Урок 28. Обучение с подкреплением

7 класс · Урок 28 из 34 · 45 минут

Обучение с подкреплением: урок для 7 класса об агенте, среде, награде и стратегии, с парным экспериментом по исследованию неизвестных действий.

Знакомый способ уже приносит очки, но неизвестный может оказаться лучше. Когда стоит потратить попытку на новое? Разберём обучение с подкреплением и сравним собственные стратегии в парном опыте.

▼ Содержание урока

1. Цель урока
2. Ход урока
3. Этап 1. Вход в тему: новый способ
4. Этап 2. От действий к стратегии
5. Этап 3. Ветер не выполняет готовую команду
6. Этап 4. Парный эксперимент со стратегиями
7. Этап 5. Итоги урока

▼ 01. Цель урока

Понять, как последовательность действий и получаемая награда помогают улучшать стратегию.

К концу урока ученики смогут:

- Различать агента, состояние, действие и награду.
- Объяснять отличие подкрепления от готовой метки правильного действия.
- Обсуждать выбор между исследованием нового и использованием известного.
- Испытывать стратегию и объяснять влияние правила награды.

▼ 02. Ход урока

Этап	Мин	Учитель	Ученики
1. Вход в тему: новый способ	5	Обсуждает два реальных выбора между пробой и знакомым способом.	Приводят свой опыт и объясняют цену попытки.
2. От действий к стратегии	10	Объясняет цикл действий, наград и изменения стратегии.	Сопоставляют награду, метку и будущую пользу.
3. Ветер не выполняет готовую команду	8	Разбирает задачу управления шаром в ветре.	Объясняют значение меняющейся среды и испытания.
4. Парный эксперимент со стратегиями	16	Организует два парных раунда и разбор стратегий.	Ведут журнал, объясняют цену исследования и смену награды.
5. Итоги урока	6	Обсуждает общие выводы и ответы на итоговые вопросы.	Формулируют знания и объясняют ответы.

03. Этап 1. Вход в тему: новый способ 5 минут

Говорим ученикам

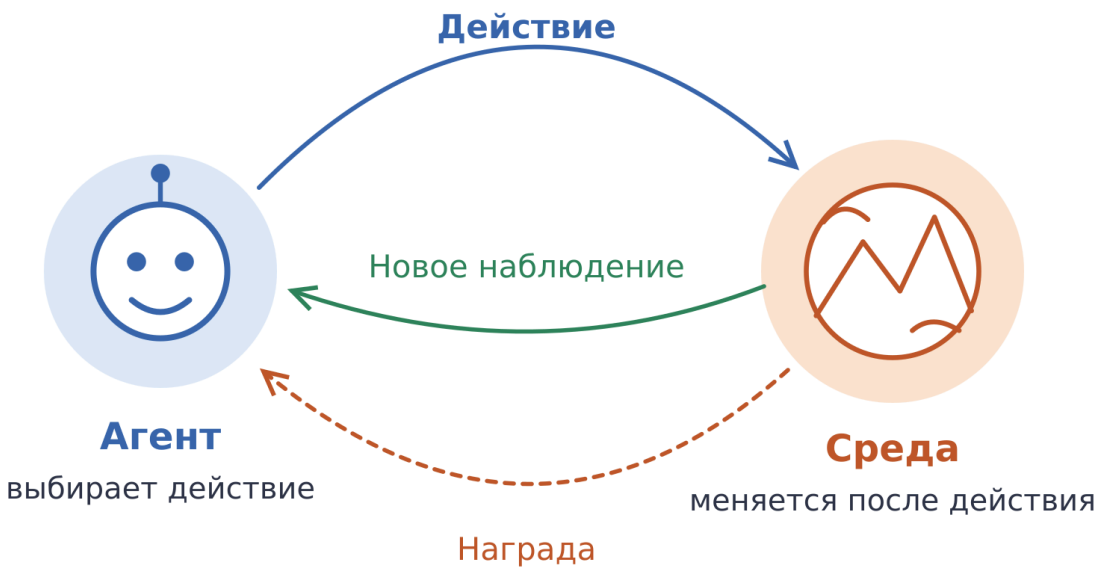
Вспомните, как пробовали незнакомую возможность в игре, конструкторе или творческой программе. Иногда знакомый способ уже даёт результат, но другой может оказаться лучше. Как вы решаете, стоит ли потратить попытку на исследование? Назовите конкретный выбор из своего опыта.

Обучение с подкреплением связано с похожим вопросом: как выбирать действия, когда результат узнаёшь после попытки. В нашем опыте решения принимают люди; программа выбирала бы действия по своему алгоритму. Мы разберём устройство такого обучения, а затем исследуем стратегии в небольшом опыте.

04. Этап 2. От действий к стратегии 10 минут

Говорим ученикам

Агент — система, которая выбирает действия. Среда — то, с чем она взаимодействует. Состояние — текущая ситуация в среде. Агент получает сведения о ней; иногда ему видна только часть состояния. После действия среда меняется, а агент получает новое наблюдение и награду — числовую оценку результата по заданному правилу.



Наблюдение сообщает о состоянии; награда оценивает результат

Агент действует, и состояние среды меняется. Затем он получает наблюдение — сведения о новом состоянии — и награду по заданному разработчиками правилу. Это не готовая метка правильного действия.

Стратегия задаёт способ выбора действий по состоянию. Цель — получить хороший суммарный результат за последовательность шагов. Это отличается от подсказки «на каждом шаге нажми сюда»: правильное действие заранее может быть неизвестно. Награда оценивает результат, но сама по себе не указывает, какой шаг к нему привёл.

В обучении с учителем часто известен требуемый ответ для примера. Здесь агент пробует действия и использует последствия. Ближайшая награда и долгосрочная польза могут различаться: полезная подготовка иногда не даёт немедленного выигрыша. Поэтому важна вся последовательность.

Есть ещё выбор между исследованием и использованием. Пробуя неизвестное действие, агент рискует потратить попытку, зато может открыть лучший способ. Повторяя известное, он получает ожидаемый результат, но может пропустить улучшение. Уместный баланс зависит от числа оставшихся попыток и изменчивости среды.

05. Этап 3. Ветер не выполняет готовую команду

8 минут

Говорим ученикам

Воздушный шар не может просто повернуть мотором туда, куда хочет оператор. На разных высотах ветер дует по-разному. Меняя высоту, система выбирает поток, который понесёт шар в нужную сторону. Но сведения о ветре неполны и условия меняются, поэтому одного заранее записанного маршрута недостаточно.

2 декабря 2020 года исследователи проекта Loon описали применение обучения с подкреплением для управления стратосферными шарами. В публикации сообщалось и об управляемом 39-дневном эксперименте над Тихим океаном. Система училась действовать в

изменчивой среде, а не воспроизводить готовую последовательность команд.

Действие может изменить то, что агент сможет сделать дальше. Поэтому при обучении учитывают и более поздние результаты. Испытание в реальной среде показывает, переносится ли стратегия за пределы условий, на которых её подбирали.

Реальный случай: управление шарами в меняющемся ветре

Исследователи Loon и Google проверяли обученную систему управления в настоящих полётах, включая эксперимент над Тихим океаном. В статье сравнили его с прежним алгоритмом управления.

Этот случай показывает, зачем оценивать стратегию по последовательности результатов в меняющихся условиях. Одного удачного решения в одной точке недостаточно.

Для учителя

Речь об исследовании 2020 года; не представлять Loon как действующий сегодня сервис. Для объяснения достаточно зависимости направления переноса от высоты. Не приписывайте уроку управление реальным аппаратом.

06. Этап 4. Парный эксперимент со стратегиями 16 минут

Говорим ученикам

Мы исследуем ограниченную задачу: выбор среди неизвестных действий, у которых есть постоянная отдача. Это упрощённая задача: здесь нет меняющегося состояния и длинной цепочки последствий. Зато можно увидеть цену исследования. Человек в опыте принимает решения осознанно; компьютерная программа настраивала бы способ выбора по своим правилам.

Задание. Что стоит исследовать?

Один участник готовит три непрозрачных конверта А, Б и В. На скрытых карточках он записывает разные постоянные награды: 1, 3 и 5 очков, распределяя их в неизвестном партнёру порядке. Конверты не содержат предметов: выбор — только запись буквы, а ведущий сообщает соответствующее число. Затем роли меняются с новым распределением.

1. До начала игрок записывает план: когда пробовать новое, когда повторять знакомое. Ему известны возможные награды, но неизвестно их расположение.
2. Сделайте восемь выборов. После каждого запишите действие, награду и накопленную сумму; отмечайте момент изменения стратегии.
3. Сравните первоначальный план с тем, что произошло. Когда полученными сведениями ещё можно было воспользоваться в следующих ходах, а когда будущих ходов уже не осталось?
4. Обсудите другой критерий: если поощрять только число разных открытых конвертов, какую стратегию это сделает выгодной?

5. Поменяйтесь ролями и повторите короткую серию. Сравните рассуждения, а не объявляйте победителя по одной сумме.

Результат — журнал и объяснение одного решения. Критерии: стратегия названа; получение сведений отличено от получения очков; вывод учитывает оставшиеся попытки; ограничение опыта указано.

Для учителя

Карточки наград ведущий распределяет сам и держит скрытыми. При награде 5 игрок знает, что лучше действия в этом опыте нет; при 1 или 3 исследование может оказаться полезным. На последнем ходе неизвестное действие может принести больше или меньше очков. Но воспользоваться полученными сведениями в следующем ходе уже нельзя. Это не означает, что его немедленная ожидаемая выгода всегда отрицательна. Не подменяйте анализ суммой очков: порядок открытия влияет на результат. Два раунда по восемь быстрых записей и обсуждение укладываются в этап.

07. Этап 5. Итоги урока 6 минут

Говорим ученикам

Обучение с подкреплением соединяет действие, последствия и изменение стратегии. Но числовая награда — наше описание цели. Если описать цель неудачно, система может научиться успешно набирать баллы без нужной человеку пользы. Поэтому проверяют и выбранную стратегию, и само правило оценки.

Главное за урок

- При обучении с подкреплением агент использует последствия действий и награды для улучшения стратегии.
- Полезное действие оценивают с учётом будущих результатов, а не только ближайшего выигрыша.
- Исследование нового помогает находить лучшие действия, но требует попыток и может уменьшать текущую отдачу.
- Изменение правила награды меняет выгодное поведение; награда должна отражать нужный результат.

Итоговые вопросы

Ответьте коротко. Затем обсудим объяснения двух-трёх участников.

1. Чем награда после действия отличается от готового правильного ответа при обучении с учителем?

2. Почему действие с маленькой немедленной наградой иногда полезно для будущего результата?
3. Как число оставшихся попыток влияет на ценность исследования?
4. Что нужно проверить, если система хорошо набирает очки, но не решает задачу человека?

Для учителя

1. Награда оценивает последствия, а метка задаёт требуемый ответ на пример; награда не обязана указывать верное действие.
2. Оно может давать сведения или улучшать будущие возможности.
3. Когда будущих попыток мало, остаётся меньше времени воспользоваться найденным знанием.
4. Соответствие награды цели и способы получения очков, а не только величину суммы.

Принимаются ответы по смыслу. Важна связь с общим знанием, а не воспроизведение деталей реального случая.

Для учителя: источники

1. Pablo Samuel Castro. Autonomous navigation of stratospheric balloons using reinforcement learning. Дата материала: 02.12.2020. Материал старше 2024 года. Подтверждающая фраза: «including a 39-day controlled experiment over the Pacific Ocean».