Методы policy gradient. Алгоритм Reinforce

Теория игр, 2022

Содержание

Содержание

МППР. Определение

Марковский процесс принятия решений (МППР) используется для описания среды (environment) в том случае, если выполнено следующее свойство (Markov property): процесс зависит только от текущего состояния и не зависит от всей предыдущей истории.

Определение

Марковский процесс принятия решений задается набором следующих элементов $< \mathcal{S}, \mathcal{A}, p >$:

- множество $\mathcal{S}$ состояний среды $s \in \mathcal{S}$
- множество $\mathcal{A}$ доступных действий агента $a \in \mathcal{A}$
- распределение вероятностей

$$p(s', r|s, a) = \Pr[S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a]$$

перехода на шаге t в состояние s' и получения награждения r при условии нахождения в состоянии s и выполнении действия a

Бесконечный МППР

Стратегия

$$\pi(s \mid a) = \mathbb{P}[A_t = a \mid S_t = s]$$

Траектория

Стратегия вместе с моделью МППР задают вероятностное распределение над множеством траекторий $\Omega = \{\tau\}$

$$\tau = (s_0, a_1, r_1, s_1, a_2, \ldots, s_{t-1}, a_t, r_t, s_t, \ldots)$$

$$a_t \sim \pi(\cdot \mid s_t)$$

$$s_{t+1}, r_{t+1} \sim p(\cdot|s_t, a_t)$$

$S_t(\tau) = s_t, A_t(\tau) = a_t, R_t(\tau) = r_t$

Бесконечный МППР

Стратегия

$$\pi(s \mid a) = \mathbb{P}[A_t = a \mid S_t = s]$$

Траектория

$$\tau = (s_0, a_1, r_1, s_1, a_2, \ldots, s_{t-1}, a_t, r_t, s_t, \ldots)$$

Сумма полученных наград

$$G(\tau) = \sum_{k=1}^{\infty} \gamma^k R_k(\tau) = \lim_{t \to \infty} \sum_{k=1}^{t} \gamma^k R_k(\tau)$$

Функция ценности состояния $v^\pi : \mathcal{S} \to \mathbb{R}$

$$v^\pi(s) = \mathop{\mathbb{E}}_{\tau}[G(\tau) | \pi, S_0(\tau) = s]$$

Бесконечный МППР

Стратегия

$$\pi(s \mid a) = \mathbb{P}[A_t = a \mid S_t = s]$$

Траектория

$$\tau = (s_0, a_1, r_1, s_1, a_2, \ldots, s_{t-1}, a_t, r_t, s_t, \ldots)$$

Сумма полученных наград

$$G(\tau) = \sum_{k=1}^{\infty} \gamma^k R_k = \lim_{t \to \infty} \sum_{k=1}^{t} \gamma^k R_k$$

Q функция

$$q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

$$q^\pi(s, a) = \mathop{\mathbb{E}}_{\tau}[G(\tau) | \pi, S^0 = s, A^0 = a]$$

Policy gradient подход

Параметризуемая стратегия

$$\pi : \Theta \to (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$$

$$\theta \mapsto \pi(s \mid a, \theta) \equiv \pi_\theta$$

Максимизация функционала

$$J(\theta) \equiv \mathbb{E}_s \left[v^{\pi_\theta}(s) \right]$$

Policy Gradient Theorem

Выражение для градиента оптимизируемого функционала можно записать следующим образом:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi} \sum_{t \geq 0} \gamma^t \nabla_\theta \log \pi_\theta(A_t \mid S_t) q^\pi(S_t, A_t) \tag{1}$$

Стохастический градиентный спуск

Стохастический градиентный спуск

Управляемое воздействие $x_k \in \mathcal{X}$

Случайная величина $\xi \sim \mathbb{P}_\xi$ (добавляется стохастичность)

На выходе случайная функция, выдаваемая $f(x_k, \xi)$, такая что

$$\mathbb{E}_{\xi \sim \mathbb{P}_\xi}[f(x^k, \xi)] = \nabla_x|_{x=x_k}$$

Обновление $x_{k+1} \leftarrow x_k + \alpha_k f(x_k, \xi_k)$

Policy Gradient Theorem

$$f(x_k, \tau) = \sum_{t \geq 0} \gamma^t \nabla_\theta \log \pi_\theta(A_t \mid S_t) q^\pi(S_t, A_t)$$

$G_t(\tau) = \sum_{k=t}^{T} \gamma^{k-t} R_k$

$$\mathbb{E}_{\tau\sim\pi|s_0=s} \sum_{t\geq 0} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) G_t =$$

$$= \sum_{t\geq 0} \mathbb{E}_{\tau\sim\pi|s_0=s} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) G_t =$$

$$= \sum_{t\geq 0} \mathbb{E}_{a_0,s_1\ldots s_t,a_t} \mathbb{E}_{s_{t+1},a_{t+1}\ldots} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) G_t =$$

$$= \sum_{t\geq 0} \mathbb{E}_{a_0,s_1\ldots s_t,a_t} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) \mathbb{E}_{s_{t+1},a_{t+1}\ldots} G_t =$$

$$= \sum_{t\geq 0} \mathbb{E}_{a_0,s_1\ldots s_t,a_t} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^\pi(s_t, a_t) =$$

$$= \sum_{t\geq 0} \mathbb{E}_{\tau\sim\pi|s_0=s} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^\pi(s_t, a_t) =$$

$$= \mathbb{E}_{\tau\sim\pi|s_0=s} \sum_{t\geq 0} \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^\pi(s_t, a_t)$$

Алгоритм

REINFORCE

Гиперпараметры: N — количество игр, $\pi(a \mid s, \theta)$ — стратегия с параметрами θ, SGD-оптимизатор.

Инициализировать θ произвольно

На очередном шаге t:

1. играем N игр $\tau_1, \tau_2 \dots \tau_N \sim \pi$, $\tau = (s_0, a_0, r_1, s_1, \dots, r_T, s_T)$
2. для каждого t в каждой игре τ считаем reward-to-go: $G_t(\tau) = \sum_{k=t}^{T} \gamma^{k-t} r_k$
3. считаем оценку градиента:

$$\nabla_\theta J(\pi) = \frac{1}{N} \sum_{\tau} \sum_{t \geq 0} \gamma^t \nabla_\theta \log \pi(a_t \mid s_t, \theta) G_t(\tau)$$

4. делаем шаг градиентного подъёма по θ, используя $\nabla_\theta J(\theta)$