Лекция 8. Обучение с подкреплением

2024/2025 учебный год

Доцент кафедры ИВЭ, Махно В.В.

©Создано при помощи https://sberuniversity.ru/

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК
имени И.И. Воровича

Обучение с подкреплением

Общая высокоуровневая схема в задачах машинного обучения выглядит так:

Сбор данных → Обучение алгоритма по данным → Применение готового алгоритма.

Приведенная схема позволяет сначала внимательно обработать данные, затем аккуратно построить алгоритм на основе данных, и затем оценить качество работы алгоритма с использованием понятных метрик. Однако есть ряд задач, в которых такая схема с тремя отдельными шагами неэффективна, и сбор данных и обучение происходит прямо в процессе применения алгоритма. Область искусственного интеллекта, изучающая такие задачи, называется **обучением с подкреплением**

Общая схема обучения с подкреплением

В обучении с подкреплением алгоритм обучается не на основе заранее собранных данных, а в процессе взаимодействия со средой. Сам алгоритм принято называть агентом. Примеры агентов и сред:

- Робототехника: агент — робот, среда — окружающий мир.
- Компьютерные игры: агент — игрок, среда — сама игра.
- Алгоритмический трейдинг: агент — трейдер (алгоритм), среда — биржа.

Схема обучения с подкреплением

Схема обучения с подкреплением состоит из повторяющихся шагов:

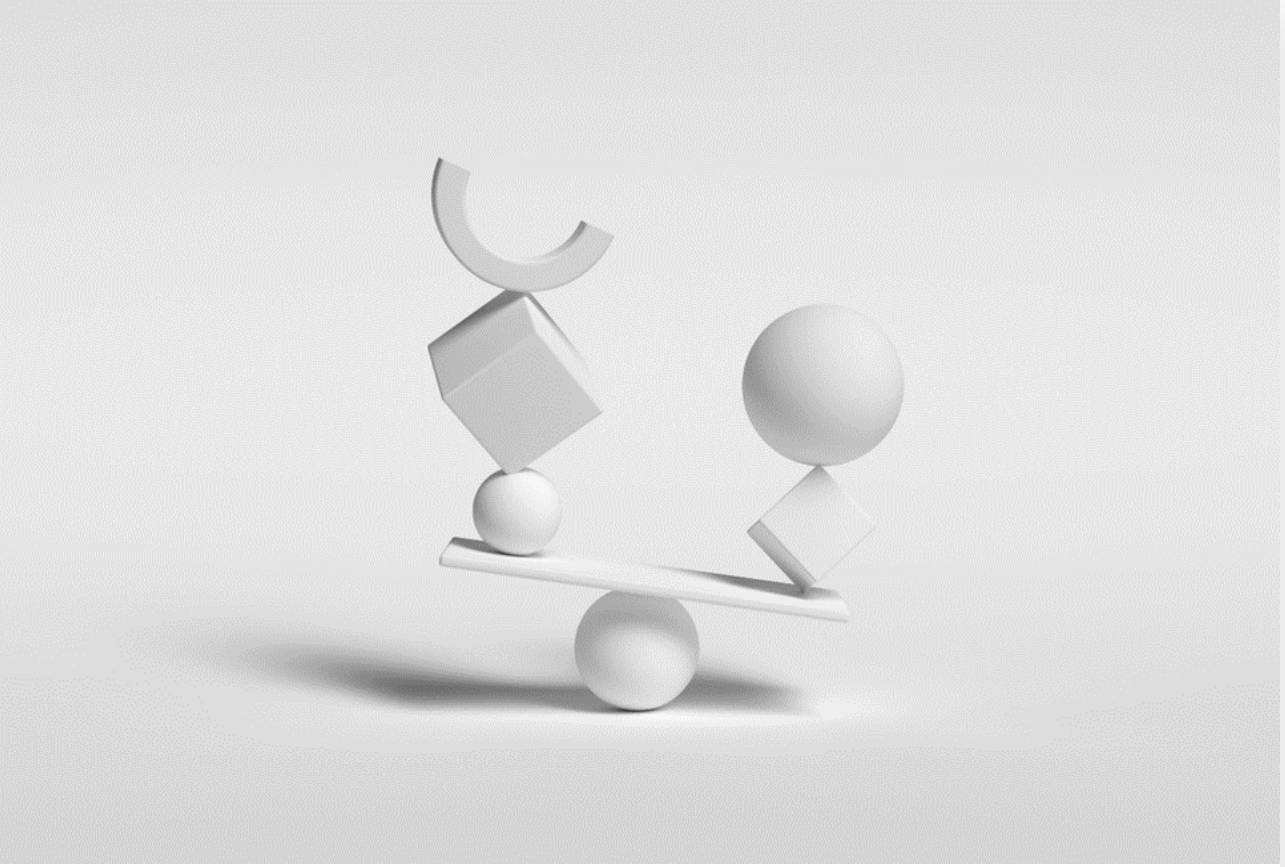
- агент выполняет действие со средой;
- среда генерирует награду и (необязательно) обновляет состояние;
- агент корректирует стратегию.

Кумулятивная награда

Важной особенностью обучения с подкреплением является наличие отложенных наград: некоторые действия могут быть невыгодными в краткосрочной перспективе, но приносить большой выигрыш в долгосрочной перспективе.

- В обучении с подкреплением обычно используют **кумулятивную награду**: разрабатывают стратегии, которые оптимизируют ожидаемую сумму всех последующих наград, а не только следующую.

- Часто также используют **дисконтирование**: при расчете кумулятивной награды далекую награду включают с меньшим весом, чем близкую. Дисконтирование будто бы комбинирует оба подхода: учитывает будущие награды, но уделяет больше внимания ближайшим наградам.



Баланс исследования-использования (exploration-exploitation trade-off)

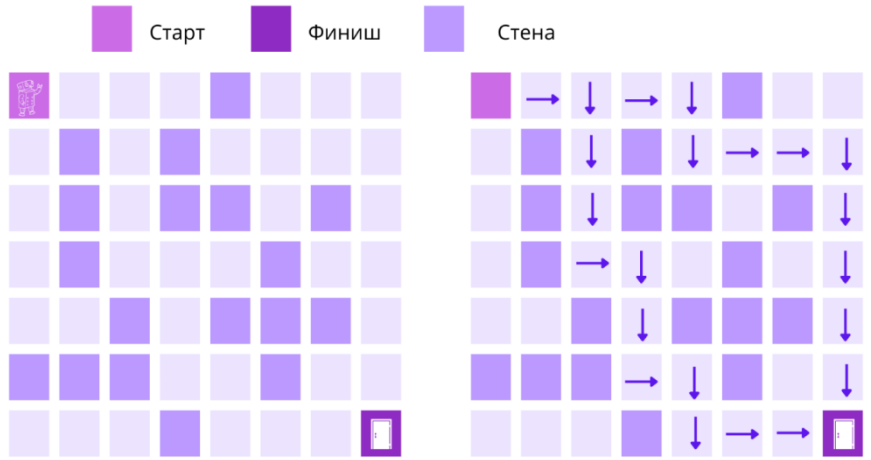
В начале работы агент пробует выполнять разные действия и запоминает, какие из них приносили награды в каких состояниях среды. После нескольких шагов агент попробовал пару-тройку действий, а про пользу или вред других действий агенту ничего не известно, потому что он их еще не пробовал. Далее агенту нужно выбрать следующее действие, и возникает выбор: либо выбирать наиболее оптимальное действие из ранее сделанных (оптимальность других действий оценить нельзя, так как агент их еще не пробовал выполнять), либо случайно выбрать из остальных, неизведанных действий.

Первый подход называется **стратегией использования (exploitation)**, а второй — **стратегией исследования (exploration)**.

Алгоритмы обучения с подкреплением

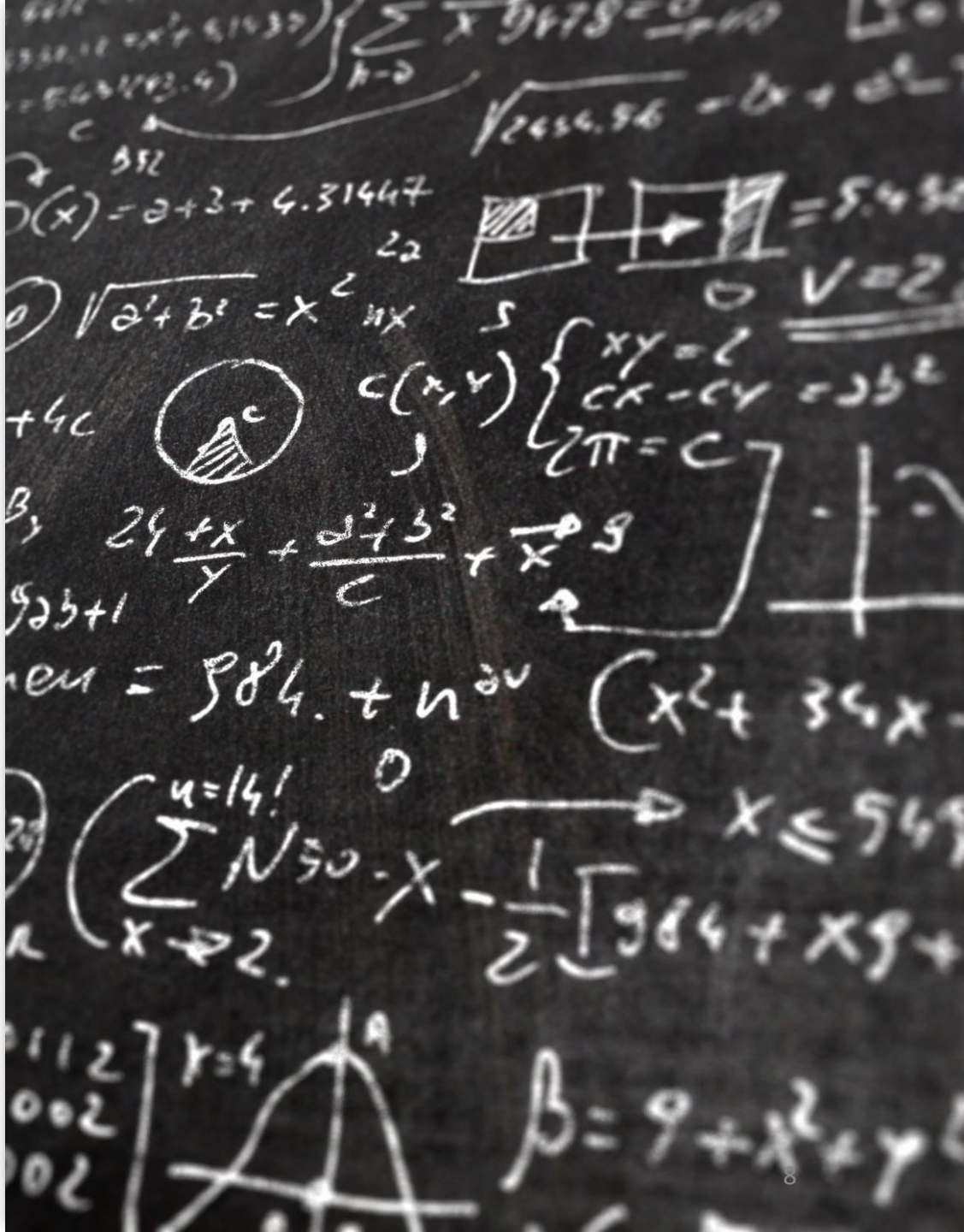
Q-обучение (Q-learning)

Пример. Игра лабиринт



Алгоритмы обучения с подкреплением

- Если состояний среды может быть много, а действий все еще не очень много (например, в простых компьютерных играх, таких как Atari) то можно использовать нейронную сеть, которая по состоянию среды будет предсказывать кумулятивную награду от всех возможных действий — такой алгоритм называется Deep Q-Network (DQN). Среди концептуально похожих алгоритмов можно выделить алгоритм SARSA (State-Action-Reward-State-Action).
- Если количество возможных действий очень велико, используют подходы, в которых нейронная сеть учится предсказывать следующее действие по состоянию среды, например алгоритм REINFORCE. Наиболее успешными часто оказываются комбинированные подходы, в которых одна нейросеть предсказывает кумулятивную награду, другая — следующее действие, и при обучении они «помогают» друг другу. Пример такого комбинированного подхода — метод A3C (Asynchronous Advantage Actor Critic). Например, A3C можно использовать в трейдинге — состояний биржи может быть бесконечно много (наборов котировок).



Итоги

Обучение с подкреплением — это отдельная область искусственного интеллекта, в которой агент учится выполнять выигрышные действия в результате взаимодействия со средой. Обучение с подкреплением — многообещающая технология, позволившая достичь высоких успехов ИИ в играх, однако в более прикладных сферах пока не достигшая уровня качества человека. Обучение с подкреплением является более сложной задачей, чем классическое обучение с учителем, поэтому если в бизнес-задачах можно использовать последнее, то лучше так и сделать.

Темы докладов

1. В 2019 году нейросети выиграли у человека в покер (разработка компании Facebook), в игре Го, в шахматах
2. Активно ведутся исследования по применению обучения с подкреплением в робототехнике
3. Ассоциативные правила и алгоритмы составления ассоциаций ссылке.
4. Рекомендательные системы и их использование на практике, ссылке.
5. Методы кластеризации (например, DBSCAN)
6. Методы отбора признаков, ссылке.

Бесплатный курс от Сбера по генеративному искусству

https://courses.sberuniversity.ru/generative_art?utm_source=tg&utm_medium=organic&utm_campaign=courses&utm_content=gen_i&utm_term=01_09_2023