Введение в обучение с подкреплением

Михненко Наталья БПМИ182

Reinforcement learning

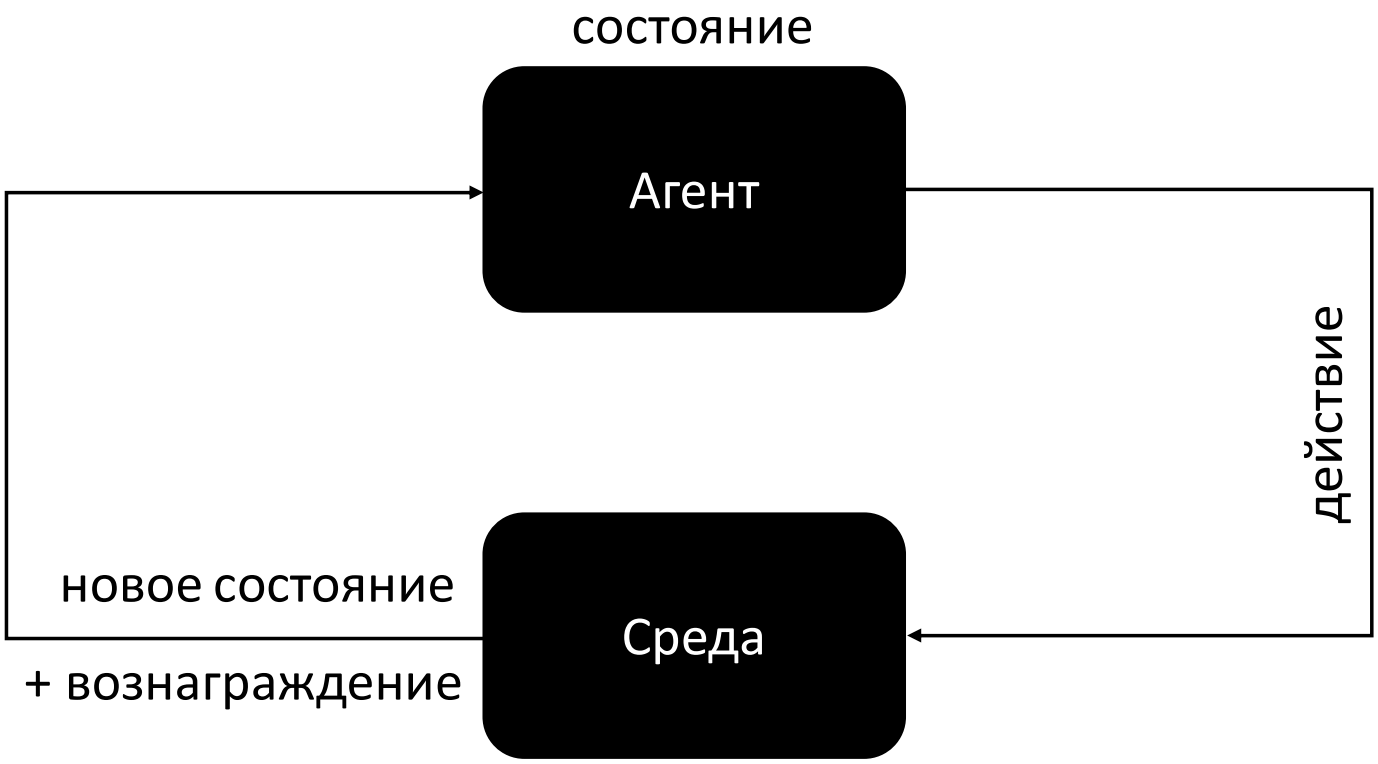
Обучение с подкреплением – это способ машинного обучения, при котором система обучается, взаимодействуя с некоторой средой.

- Обучается в процессе взаимодействия со средой, без исторических данных
- Основная идея – поощрать действий, ведущие к награде, избегать ведущих к неудаче

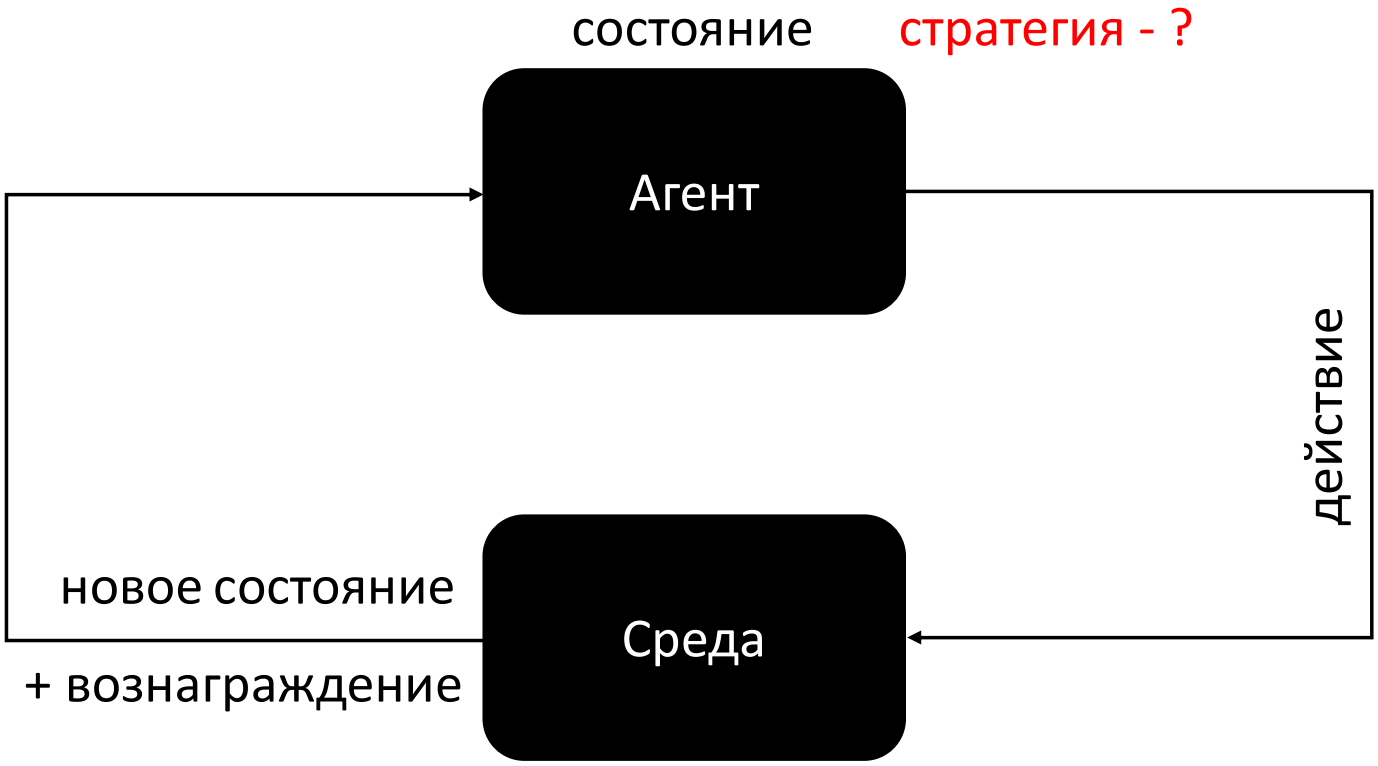
Примеры прикладных задач

- Стратегические игры: шахматы, Go, дота
- Управление роботами
- Управление ценами и ассортиментом в сетях продаж
- Создание чат ботов
- Обучение трейдинговых ботов
- Обучение беспилотников

Постановка задачи



Постановка задачи



Цель агента – максимизировать суммарное вознаграждение

Постановка задачи

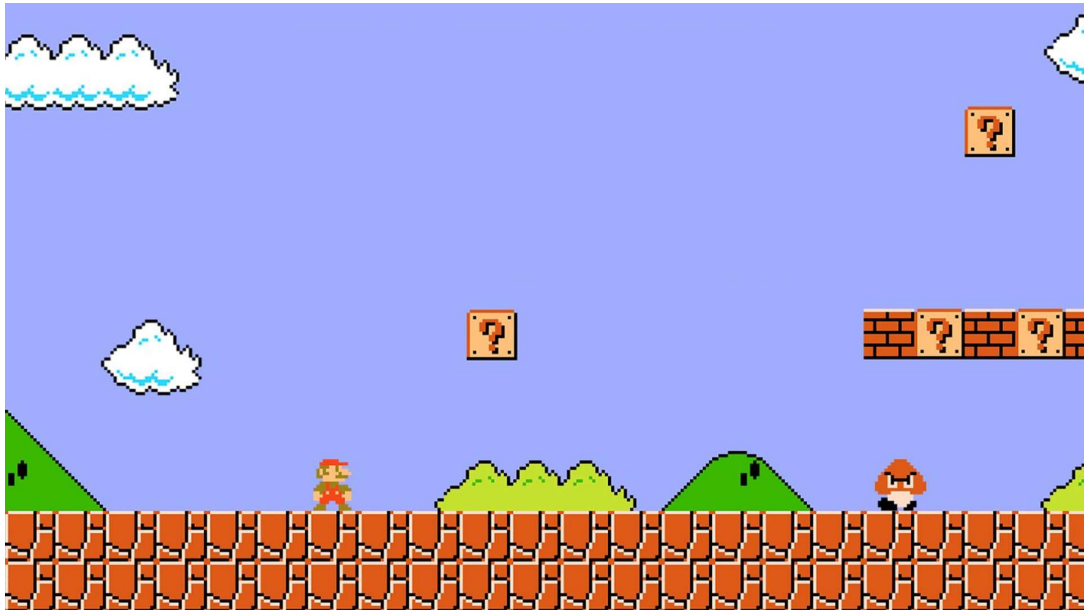
1. На каждом шаге среда находится в некотором состоянии s ∈ S
2. На каждом шаге агент выбирает из имеющегося набора действий a ∈ A согласно некоторой стратегии π
3. Окружающая среда сообщает какое вознаграждение r получил агент и новое состояние среды s* ∈ S
4. Агент корректирует стратегию π

Задача агента выработать стратегию, максимизирующую R = $\sum_t \gamma^t r_t$

Типы задач

Эпизодические

В этом случае у нас есть начальная точка и конечная (конечное состояние). Это создает эпизод: список состояний, действий, вознаграждений и новых состояний.



Типы задач

Непрерывные

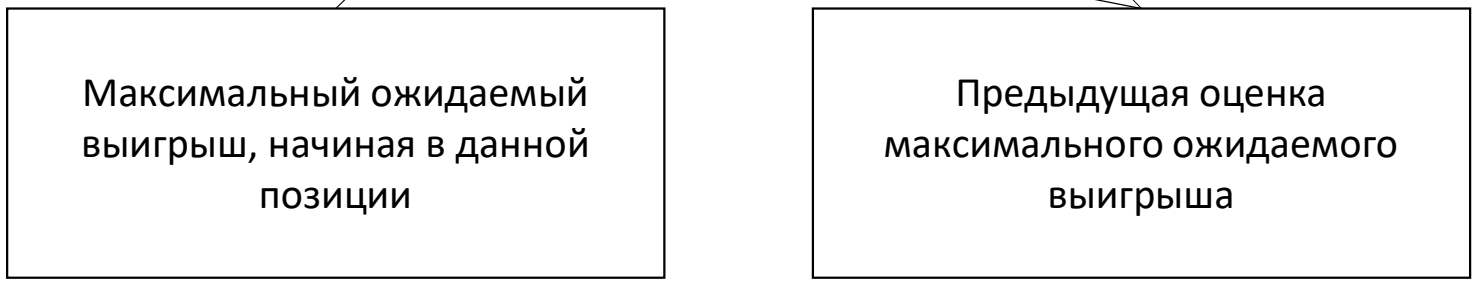
Это задачи, которые продолжаются вечно (без состояния окончания). В этом случае агент должен научиться выбирать лучшие действия и одновременно взаимодействовать со средой.

Методы обучения

Monte Carlo

Когда эпизод заканчивается, агент смотрит на общее накопленное вознаграждение чтобы увидеть, насколько хорошо он прошел этот эпизод. Таким образом, награды получают только в конце игры.

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$$

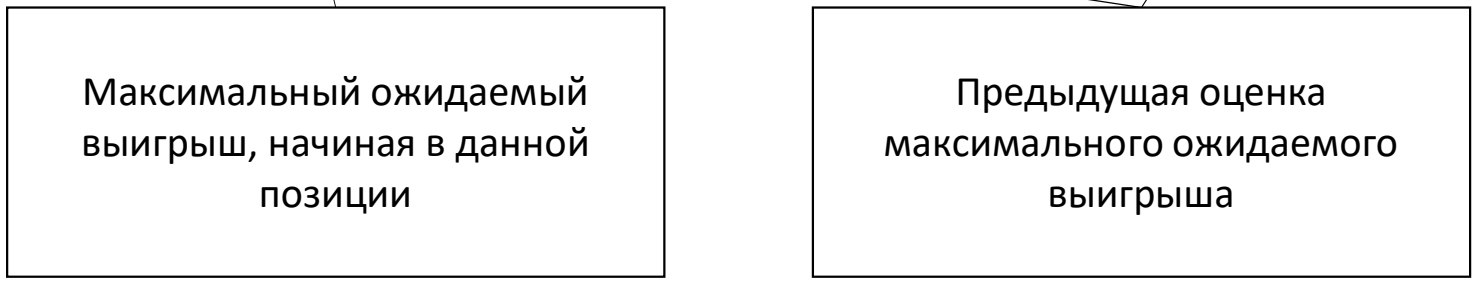


Методы обучения

Temporal Difference Learning

TD Learning не будет ждать конца эпизода, чтобы обновить максимальную ожидаемую оценку вознаграждения: он будет обновлять свою оценку значения V на каждом шаге t.

$$V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$$



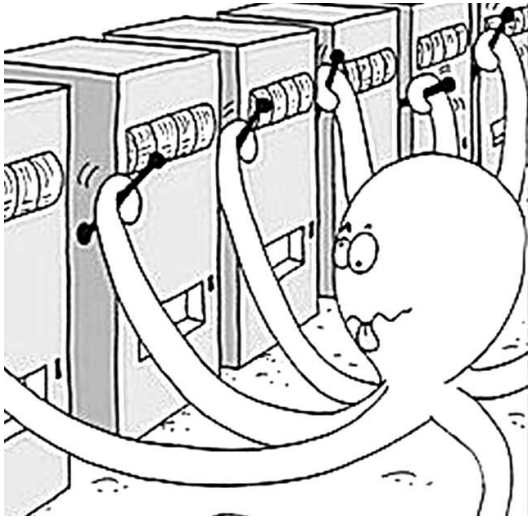
Жадный алгоритм

Многорукий бандит

A – множество возможных действий (ручек)
$p(r|a)$ – неизвестное распределение вознаграждения r ∈ R для a ∈ A
π_t - стратегия агента в момент времени t

Взаимодействие агента со средой в момент времени t:
1. Агент выбирает действие a_t
2. Среда генерирует вознаграждение r_t
3. Агент корректирует стратегию

$Q_t(a) = \frac{\sum_{i=1}^{t} r_i[a_i=a]}{\sum_{i=1}^{t}[a_i=a]}$ - средняя премия за t раундов



Жадный алгоритм

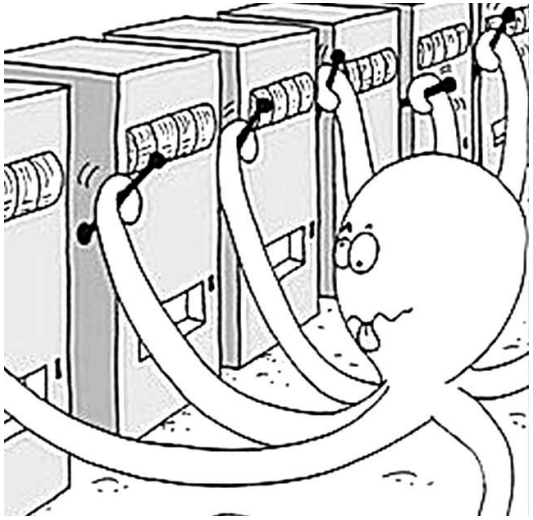
Многорукий бандит

$Q_t(a) = \frac{\sum_{i=1}^{t} r_i[a_i=a]}{\sum_{i=1}^{t}[a_i=a]}$ - средняя премия за t раундов

$\mathrm{A}_t = Argmax_{a \in A} Q_t(a)$ – множество действий с текущей максимальной ценностью

Жадная стратегия – выбирать любое действие из A_t

Недостаток – по некоторым действиям можно не набрать статистику



ε-жадный алгоритм

$\varepsilon \in [0; 1]$

random_number $\in [0; 1]$

```
If random_number < ε:
   explore()
else:
   exploit()
```

- Со временем предлагается уменьшать ε
- Достигается компромисс “изучение - применение”

Табличный метод

	Действие a1	Действие a2	Действие a3	…
Состояние s1	Q(s1, a1)	Q(s1, a2)	Q(s1, a3)	
Состояние s2	Q(s2, a1)	Q(s2, a2)	Q(s2, a3)	
Состояние s3	Q(s3, a1)	Q(s3, a2)	Q(s3, a3)	
…				

Недостатки:

1. Храним много данных
2. Подходит только для эпизодических задач

Метод кросс-энтропии

repeat:

- провести N эпизодов испытаний
- выбрать M лучших эпизодов (Elite)
- поменять стратегию, отдавая приоритет действиям из лучших эпизодов

$$\pi(a|S) = \frac{\sum_{s_t,a_t \in Elite}[s_t=S][a_t=a]}{\sum_{s_t \in Elite}[s_t=S]}$$ - пересчет стратегии

Метод кросс-энтропии

Проблема: редкие состояния

Решение: сглаживание

$$\pi(a|S) = \frac{\sum_{s_t, a_t \in Elite} [s_t = S][a_t = a] + \lambda}{\sum_{s_t \in Elite} [s_t = S] + \lambda N}$$ - пересчет стратегии

Метод кросс-энтропии

Проблема: стохастические выигрыши

Решение: сэмплировать действия для каждого состояния и усреднить результат

Некоторые итоги

- В обучении с подкреплением нет правильных ответов, есть только реакция среды
- Для большей части задач не работают табличные методы
- Нужно использовать алгоритм обучения с временными воздействиями
- Компромисс изучение/применение нужно подбирать экспериментально

Источники

https://coursera.org/share/c7bb0ac37e0b179b91ad96a4d5953bd8 – отличный курс, доклад по неделе 1

https://www.coursera.org/lecture/practical-rl/crossentropy-method-TAT8g - конкретно про кросс-энтропию

https://datascience.org.ua/vvedenie-v-reinforcement-learning-ili-obuchenie-s-podkrepleniem - методы обучения

http://www.machinelearning.ru/wiki/images/3/35/Voron-ML-RL-slides.pdf - многорукий бандит