Policy gradient methods

Семененя Яна, БПМИ-172

Reminder

Простейшая модель обучения с подкреплением состоит из :

- Множество состояний S

- Множество действий A

- Функция вознаграждения(подкрепление) $R : S \times A \to \mathbb{R}$

- Функции перехода между состояниями $p_{ss'}^{a} : S \times A \to \Pi(S)$,

$\Pi(S)$ - множество распределений вероятностей над S

Reminder

- В произвольный момент времени **t** агент характеризуется состоянием $\boldsymbol{s_t}$ и множеством возможных действий $\boldsymbol{A(s_t)}$

- Выбирая действие $\boldsymbol{a \in A(s_t)}$, он переходит в состояние $\boldsymbol{s_{t+1}}$ и получает выигрыш $\boldsymbol{r_t}$.

- Агент должен выработать стратегию $\boldsymbol{\pi: S \rightarrow A}$, которая максимизирует величину :

$$\boldsymbol{R = \sum_t \gamma^t r_t}$$

Мотивация

Игра агента со средой:

- Инициализация стратегии $\pi_1(a|s)$ и состояния среды s_1
- для всех $t = 1 \ldots T$:
  - агент выбирает действие $a_t \sim \pi_t(a|s_t)$
  - среда генерирует награду $r_{t+1} \sim p(r|a_t, s_t)$ и новое состояние $s_{t+1} \sim p(s|a_t, s_t)$
  - агент корректирует стратегию $\pi_{t+1}(a|s)$

Нужно определить алгоритм, который будет искать стратегии, обеспечивающие наибольшую награду.

Policy gradient method

Идея: оптимизировать стратегию $\pi_\theta(a|s)$ напрямую

- Сценарий $\tau = (s_1, a_1, s_2, a_2, \dots, s_T, a_T)$
- Сумма выигрыша в ходе сценария $R_\tau = \sum_\tau r(s_t, a_t)$
- Вероятность реализации сценария:

$$p_\theta(\tau) = p_\theta(s_1, a_1, s_2, a_2, \dots, s_T, a_T) = p(s_1 \prod_{t=1}^{T} \pi_\theta(a_t|s_t) p_\theta(s_{t+1}|s_{t,} a_t))$$

Policy gradient method

Задача : нужно выбрать такой набор параметров агента θ, задающий $\pi_\theta(a|s)$, чтобы максимизировать сумму полученных выигрышей:

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)}[R_\tau] = \int p_\theta(\tau) R_\tau d\tau \to \max$$

Максимизируем выигрыш с помощью градиентного подъема

$$\nabla J(\theta) = \int \nabla p_\theta(\tau) R_\tau d\tau$$

Log derivative trick

Проблема: Не можем подсчитать $\nabla p_\theta(\tau)$ напрямую

Решение:

$$\nabla_\theta p_\theta(\tau) = p_\theta(\tau)\frac{\nabla_\theta p_\theta(\tau)}{p_\theta(\tau)} = p_\theta(\tau)\nabla \log p_\theta(\tau)$$

Тогда заменим $\nabla_\theta p_\theta(\tau)$ на $p_\theta(\tau)\nabla \log p_\theta(\tau)$:

$$\nabla J(\theta) = \int p_\theta(\tau)\nabla \log p_\theta(\tau) R_\tau d\tau$$

Policy gradient method

Рассмотрим $\nabla_\theta \log p_\theta(\tau)$:

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta \log p_\theta(s_1) + \sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t|s_t) + \nabla_\theta \log p(s_{t+1}|s_t, a_t)\right) =$$
$$\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t|s_t)\right)$$

Подставляя в определение $\nabla_\theta J(\theta)$:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)}\left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right) R_\tau\right]$$

Метод Монте-Карло

Если у нас есть выборка из N уже известных сценариев $\tau_i = \left(s_1^i, a_1^i, \ldots, s_T^i, a_T^i\right)$ полученная из распределения $\tau \sim p_\theta(\tau)$, то мы можем приблизить посчитать приблизительное значение $\nabla_\theta J(\theta)$:

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=0}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log\pi_\theta\left(a_t^i \middle| s_t^i\right)\right)R_{\tau^i} =$$

$$\frac{1}{N}\sum_{i=0}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log\pi_\theta\left(a_t^i \middle| s_t^i\right)\right)\left(\sum_{t=1}^{T} r\left(a_t^i, s_t^i\right)\right)$$

REINFORCE

Оптимизицая $J(\theta)$ алгоритмом REINFORCE:

1. Прогнать N сценариев τ_i со стратегией $\pi_\theta(a/s)$
2. Посчитать среднее арифметическое $\nabla_\theta J(\theta)$ по методу Монте-Карло
3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
4. Если не сошлись к экстремуму, повторить с пункта 1.

Недостатки метода REINFORCE

- Для получения всего одного семпла требуется произвести T взаимодействий со средой
- $\nabla_\theta \log p_\theta(\tau)\, R_\tau$ имеет большую дисперсию, поэтому требуется много семплов сценариев
- семплы, собранные для предыдущих значений θ, никак не переиспользуются на следующем шаге

Baseline method

Заметим, что если b - константа относительно τ, то:

$$\mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log p_\theta(\tau)(R_\tau - b)] = \mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log p_\theta(\tau) R_\tau]$$

так как:

$$\mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log p_\theta(\tau) b\,] = \int p_\theta(\tau)\, \nabla_\theta \log p_\theta(\tau) b d\tau = \int \nabla_\theta p_\theta(\tau) b d\tau =$$

$$b\nabla_\theta \int p_\theta(\tau) d\tau = b\nabla_\theta 1 = 0$$

Однако дисперсия $Var_{\tau\sim p_\theta(\tau)}(\nabla_\theta J(\theta))$ зависит b

$$Var_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log p_\theta(\tau)(R_\tau - b)] =$$

$$\mathbb{E}_{\tau\sim p_\theta(\tau)}[(\nabla_\theta \log p_\theta(\tau)(R_\tau - b))^2] - \mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log p_\theta(\tau)(R_\tau - b)]^2$$

Q Actor Critic Method

Градиент суммы полученных выигрышей :

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)}[(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t))R_\tau]$$

Заметим, что:

$$\mathbb{E}_{\tau \sim p_\theta(\tau)}(R_\tau) = Q(s_t, a_t)$$

Где $Q^\pi(s_t, a_t) = \sum_{t'=t}^{T} \mathbb{E}_{\pi_\theta}[r(s_{t'}, a_{t'})|s_t, a_t]$ - оценка будущего выигрыша

Тогда мы можем переписать градиенты суммы выигрышей:

$$\nabla_\theta J(\theta) \propto \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta log \pi_\theta \left(a_t^i | s_t^i\right) Q_{\tau_i, t},$$

Список источников:

- К. А. А. Е. Николенко С., Глубокое обучение. Погружение в мир нейронных сетей, Спб: Питер, 2020, pp. 372-402
- «Методы policy gradient и алгоритм асинхронного актора-критика» [В Интернете].Available:http://neerc.ifmo.ru/wiki/index.php?title=Методы_policy_gradient_и_алгоритм_асинхронного_актора-критика
- Chris Yoon, «Understanding Actor Critic Methods and A2C» 06 02 2019. [В Интернете]. Available: https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f.

Вопросы

1. Описать суть трюка с лог производной (Log-derivative trick)
2. Напишите теорему градиента стратегии, поясните все составляющие. Что обновляется с каждым новым подсчетом градиента?
3. В чем суть модификации градиента по стратегии – baseline? Какие недостатки удается устранить с помощью этой модификации?