Марковский процесс принятия решений. Динамическое программирование

Теория игр, март 2022

Содержание

Содержание

Математическое ожидание

Определение

Математическое ожидание - понятие в теории вероятностей, означающее среднее (взвешенное по вероятностям возможных значений) значение случайной величины

Для дискретной случайной величины

$$\underset{X\sim p_X}{\mathbb{E}}[X] = \sum_i x_i p_X(x_i)$$

Для функции дискретной случайной величины

$$\underset{X\sim p_X}{\mathbb{E}}[f(X)] = \sum_i f(x_i) p_X(x_i)$$

Здесь $p_X(x_i) = \mathbb{P}[X = x_i]$ - функция вероятности дискретной случайной величины X

Принятие решения в условиях неопределенности

Функция полезности $U : \mathcal{O} \to \mathbb{R}$ ставит каждому исходу в соответствие его полезность, где $\mathcal{O} = \{o_1, \ldots, o_m\}$ - множество исходов
Принятие решение в условиях неопределенности характеризуется тем, что каждому решению (действию) a соответствует вероятностное распределение исходов $p(O \mid a) = (\mathbb{P}[o_1 \mid a], \ldots, \mathbb{P}[o_m \mid a])$
$a \succsim a'$ тогда и только тогда

$$\mathop{\mathbb{E}}_{O \sim p(|a)}[U(O) \mid a] \geq \mathop{\mathbb{E}}_{O \sim p(|a')}[U(O) \mid a']$$

$$\sum_i \mathbb{P}[o_i \mid a] U(o_i) \geq \sum_i \mathbb{P}[o_i \mid a'] U(o_i)$$

Содержание

Архитектура задачи последовательного принятия решения

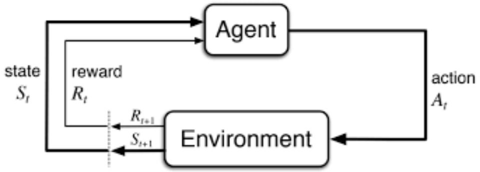


Рис.: Взаимодействие агента и среды

- S_t - состояние (наблюдение состояния) среды в момент t
- R_t - награждение, которое получает агент в момент t
- A_t - действия агента в момент t

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 $< \mathcal{S}, \mathcal{A}, p >$:

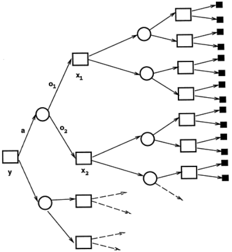
- множество $\mathcal{S}$ состояний среды $s \in \mathcal{S}$
- множество $\mathcal{A}$ доступных действий агента $a \in \mathcal{A}$
- распределение вероятностей

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

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

Эпизодический (конечный) МППР

- Представим в виде дерева конечной глубины T
- Если из состояния s_1 можно перейти в состояние s_2, то из состояния s_2 нельзя перейти в состояние s_1
- Реализация каждого эпизода представляется последовательностью (траекторией), которая имеет вид $\tau = (S_0, A_1, R_1, S_1, A_2, \dots, S_{T-1}, A_T, R_T, S_T)$
- Реализация каждого эпизода численно характеризуется суммой полученных наград $G(\tau) = \sum_{k=1}^{T} R_k$, из-за стохастичности траектории являющейся случайной величиной.



Стратегия агента. Функция ценности состояния

Определение

Функция $\pi : \mathcal{S} \to \mathcal{A}$, которая каждому состоянию s ставит в соответствие действие a, называется стратегией агента

Фиксация определённой стратегии $\pi = (\pi_1, \ldots, \pi_T)$ в МППР позволяет численно оценивать состояние математическим ожиданием суммы $G_t(\tau) = \sum_{k=t}^{T} R_k = R_t + R_{t+1} + \cdots + R_T$ полученных наград траектории, начинающейся на шаге t с оцениваемого состояния $\tau = (S_t = s, A_{t+1}, R_{t+1}, S_{t+1}, \ldots, S_{T-1}, A_T, R_T, S_T)$

Функция ценности состояния V

$$v_t^\pi : \mathcal{S} \to \mathbb{R}$$

$$v_t^\pi(s) = \mathop{\mathbb{E}}_{\tau \sim p}[G_t(\tau) \mid A_{t+1} = \pi_k(S_t), \ldots, A_T = \pi_k(S_{T-1}), S^0 = s]$$

Функция ценности состояния (продолжение)

Определение

Функция $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$,

$$R(s, a, s') = \underset{(s', r) \sim p(|s,a)}{\mathbb{E}} [R^{t+1} \mid S^t = s, A^t = a, S^{t+1} = s']$$
$$= \sum_{r \in \mathbb{R}} r \frac{p(s', r | s, a)}{\sum_{s \in \mathcal{S}} p(s', r \mid s, a)}$$

называется функцией награды

Так как зафиксировав π, мы получаем марковскую цепь, то функцию ценности можно как мат ожидание функции награды на множестве всех возможных последовательностей состояний начиная с $S^0 = s$

$$v^{\pi}(s) = \underset{S^1, \ldots, S^T}{\mathbb{E}} \left[\sum_{k=0}^{T} R(S^k, \pi(S^k), S^{k+1}) \right]$$

Оптимальная функция ценности состояния

Определение

Оптимальной стратегией называется стратегия π^*, *для которой для каждого* $s \in \mathcal{S}$ *и для любой* π *верно*

$$v^{\pi^*}(s) \geq v^{\pi}(s)$$

Оптимальных стратегий может быть несколько, но всем им соответствует оптимальная функция ценности

Определение

Оптимальной функцией ценности называется функция

$$v^*(s) = \max_{\pi_1,\dots,\pi_T} v^{\pi}(s)$$

для каждого $s \in \mathcal{S}$

Содержание

Задача оптимизации в МППР

Постановка задачи

Нахождение оптимальной стратегии $\pi^* : \mathcal{S} \to \mathcal{A}$

достигается решением

Постановка задачи

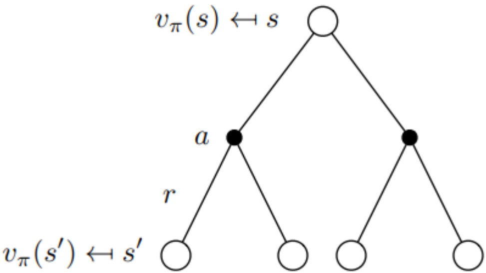
Нахождение оптимальной функции ценности $v^* : \mathcal{S} \to \mathbb{R}$

Нахождение оптимальной функции ценности

Для конечного МППР может быть решена методом обратной индукции (метод динамического программирования)

уравнение оптимальности Беллмана

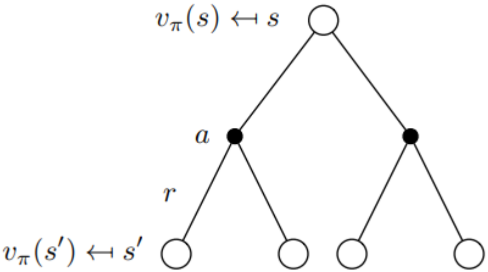
$$v^*_{t+1}(s) = \max_a \mathop{\mathbb{E}}_{(s',r)\sim p(|s,a)} [r + v^*_t(s')] = \max_a \sum_{(s',r)} p(s', r \mid s, a)\Big(r + v^*_t(s')\Big)$$



Нахождение оптимальной стратегии

Оптимальная стратегия

$$\pi^*_{t+1}(s) = \operatorname*{argmax}_a \sum_{(s',r)} p(s', r \mid s, a)\Big(r + v^*_t(s')\Big)$$



Функция, ставящая в соответствие состоянию s и возможному действию a в данном состоянии математическое ожидание суммы $G(\tau) = \sum_{k=1}^{T} R_k$ полученных наград траектории, начинающейся с оцениваемого состояния и действия
$\tau = (S^t = s, A^{t+1} = a, R^{t+1}, S^{t+1}, \dots, S^{T-1}, A^T, R^T, S^T)$

Q функция

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

$$q_t^\pi(s, a) =$$
$$\mathop{\mathbb{E}}_{\tau \sim p}[G_t(\tau) \mid \pi, S^t = s, A^t = a, A_{t+1} = \pi_k(S_t), \dots, A_T = \pi_k(S_{T-1})]$$

Q функция и функция ценности

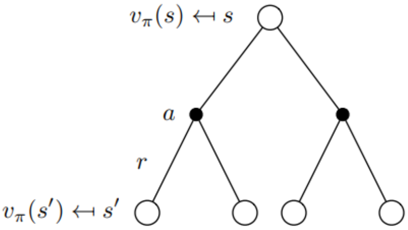
Для любого $s \in \mathcal{S}$

$$v^*_{t+1}(s) = \max_a q^*_{t+1}(s, a)$$

$$q^*_{t+1}(s, a) = \mathop{\mathbb{E}}_{(s',r)\sim p(|s,a)} [r + v^*_t(s')]$$

$$= \sum_{(s',r)} p(s', r \mid s, a)\Big(r + v^*_t(s')\Big)$$

$$\pi^*_{t+1}(s) = \operatorname*{argmax}_a q^*_{t+1}(s, a)$$



Нахождение оптимальной функции ценности для конечного МППР

Algorithm 1 Алгоритм динамического программирования

procedure ДП
    Input: МППР $< \mathcal{S}, \mathcal{A}, P(s', r \mid s, a) >$
    Output: v^*
    $v^0(s) \leftarrow 0$ для каждого $s \in \mathcal{S}$
    for $i = 1, 2, \ldots, T$ **do**
        $v^{k+1} \leftarrow \max \mathbb{E}_{(s',r)\sim p(|s,a)}[r + v^k(s')]$
        где $\mathbb{E}_{(s',r)\sim p(|s,a)}[r + v^k(s')] = \sum_{(s',r)} p(s', r \mid s, a)\left(r + v^k(s')\right)$
    end for
    Return: v^T
end procedure

Содержание

Бесконечный МППР

Траектория имеет вид $\tau = (S^0, A^1, R^1, S^1, A^2, \dots, S^{t-1}, A^t, R^t, S^t, \dots)$
Сумма полученных наград

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

для $\gamma < 1$ (условие необходимо для сходимости ряда)
Стратегия стационарна (не зависит от шага t) и имеет вид $\pi = (\pi, \pi, \dots)$

Функция ценности состояния V

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

$$v^\pi(s) = \mathop{\mathbb{E}}_{\tau\sim p}[G(\tau)|\pi, S^0 = s]$$

Определение

Оптимальной функцией ценности называется функция

$$v^*(s) = \max_{\pi} v^{\pi}(s)$$

для каждого $s \in \mathcal{S}$

Q функция

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

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

Нахождение оптимальной функции ценности

Для бесконечного МППР может быть решена методом value iteration

уравнение оптимальности Беллмана

На множестве функций $v : \mathcal{S} \to \mathbb{R}$ зададим отображение

$$T(v)(s) = \max_a \sum_{(s',r)} p(s', r \mid s, a)\Big(r + \gamma v(s')\Big)$$

Тогда

$$v* = T(v^*) = \lim_{k\to\infty} T^k(v) = \lim_{k\to\infty} T(T(\dots T(v)))$$

При этом для любых v, v'

$$||T(v) - T(v')|| \leq ||v - v'||$$

Нахождение оптимальной функции ценности для бесконечного МППР

Algorithm 2 Value iteration

procedure ДП
Input: МППР $< \mathcal{S}, \mathcal{A}, P(s', r \mid s, a), \gamma >$, требуемая точность решения ϵ
Output: v^*
$v^0(s) \leftarrow 0$ для каждого $s \in \mathcal{S}$
repeat
$v^{k-1} \leftarrow v^k$
$v^k \leftarrow \sum_{(s',r)} p(s', r \mid s, a)\left(r + \gamma v^{k-1}(s')\right)$
until $||v^k - v^{k-1}|| \leq \epsilon$
Return:v^k
end procedure