение вероятностей выбора действий. Для конечного набора действий $\mathcal{A} = \{a_1,\ \ldots,\ a_n\}$ распределение вероятностей является **категориальным распределением**, которое задаётся набором вероятностей $(\pi_\theta(a_1|\, s),\ \ldots,\ \pi_\theta(a_n|\, s))$.

Во время обучения те действия, которые приводят к хорошим результатам, должны иметь большую вероятность их выбора, то есть они **положительно подкрепляются**. Действия же ведущие к плохим результатам должны иметь меньшую вероятность их выбора.

Для параметрической модели изменение распределения вероятностей происходит через изменение параметров модели.
При успешном обучении распределение вероятностей выбора действий у параметрической модели π_θ будут сдвигаться к распределению, которое обеспечивает всё более и более хорошее поведение агента в среде.

Изменения параметров модели осуществляется в ходе градиентного подъёма в ходе максимизации ожидаемого дохода $\mathbb{E}_{\pi_\theta}[G_0]$ от действий агент по стратегии π_θ. Метод REINFORCE также известен как **Policy Gradient Algorithm**.

REINFORCE

Для реализации policy-based методов (и в частности REINFORCE) требуется определить:

1. вид параметрической модели для стратегий π_θ;
2. целевую функцию или функционал для максимизации $J(\theta)$;
3. формулу обновления параметров стратегии по методу градиентного подъёма.

В качестве параметрической для стратегий π_θ модели рассмотрим ИНС. Каждый набор параметров ИНС θ задаёт отдельную стратегию. Для дискретного набора действий результат работы модели π_θ должен быть вектором вероятностей.

Чтобы произвольный вектор перевести в набор вероятностей, можно применить к этому вектору функцию **Softmax**:

$$\text{Softmax}(z) = \frac{1}{\sum_{i=1}^{d} e^{z_i}} \left(e^{z_1}, \ \ldots, \ e^{z_d} \right) \quad \text{для вектора } z = (z_1, \ \ldots, \ z_d) \in \mathbb{R}^d.$$

REINFORCE

Целевая функция для максимизации — это ожидаемый доход по всем траекториям, порождённым стратегией π_θ

$$J(\pi_\theta) \;=\; \mathbb{E}_{\pi_\theta}[G_0] \;=\; \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{T-1}\gamma^t R_{t+1}\Big].$$

Фактически посчитать значение этой функции мы не можем, но функцию J легко оценить, генерируя траектории по зафиксированной стратегии π_θ, считая доход и усредняя результат.

Цель обучения заключается в подборе таких параметров θ, которые максимизируют функцию $J(\pi_\theta)$:

$$\theta \;=\; \arg\max_\theta J(\pi_\theta).$$

Классический метод решения такой задачи — это метод градиентного подъёма. Суть его в том, чтобы сдвигать параметры θ в сторону направления градиента (сторону наибольшего роста функции), а именно

$$\theta \leftarrow \theta + \alpha\, \mathrm{grad}_\theta J(\pi_\theta), \quad \text{где } \alpha \text{ шаг обучения.}$$

Теорема (о градиенте стратегии). Формула для $\mathrm{grad}_\theta J(\pi_\theta)$ имеет вид

$$\mathrm{grad}_\theta J(\pi_\theta) \;=\; \sum_{k=0}^{T-1}\gamma^k\, \mathbb{E}_{\pi_\theta}[G_k \,\cdot\, \mathrm{grad}_\theta \ln \pi_\theta(A_k \,|\, S_k)].$$

Доказательство

Доказательство. В силу линейности математического ожидания

$$\operatorname{grad}_\theta J(\pi_\theta) \;=\; \operatorname{grad}_\theta \mathbb{E}_{\pi_\theta}[G_0] \;= \operatorname{grad}_\theta \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{T-1} \gamma^t R_{t+1}\Big] \;=\; \sum_{t=0}^{T-1} \gamma^t \operatorname{grad}_\theta \mathbb{E}_{\pi_\theta}[R_{t+1}].$$

Не ясно, как дифференцировать R_t по θ, поскольку вознаграждения R_t порождаются неизвестной нам функцией $p(s',\ r \mid s,\ a)$, задающей модель среды. Но можно отметить, что изменение параметров θ у стратегии π_θ влияет на итоговый доход G_0 через изменение вероятностей совершения действий, что и приводит к изменению дохода агента. Модель среды не зависит от этих параметров, поэтому найти градиент можно, но надо совершить ряд преобразований.

Для упрощения выкладок рассмотрим эпизодический конечный МППР. Для текущей стратегии π_θ начнём генерировать траектории. Траектория — это набор значений состояний, действий, вознаграждений от начала и до конца эпизода:

$\tau \;=\; \{s_0, a_0, r_1, s_1, a_1, \;\ldots, r_T, s_T\}.$

Каждая траектория имеет свою вероятность появления. Это произведение следующих величин:

вероятности появления начального состояния s_0, обозначим его $p(s_0)$,

вероятности совершить действие a_0 в состоянии s_0, то есть $\pi_\theta(a_0 \mid s_0)$,

вероятности получить вознаграждение r_1 и состояние s_1, то есть $p(s_1,\ r_1 \mid s_0,\ a_0)$,

совершить действие a_1 в состоянии s_1, то есть $\pi_\theta(a_1 \mid s_1)$,

вероятности получить вознаграждение r_2 и состояние s_2, то есть $p(s_2,\ r_2 \mid s_0,\ a_0)$,

и т.д.

Доказательство

$$\mathbb{P}[\tau \mid \pi_\theta] = p(s_0) \prod_{k=0}^{T-1} \pi_\theta(a_k \mid s_k)\, p(s_{k+1}, r_{k+1} \mid s_k, a_k).$$

Зафиксируем $t = 0, \ldots, T-1$. Тогда

$$\operatorname{grad}_\theta \mathbb{E}_{\pi_\theta}[R_{t+1}] = \operatorname{grad}_\theta \sum_\tau \mathbb{P}[\tau \mid \pi_\theta]\, R_{t+1}(\tau) =$$

$$\sum_\tau \operatorname{grad}_\theta \mathbb{P}[\tau \mid \pi_\theta]\, R_{t+1}(\tau) = \sum_\tau \mathbb{P}[\tau \mid \pi_\theta] \frac{\operatorname{grad}_\theta \mathbb{P}[\tau \mid \pi_\theta]}{\mathbb{P}[\tau \mid \pi_\theta]} R_{t+1}(\tau) =$$

$$\sum_\tau \mathbb{P}[\tau \mid \pi_\theta] \operatorname{grad}_\theta \ln(\mathbb{P}[\tau \mid \pi_\theta])\, R_{t+1}(\tau),$$

где сумма берётся по всем возможным траекториям.

$$\ln(\mathbb{P}[\tau \mid \pi_\theta]) = \ln\left(p(s_0) \prod_{k=0}^{T-1} \pi_\theta(a_k \mid s_k)\, p(s_{k+1}, r_{k+1} \mid s_k, a_k)\right) = \ln p(s_0) + \sum_{k=0}^{T-1} \left(\ln \pi_\theta(a_k \mid s_k) + \ln p(s_{k+1}, r_{k+1} \mid s_k, a_k)\right).$$

Но вероятности переходов среды не зависят от стратегии, а значит и от параметров θ. Тогда после применения градиента $\operatorname{grad}_\theta$ к указанной выше сумме останутся только слагаемые вида $\ln \pi_\theta(a_k \mid s_k)$.

Доказательство

Итак, $\operatorname{grad}_\theta \ln(\mathbb{P}[\tau \mid \pi_\theta]) = \sum_{k=0}^{T-1} \operatorname{grad}_\theta \ln \pi_\theta(A_k(\tau) \mid S_k(\tau))$.

Продолжая равенство для $\operatorname{grad}_\theta \mathbb{E}_{\pi_\theta}[R_{t+1}]$, получим

$$\operatorname{grad}_\theta \mathbb{E}_{\pi_\theta}[R_{t+1}] = \sum_\tau \mathbb{P}[\tau \mid \pi_\theta] \sum_{t=0}^{T-1} \operatorname{grad}_\theta \ln \pi_\theta(A_k(\tau) \mid S_k(\tau))\, R_{t+1}(\tau) = \sum_{k=0}^{T-1} \mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}].$$

Подставим выражение для $\operatorname{grad}_\theta \mathbb{E}_{\pi_\theta}[R_{t+1}]$ в $\operatorname{grad}_\theta J(\pi_\theta)$:

$$\operatorname{grad}_\theta J(\pi_\theta) = \sum_{t=0}^{T-1} \gamma^t \sum_{k=0}^{T-1} \mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}] = \sum_{k=0}^{T-1} \left(\sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}] \right) =$$

$$\sum_{k=0}^{T-1} \left(\sum_{t=0}^{k-1} \gamma^t \mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}] + \sum_{t=k}^{T-1} \gamma^t \mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}] \right) \overset{???}{=} \sum_{k=0}^{T-1} \left(0 + \gamma^k \mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, G_k] \right),$$

то есть надо показать, что $\mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}]$ будет равно нулю, при $t = 0, \ldots, k-1$. Эвристическое основание этого факта в том, что сдвиг вероятностей выбора будущих действий не повлияет на вознаграждения, которые были получены в прошлом.

Можно установить, что $\mathbb{E}_{\pi_\theta}[\operatorname{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}] = 0$ при $t = 0, \ldots, k-1$ формально.

Доказательство

Зафиксируем k и $t < k$.

Рассмотрим всевозможные траектории τ, которые в момент времени k принимают значения $S_k(\tau) = s$, $A_k(\tau) = a$.

Вероятность появления таких траекторий при действии по стратегии π_θ равна

$\mathbb{P}[S_k(\tau) = s,\ A_k(\tau) = a \mid \pi_\theta]$. Тогда

$$\mathbb{E}_{\pi_\theta}\left[\mathrm{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}\right] = \sum_{s_k \in \mathcal{S}} \sum_{a_k \in \mathcal{A}} \sum_{\tau} \mathbb{P}[S_k(\tau) = s_k,\ A_k(\tau) = a_k \mid \pi_\theta]\, \mathrm{grad}_\theta \ln \pi_\theta(a_k \mid s_k)\, r_{t+1} =$$

$$\sum_{\tau} r_{t+1} \sum_{s_k \in \mathcal{S}} \mathbb{P}[S_k(\tau) = s_k \mid \pi_\theta] \sum_{a_k \in \mathcal{A}} \mathbb{P}[A_k(\tau) = a_k \mid \pi_\theta, S_k(\tau) = s_k]\, \frac{\mathrm{grad}_\theta\, \pi_\theta(a_k \mid s_k)}{\pi_\theta(a_k \mid s_k)}.$$

Так как $\mathbb{P}[A_k(\tau) = a_k \mid \pi_\theta, S_k(\tau) = s_k] = \pi_\theta(a_k \mid s_k)$, то

$$\sum_{a_k \in \mathcal{A}} \mathbb{P}[A_k(\tau) = a_k \mid \pi_\theta, S_k(\tau) = s_k]\, \frac{\mathrm{grad}_\theta\, \pi_\theta(a_k \mid s_k)}{\pi_\theta(a_k \mid s_k)} = \sum_{a_k \in \mathcal{A}} \mathrm{grad}_\theta\, \pi_\theta(a_k \mid s_k) = \mathrm{grad}_\theta\, 1 = 0$$

В итоге, $\mathbb{E}_{\pi_\theta}[\mathrm{grad}_\theta \ln \pi_\theta(A_k \mid S_k)\, R_{t+1}] = 0$ при $t < k$. •

REINFORCE

Полученная формула градиента целевой функции содержит математическое ожидание:

$$\mathrm{grad}_\theta J(\pi_\theta) \;=\; \sum_{t=0}^{T-1} \gamma^t \, \mathbb{E}_{\pi_\theta}[G_t \cdot \mathrm{grad}_\theta \, \ln \pi_\theta(A_t \,|\, S_t)].$$

На практике у нас нет всех траекторий сразу, поэтому мы будем генерировать траектории по текущей стратегии π_θ и оценивать градиент целевой функции $\mathrm{grad}_\theta J(\pi_\theta)$ в виде

$$\sum_{t=0}^{T-1} \gamma^t \, G_t \; \mathrm{grad}_\theta \ln \pi_\theta(a_t \,|\, s_t)\,, \quad \text{где } \; G_t \;=\; \sum_{q=t}^{T-1} \gamma^{q-t} \, r_{q+1}, \quad \text{где } a_t,\, s_t,\, r_t \; \text{фактические значения из траектории.}$$

Фактически на каждом шаге обновления параметры сети сдвигаются так, что максимизировать целевую функцию вида

$$J(\tau,\, \theta) \;=\; \sum_{t=0}^{T-1} \gamma^t \, G_t \ln \pi_\theta(a_t \,|\, s_t).$$

REINFORCE

Рассмотрим механизм изменения вероятностей действий при сдвиге параметров в сторону найденного выше градиента для максимизации функции:

$$J(\tau, \theta) = \sum_{t=0}^{T-1} \gamma^t G_t \ln \pi_\theta(a_t \mid s_t).$$

Отметим, что $\pi_\theta(a_t \mid s_t) \in [0, 1]$, а значит $\ln \pi_\theta(a_t \mid s_t) \in (-\infty, 0]$.

Если $G_t > 0$, то при **максимизации** $J(\tau, \theta)$ надо сдвинуть **значение** $\ln \pi_\theta(a_t \mid s_t)$ **к нулю**, а для этого надо сдвинуть **вероятность** $\pi_\theta(a_t \mid s_t)$ к 1, то есть **увеличить**.

Если $G_t < 0$, то при **максимизации** $J(\tau, \theta)$ надо сдвинуть **значение** $\ln \pi_\theta(a_t \mid s_t)$ **в сторону** $-\infty$, а для этого надо сдвинуть **вероятность** $\pi_\theta(a_t \mid s_t)$ к 0, то есть **снизить**.

Иными словами, хорошие действия поощряются, плохие действия выбираются реже.

Основной моделью для приближения стратегий π_θ будет ИНС. Различные фреймворки, как правило, занимаются минимизацией целевой функции. Для реализации метода REINFORCE целевую функцию будет формировать в виде

$$J(\tau, \theta) = -\sum_{t=0}^{T-1} \gamma^t G_t \ln \pi_\theta(a_t \mid s_t).$$

Минус появился из тех соображений, что фреймворки запрограммированы на решение задачи минимизации ошибки, а наша цель в максимизации.

REINFORCE, псевдокод

Приведём псевдокод алгоритма REINFORCE.

1. Инициализация: шаг обучения α, коэффициент обесценивания γ, веса ИНС π_θ
2. Вычисления.

Для каждого эпизода:

Генерация траектории $\tau = \{s_0, a_0, r_1, s_1, a_1, \ldots, r_T, s_T\}$ по стратегии π_θ

grad := 0

Для $k = 0, \ldots, T$:

Найти доход с момента времени k: $G_k = \sum_{t=k}^{T-1} \gamma^{t-k} r_{t+1}$

grad = grad + $\gamma^k \, \text{grad}_\theta \ln \pi_\theta(a_k \mid s_k) \, G_k$

$\theta = \theta + \alpha\,$grad

Полученный метод является методом обучения с единой стратегией, то есть обучение проходит по актуальному опыту. Сгенерированная траектория не может быть сохранена и использована для обучения в будущем. Траектории генерируются по стратегии, а стратегии меняются. Нельзя обновлять веса по старой траектории, так как для новой стратегии вероятность появления такой траектории может быть очень малой.

Особенности REINFORCE

Преимущества:

Основное преимущество policy-based методов (и, в частности, метода REINFORCE) в их простой адаптации к задачам с различным типом действий: дискретные, непрерывные или их смесь.

Кроме того, по методу REINFORCE агент при обучении напрямую достигает своей общей цели — это максимизация ожидаемого дохода.

Недостатки:

Метод REINFORCE не эффективен по работе с имеющимся опытом, поскольку его нельзя повторно использовать.

Оценка градиента по методу REINFORCE хотя и является несмещённой, но эта оценка может иметь большую дисперсию. Действительно, доходы могут сильно отличаться от траектории к траектории в рамках одной стратегии, поскольку действия порождаются стохастической стратегией, начальное состояние может меняться от эпизода к эпизоду и смена состояний может происходить с некоторой вероятностью.

Одним из простых способов снизить дисперсию является нормализация доходов G_t в рамках полученной траектории. То есть для доходов

$$G_0, G_1, \ \ldots, G_{T-1}$$

из траектории находим их среднее μ и дисперсию σ, а затем нормализуем

$$G_i^{\text{new}} = \frac{G_i - \mu}{\sigma}.$$