8.1. Формальные обозначения

В то время как контролируемое и неконтролируемое обучение находятся в противоположных концах диапазона всевозможных методов машинного обучения, *обучение с подкреплением* (reinforcement learning) находится где-то посредине. Оно не является обучением с учителем, потому что обучающие данные появляются из алгоритма принятия решения о том, исследовать или использовать. И его нельзя считать обучением без учителя, потому что алгоритм использует обратную связь с окружающей средой. Пока выполнение действия в том или ином положении приводит к успеху, можно пользоваться обучением с подкреплением, чтобы найти благоприятную последовательность действий для максимального обеспечения ожидаемых наград.

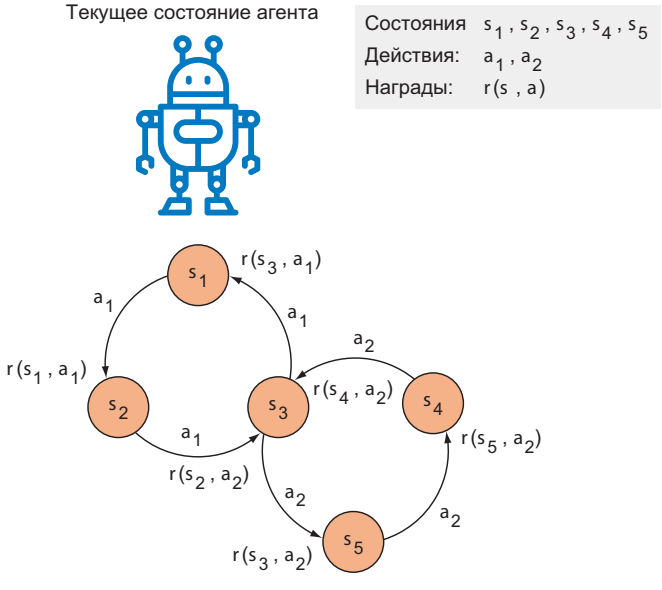

Рис. 8.2. Действия представлены стрелками, а состояния — кружками. Выполнение действия в состоянии приводит к награде. Если начать в состоянии $s1$, можно выполнить действие $a1$, чтобы получить награду $r(s1, a1)$

Нельзя не заметить, что фраза «обучение с подкреплением» содержит антропоморфизацию алгоритма, которая заключается в выполнении *действий*

в определенных *ситуациях* для *достижения успеха*. Этот алгоритм часто называют *агентом*, который *взаимодействует* с окружающей средой. Не следует удивляться тому, что большая часть теории обучения с подкреплением применяется в робототехнике. На рис. 8.2 демонстрируется взаимодействие между состояниями, действиями и результатами (успехом).

Для изменения состояния робот выполняет действия. Но как он решает, какое действие следует предпринять? Чтобы ответить на этот вопрос, в следующем подразделе вводится новая концепция, получившая название *политика* (policy).

ПОЛЬЗУЮТСЯ ЛИ ЛЮДИ ОБУЧЕНИЕМ С ПОДКРЕПЛЕНИЕМ?

Обучение с подкреплением представляется лучшим способом объяснить, как выполнять следующее действие, основываясь на текущей ситуации. Возможно, люди ведут себя так исходя из биологических причин. Но давайте не будем опережать самих себя и рассмотрим следующий пример.

Иногда люди действуют не задумываясь. Если мне хочется пить, я могу инстинктивно схватить стакан воды, чтобы утолить жажду. Я не перебираю в сознании все возможные движения и не выбираю оптимальный вариант, как именно поднести стакан, после тщательных вычислений.

Самое важное то, что выполняемые нами действия не характеризуются одними только наблюдениями в каждый момент их выполнения. Иначе получается, что мы не умнее бактерии, которая действует в зависимости от состояния окружающей среды. Кажется, что происходит что-то более сложное, и простая модель обучения с подкреплением не может полностью объяснить поведение человека.

8.1.1. Политика

Каждый приводит свою комнату в порядок по-разному. Некоторые начинают заправлять постель. Я предпочитаю приводить в порядок комнату по часовой стрелке, чтобы не пропустить ни одного угла. Вы когда-нибудь видели, как работает робот-пылесос, например, Roomba? Некоторые программируют стратегию, которой можно следовать, чтобы выполнить в комнате уборку. В стимулированном обучении процесс выбора агентом определенного действия

называется *политикой*: это набор действий, которые определяют следующее состояние (рис. 8.3).

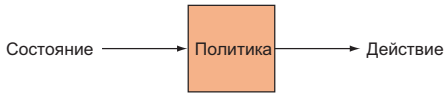

Рис. 8.3. Стратегия предлагает, какое следует выполнить действие в данном состоянии

Цель обучения с подкреплением — выявить наилучшую стратегию. Распространенным способом разработки политик является анализ долговременной последовательности действий в каждом состоянии. *Награда* является мерой результативности выполнения какого-либо действия. Наилучшую из всех возможных стратегий называют *оптимальной политикой*, и она является Священным Граалем обучения с подкреплением. Оптимальная политика указывает оптимальное действие в любом состоянии, однако она может не обеспечивать в тот момент максимальную награду.

Если измерять награду по немедленным последствиям, то есть по состоянию вещей после предпринятого действия, то ее легко просчитать. Такой подход называют *жадной стратегией*, однако не всегда стоит выбирать действие с наилучшей *немедленной* наградой. Например, при уборке комнаты вы можете сначала заправить постель, потому что так комната сразу выглядит аккуратнее. Но если вам еще нужно постирать простыни, то уборку постели нельзя считать лучшей стратегией. Вам следует посмотреть на результаты следующих нескольких действий и на конечное состояние, чтобы прийти к оптимальному варианту. Аналогичным образом при игре в шахматы взятие ферзя у противника может дать преимущество фигурам на доске, но если через несколько ходов это приведет к шаху и мату, то сделанный ход не лучший из всех возможных.

Вы можете также выбирать действие произвольным образом, и эту стратегию называют *случайной стратегией*. Если вы нашли стратегию для решения задачи обучения с подкреплением, рекомендуется устроить двойную проверку того, что полученная в процессе обучения стратегия более эффективна, чем случайная стратегия или жадная стратегия.

ОГРАНИЧЕНИЯ (МАРКОВСКОГО) ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

Большинство вариантов обучения с подкреплением предполагают, что лучшее действие может быть установлено на основе знания текущего состояния, а не исходя из учета длительной истории состояний и действий, которые привели к этому состоянию. Этот метод принятия решений основан на текущем состоянии и называется марковским, а фреймворк часто называют *марковским процессом принятия решений* (МППР).

Ситуации, в которых в определенном состоянии достаточно сведений, чтобы сделать следующий шаг, можно моделировать с помощью алгоритмов обучения с подкреплением. Но большинство ситуаций реального мира нельзя считать марковскими и для этих ситуаций нужен более реалистичный метод, такой, например, как иерархическое представление состояний и действий. Проще говоря, иерархические модели похожи на контекстно-свободную грамматику, тогда как МППР напоминают конечные автоматы. Экспрессивный скачок в моделировании проблемы, как в случае перехода от МППР к иерархическим представлениям, может значительно повысить эффективность алгоритма планирования.

8.1.2. Выгода

Долгосрочную награду называют *выгодой*. Если известна выгода выполнения определенного действия в определенном состоянии, поиск оптимальной стратегии легко выполнить с помощью обучения с подкреплением. Например, чтобы решить, какое выполнить действие, выбирается действие с максимальной выгодой. Самое сложное, как можно догадаться, состоит в раскрытии значений этой выгоды.

Выгода выполнения действия a в состоянии s записывается как функция $Q(s, a)$, которую называют *функцией выгоды* (рис. 8.4).

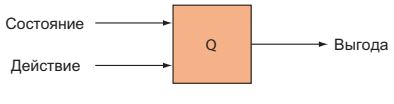

Рис. 8.4. При заданном состоянии и выполненном действии применение функции выгоды Q предсказывает ожидаемую и суммарную выгоды: ближайшая награда (следующее состояние) плюс награды, полученные позже, при следовании оптимальной стратегии

УПРАЖНЕНИЕ 8.1

Если у вас функция выгоды *Q*(s, a), как ее можно использовать, чтобы получить функцию политики?

ОТВЕТ

Policy(s) = argmax_a *Q*(s, a)

Элегантный способ рассчитать выгоду определенной пары «состояние — действие» (*s*, *a*) заключается в рекурсивном учете выгод будущих действий. На выгоду текущего действия влияет не только ближайшая награда, но и наилучшие действия (см. следующую формулу). В этой формуле *s'* обозначает следующее состояние, а *a'* — следующее действие. Награда за выполнение действия *a* в состоянии s обозначается *r*(*s*, *a*):

$$Q(s, a) = r(s, a) + \gamma \max Q(s', a').$$

Здесь γ — гиперпараметр, который следует выбрать и который называют *дисконтирующим множителем* (discount factor). Если γ равен 0, агент выбирает действие, которое приводит к максимальной ближайшей награде. Более высокие значения γ заставят агента придать больше значения учету долгосрочных последствий. Эту формулу можно прочитать следующим образом: «выгодой этого действия является ближайшая награда при условии выполнения этого действия, добавленная к произведению дисконтирующего множителя на максимальную награду, которая возникла после этого».

Оценка будущих наград относится к одному из тех гиперпараметров, которыми можно манипулировать, но есть также и другой гиперпараметр. В некоторых вариантах использования обучения с подкреплением ставшая доступной новая информация может оказаться важнее, чем архивные записи, и наоборот. Например, если робота пытаются обучить выполнять задания быстро, но не обязательно оптимально, то можно задать более высокую скорость обучения. Или, если роботу дали больше времени, чтобы изучить новые действия и варианты использования старых действий, скорость обучения придется уменьшить. Обозначим скорость обучения через α и изменим функцию выгоды (обратите внимание на то, что, когда α = 1, уравнения становятся идентичными):

$$Q(s, a) \leftarrow Q(s, a) + \alpha\,(r(s, a) + \gamma \max Q(s', a') - Q(s, a)).$$

Обучение с подкреплением может быть выполнено, если известна Q-функция, $Q(s, a)$. *Нейронные сети* (глава 7) представляют собой способ аппроксимации функций, если имеется достаточно обучающих данных. TensorFlow служит превосходным инструментом для работы с нейронными сетями, так как содержит множество важных алгоритмов, упрощающих использование нейронных сетей.

8.2. Применение обучения с подкреплением

Для использования обучения с подкреплением необходимо определить способ получения награды после того, как выполнено определенное действие из определенного состояния. Биржевой трейдер без труда выполняет эти требования, потому что покупка и продажа ценных бумаг меняет состояние брокера (остаток денежных средств), при этом каждое действие обеспечивает награду или приводит к потере.

Состояния в этой ситуации представляют собой вектор, содержащий информацию о текущем состоянии бюджета, текущем объеме ценных бумаг и недавнюю историю цен ценных бумаг (последние 200 курсов акций). Каждое состояние является 202-мерным вектором.

УПРАЖНЕНИЕ 8.2

Назовите возможные недостатки использования обучения с подкреплением для задачи по покупкам и продаже акций.

ОТВЕТ

Выполняя действия на рынке, такие как покупка или продажа акций, вы можете оказать влияние на рынок, в результате чего он резко изменится по сравнению с вашими данными обучения.

Проще говоря, на фондовом рынке есть только три действия: покупка, продажа и удержание ценных бумаг.