
Конспект по обучению с подкреплением в Dec-POMDP

james116blue@gmail.com

11 июня 2023 г.

Аннотация

Аннотация

Предполагается знакомство читателя с основами машинного обучения и глубокого обучения. Об ошибках и опечатках в тексте можно сообщать в [репозитории проекта](#).

Оглавление

1	Основные используемые определения	3
1.1	Основные определения из RL	3
1.2	Частично наблюдаемый MDP	4
1.3	Стохастическая игра	7
A	Приложение	8

Основные используемые определения

В данной главе будут введены основные определения и описана формальная постановка задачи. Под желаемым результатом мы далее будем понимать максимизацию некоторой скалярной величины, называемой **наградой** (reward). Интеллектуальную сущность (систему/робота/алгоритм), принимающую решения, будем называть **агентом** (agent). Агент взаимодействует с **средой** (environment), которая задаётся зависящим от времени **состоянием** (state). Агенту в каждый момент времени в общем случае доступно только некоторое **наблюдение** (observation) текущего состояния мира. Сам агент задаёт процедуру выбора **действия** (action) по доступным наблюдениям; эту процедуру далее будем называть **стратегией** (policy). Процесс взаимодействия агента и среды задаётся **динамикой среды** (world dynamics), определяющей правила смены состояний среды во времени и генерации награды.

Буквы s , a , r зарезервируем для состояний, действий и наград соответственно; буквой t будем обозначать время в процессе взаимодействия.

§1.1. Основные определения из RL

Средой (environment) называется тройка $(\mathcal{S}, \mathcal{A}, \mathcal{P})$, где:

- \mathcal{S} — **пространство состояний** (state space), некоторое множество.
- \mathcal{A} — **пространство действий** (action space), некоторое множество.
- \mathcal{P} — **функция переходов** (transition function) или **динамика среды** (world dynamics): вероятности $p(s' | s, a)$.

Набор $\mathcal{T} := (s_0, a_0, s_1, a_1, s_2, a_2, s_3, a_3 \dots)$ называется **траекторией**.

Стратегия(политика) - распределение $\pi(a | s)$, $a \in \mathcal{A}$, $s \in \mathcal{S}$.

Для данной среды, политики π и начального состояния $s_0 \in \mathcal{S}$ распределение, из которого приходят траектории \mathcal{T} , называется **trajectory distribution**:

$$p(\mathcal{T}) = p(a_0, s_1, a_1 \dots) = \prod_{t \geq 0} \pi(a_t | s_t) p(s_{t+1} | s_t, a_t)$$
$$\mathbb{E}_{\mathcal{T}}(\cdot) := \mathbb{E}_{\pi(a_0 | s_0)} \mathbb{E}_{p(s_1 | s_0, a_0)} \mathbb{E}_{\pi(a_1 | s_1)} \dots (\cdot) \quad (1.1)$$

Поскольку часто придётся раскладывать эту цепочку, договоримся о следующем сокращении:

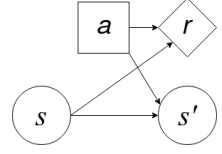
$$\mathbb{E}_{\mathcal{T}}(\cdot) = \mathbb{E}_{a_0} \mathbb{E}_{s_1} \mathbb{E}_{a_1} \dots (\cdot)$$

Однако в такой записи стоит помнить, что действия приходят из некоторой зафиксированной политики π , которая неявно присутствует в выражении. Для напоминания об этом будет, где уместно, использоваться запись $\mathbb{E}_{\mathcal{T} \sim \pi}$.

Марковский процесс принятия решений (Markov Decision Process, MDP)

— это четвёрка $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, где:

- $\mathcal{S}, \mathcal{A}, \mathcal{P}$ — среда.
- $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ — **функция награды** (reward function).



Дисконтированной кумулятивной наградой (discounted cumulative reward) или **total return** для траектории \mathcal{T} с коэффициентом $\gamma \in (0, 1]$ называется

$$R(\mathcal{T}) := \sum_{t \geq 0} \gamma^t r(s_t, a_t) \quad (1.2)$$

Для траектории \mathcal{T} величина

$$R_t := R(\mathcal{T}_{t:}) = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} r_{\hat{t}} \quad (1.3)$$

называется **reward-to-go** с момента времени t .

Определение 1: **Скором** (score или performance) стратегии π в данном MDP называется

$$J(\pi) := \mathbb{E}_{\mathcal{T} \sim \pi} R(\mathcal{T}) \quad (1.4)$$

Итак, задачей обучения с подкреплением является оптимизация для заданного MDP средней дисконтированной кумулятивной награды:

$$J(\pi) \rightarrow \max_{\pi}$$

Определение 2: Для данного MDP **V-функцией** (value function) или оценочной функцией состояний (state value function) для данной стратегии π называется величина

$$V^{\pi}(s) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s} R(\mathcal{T}) \quad (1.5)$$

Определение 3: Для данного MDP **Q-функцией** (state-action value function, action quality function) для данной стратегии π называется

$$Q^{\pi}(s, a) := \mathbb{E}_{\mathcal{T} \sim \pi | s_0 = s, a_0 = a} \sum_{t \geq 0} \gamma^t r_t$$

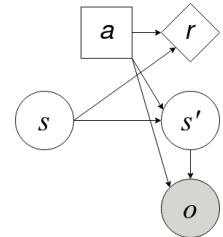
Определение 4: Для данного MDP **Advantage-функцией** политики π называется

$$A^{\pi}(s, a) := Q^{\pi}(s, a) - V^{\pi}(s) \quad (1.6)$$

§1.2. Частично наблюдаемый MDP

Частично наблюдаемые MDP являются математическим инструментом для моделирования процесса принятия решения в ситуациях, когда результат зависит как от случайности, заложенной в самой среде, так и от отсутствия агента полной информации об этой среде.

MDP называется **частично наблюдаемым** (partially observable, принятое сокращение — PoMDP), если дополнительно задано множество \mathcal{O} , называемое **пространством наблюдений** (observation space), и распределение $p(o | s', a)$, определяющая вероятность получить то или иное наблюдение агента $o \in \mathcal{O}$ в момент времени, когда мир находится в состоянии $s' \in \mathcal{S}$, в которое он попал при выполнении агентом действия a .



- $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ — Марковский процесс принятия решений.
- \mathcal{O} — пространство наблюдений.
- $p(o | s', a)$ — **распределение наблюдений** (observation function).

Также как и для случая наблюдаемого MDP, задачей обучения с подкреплением в PoMDP является

$$\mathbb{E}_{\mathcal{T} \sim \pi} \sum_{t \geq 0} \gamma^t r(s_t, a_t) \rightarrow \max_{\pi}$$

Препятствием для использования классических методов RL к POMDP является то, что агенту неизвестно состояние среды s , на основе которого в случае наблюдаемого MDP агент получал бы распределение вероятностей по действиям в $\pi(a | s)$. Состояние s важно тем, что в оптимизируемом функционале (1.4) функция r для

каждого момента времени t помимо действия a_t зависит именно от s_t , а сумма функций для всех последующих моментов времени $\hat{t} \geq t$ зависит от соответствующих состояний $s_{\hat{t}}$, вероятность попадания которых также определяется состоянием s_t .

Возможным решением здесь тогда будет переход от построения стратегии на основе состояния к стратегии на основе некоторой информации, известной нам о состоянии. Это может быть достигнуто применением Байесовской статистики — теории в области статистики, основанной на Байесовской интерпретации вероятности, когда вероятность отражает степень доверия событию, которая может измениться, когда будет собрана новая информация. В этом случае PoMDP рассматривается в виде графовой вероятностной модели — Байесовской сети (Bayesian Networks), где каждой вершине ориентированного графа соответствует случайная переменная, а дуге — зависимость между этими переменными.

Тогда состояние s заменяется распределением вероятностей по s , отражающим степень доверия к состоянию — *belief*. Так в случае конечного пространства состояний \mathcal{S} скалярное значение состояния s заменяется на вектор размерности $|\mathcal{S}|$.

Следует отметить, что указанное распределение вероятностей является обусловленным всей уже известной агенту информацией — последовательностью его наблюдений и действий, называемой историей агента (action-observation history).

Определение 5: Последовательность наблюдений и действий агента до момента t называется историей агента $h_t = (o_0, a_0, o_1, a_1, \dots, a_{t-1}, o_t)$

Фактически история является "обрезанной" до момента t траекторией, но ввиду важности ее использования в стратегиях для PoMDP имеет смысл использовать отдельное определение и обозначение. История может задаваться рекурсивно следующим выражением

$$\begin{cases} h_0 = o_0 \\ h_{t+1} = \langle h_t, a_t, o_{t+1} \rangle \end{cases} \quad (1.7)$$

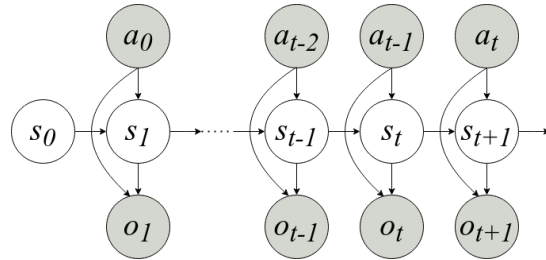
Определение 6: Вероятность пребывания в том или ином состоянии при условии наблюдаемой истории называется *belief state*

$$b_t = p(s_t | h_t)$$

Рассмотрение процесса в виде Байесовской сети, относящейся без учета функции награды к расширению скрытой марковской цепи — input output HMM, и анализа условной независимости, представленной здесь графическим свойством d-разделённости (d-separation), позволяет обосновать недостаточность использования только последнего наблюдения в качестве входного значения стратегии.

Для этого необходимо выделить три множества случайных величин и соответствующих им вершин графа:

- $A = \{s_t\}$
- $B = h_{t-1} = \{a_0, o_1, \dots, a_{t-2}, o_{t-1}\}$
- $C = \{a_{t-1}, o_t\}$



Тогда условие достаточности последнего наблюдения для принятия решения a_t можно описать своими словами следующим образом: знание о предыдущей истории h_{t-1} не добавит новой информации о распределении $p(s_t)$ помимо того, что уже известно на основе наблюдения o_t и предыдущего действия a_{t-1} . Более формально можно выразить с использованием теории информации через понятие относительной взаимной информации, которая определяет, насколько изменится условная энтропия состояния s_t

$$\begin{aligned} I(s_t; h_{t-1} | a_{t-1}, o_t) &:= H(s_t | a_{t-1}, o_t) - H(s_t | h_{t-1}, a_{t-1}, o_t) \\ &= D_{KL}(p(s_t, h_{t-1} | a_{t-1}, o_t) \| p(s_t | a_{t-1}, o_t)p(h_{t-1} | a_{t-1}, o_t)) \end{aligned}$$

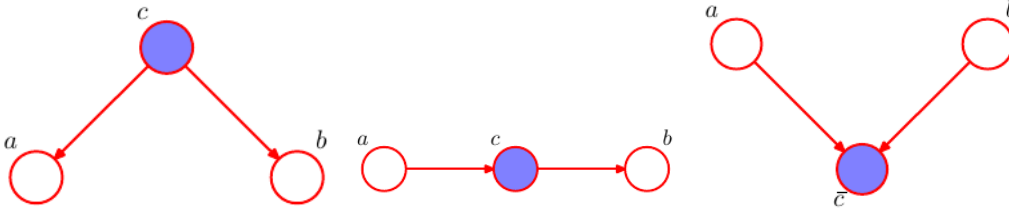
В соответствии со свойствами расстояния Кульбака — Лейблера, выражение ?? и указанная в нем относительная взаимная информация равно нулю тогда для всех $s_t, h_{t-1}, a_{t-1}, o_t$ верно $p(s_t, h_{t-1} | a_{t-1}, o_t) = p(s_t | a_{t-1}, o_t)p(h_{t-1} | a_{t-1}, o_t)$, что равнозначно

$$p(s_t | h_{t-1}, a_{t-1}, o_t) = p(s_t | a_{t-1}, o_t), \quad (1.8)$$

или выражая через множества случайных величин, $p(A | B, C) = p(A | C)$, что обозначается понятием условной независимости наборов случайных величин A и B по набору C . Это в свою очередь выполняется, если в соответствующем графе множество вершин C разделяет A и B , для чего множество вершин C должно

блокировать все пути из любой вершины, принадлежащей \mathbf{A} в любую вершину, принадлежащую \mathbf{B} (путь рассматривается для соответствующего неориентированного графа). Блокированием пути p множеством вершин \mathbf{C} называется выполнение одного из следующих условий:

- p содержит цепь $a \rightarrow c \rightarrow b$ или разветвление $a \leftarrow c \rightarrow b$ такие, что c принадлежит \mathbf{C} ; или
- p содержит инвертированное разветвление $a \rightarrow \bar{c} \leftarrow b$, такое, что ни \bar{c} , ни ее потомок не принадлежит \mathbf{C} .



Графовая модель PoMDP позволяет сделать вывод о невыполнении условий d-разделенности \mathbf{A} и \mathbf{B} множеством вершин \mathbf{C} (ни один из путей из \mathbf{A} в \mathbf{B} — например, $s_t \leftarrow s_{t-1} \rightarrow o_{t-1}$ — не блокируется множеством \mathbf{C}), а, следовательно, и том, что равенство (1.8) в общем случае неверно.

Встает вопрос — возможно ли вместо оценки доверия по всей истории рекурсивно ее обновлять подобно обновлению истории (1.7)? В Байесовской статистике для обновления вероятностей оцениваемого параметра θ , являющихся, как было отмечено выше, степенью доверия, после получения новых данных \mathcal{D} используется теорема Байеса

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D} | \theta)p(\theta)$$

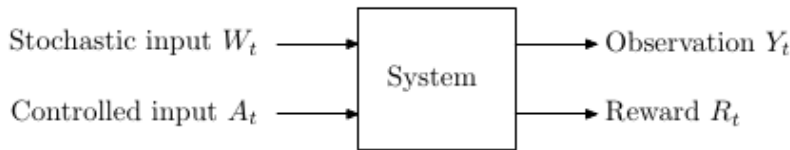
. На основе указанной теоремы строится процедура рекурсивной Байесовской оценки (Recursive Bayesian estimation, Recursive Bayesian filter), позволяющая обновлять значение belief state с точностью до нормализующей константы:

$$\begin{aligned} b_{t+1} &= p(s_{t+1} | h_{t+1}) \\ &= p(s_{t+1} | o_{t+1}, a_t, h_t,) \\ &= \{ \text{здесь оцениваемым параметром } \theta \text{ является } s_{t+1}, \text{ а новыми данными } \mathcal{D} \text{ является } o_{t+1} \} = \\ &\propto p(o_{t+1} | s_{t+1}, a_t, h_t) p(s_{t+1} | a_t, h_t) = \\ &= \{ \text{по марковости MDP и формуле полной вероятности} \} = \\ &= p(o_{t+1} | s_{t+1}, a_t) \left[\sum_{s_t} p(s_{t+1} | s_t, a_t, h_t) p(s_t | a_t, h_t) \right] = \\ &= \{ \text{состояние не зависит от действия, стоящего в траектории после состояния } p(s_t | a_t, h_t) = p(s_t | h_t) = b_t \} = \\ &= p(o_{t+1} | s_{t+1}, a_t) \left[\sum_{s_t} p(s_{t+1} | s_t, a_t) b_t \right] \end{aligned}$$

То, что belief state является аналогом состояния s наблюдаемого MDP, можно понять, попробовав обобщить понятие состояние. Для рассмотрения генезиса понятия состояния следует рассмотреть анализируемый процесс как процесс, порождаемый стохастической системой "вход/выход". Модель такой системы представляет из себя черный ящик (black box), который в каждый момент времени t в ответ на контролируемое воздействие (вход) \mathbf{A}_t генерирует наблюдение \mathbf{Y}_t (выход) и награду \mathbf{R}_t и вообще не обладает состоянием. Стохастичность зависимости между входом и выходом обеспечивается случайным входным возмущением \mathbf{W}_t .

$$\mathbf{Y}_t = f_t(\mathbf{A}_{1:t}, \mathbf{W}_{1:t})$$

$$\mathbf{R}_t = r_t(\mathbf{A}_{1:t}, \mathbf{W}_{1:t})$$



МОЖНО

Для PoMDP стратегия имеет вид $\pi(a_t | h_t), a \in \mathcal{A}, o \in \mathcal{O}$.

§1.3. Стохастическая игра

ГЛАВА А

Приложение
