1 Что такое обучение с подкреплением

Обучение с подкреплением (Reinforcement Learning, RL) — способ машинного обучения (Machine Learning, ML), при котором выполняется автоматическое обучение процессу принятия решений во времени. Эта задача получила широкое распространение во многих научных и инженерных областях.

В нашем изменчивом мире даже задачи, кажущиеся стационарными, со временем приобретают динамический характер. Рассмотрим в качестве примера классическую задачу обучения с учителем, когда нужно классифицировать изображения домашних животных по двум категориям: собаки и кошки. У вас есть тренировочный набор данных и классификатор, реализованный с помощью вашего любимого инструментария для глубокого обучения. Спустя некоторое время модель сошлась и дает превосходные результаты. Хорошо? Разумеется, да! Вы развернули ее на боевых серверах и оставили работать. Затем, после отдыха на морском побережье, вы вдруг обнаружили, что сменилась мода на стрижки собак, и значительная часть ваших запросов теперь классифицируется ошибочно. Поэтому вам нужно обновить тренировочные изображения и повторить весь процесс заново. Хорошо? Разумеется, нет!

Этот пример демонстрирует тот факт, что даже у простых задач машинного обучения есть скрытое измерение времени, которое обычно не учитывается, но может стать причиной проблем в промышленных системах.

Обучение с подкреплением — подход, который изначально включает это дополнительное измерение (чаще всего это время) в процесс обучения, что делает его значительно ближе к человеческому восприятию искусственного интеллекта. В данной главе вы узнаете:

- о связи и различиях между обучением с подкреплением и другими областями машинного обучения (обучением с учителем и без учителя);
- об основных формализмах и моделях обучения с подкреплением и их взаимосвязях;
- теоретические основы обучения с подкреплением — разберем марковские процессы принятия решений.

Обучение с учителем, без учителя и с подкреплением

Возможно, вам знакомо понятие обучения с учителем. Это наиболее изученная и широко известная задача машинного обучения. Основная идея этого метода заключается в том, чтобы автоматически построить функцию, сопоставляющую входным данным выходные данные при заданном наборе примеров. В такой формулировке эта задача кажется простой, но с ней связано множество сложных частных случаев, с которыми компьютеры только недавно стали более или менее справляться.

Существует множество задач обучения с учителем, включая следующие.

- **Классификация текстов.** Является ли полученное по электронной почте сообщение спамом?
- **Классификация изображений и определение местоположения объектов.** На этой картинке изображена кошка, собака или кто-нибудь еще?
- **Предсказания.** Располагая историей наблюдений с разных датчиков, можно ли сделать прогноз погоды на завтра?
- **Анализ эмоций.** Как определить степень удовлетворенности клиента по тексту отзыва?

Вопросы могут показаться разноплановыми, но их объединяет одна идея: есть множество примеров входных данных и желаемых результатов и нужно научиться генерировать выходные данные по неизвестным в настоящий момент входным данным. Из самого термина *«обучение с учителем»* следует, что обучение строится на известных данных, которые были получены от эксперта, предоставившего правильные ответы.

С другой стороны, существует так называемое обучение без учителя, которое предполагает, что поскольку эксперта нет, то нет и известных меток. Следовательно, необходимо изучить скрытую структуру предоставленного набора данных. Одним из широко известных примеров такого подхода к обучению является кластеризация. В этом случае алгоритм пытается объединить данные в набор кластеров в соответствии с некоторыми зависимостями между отдельными примерами.

Еще один метод обучения без учителя, который набирает все большую популярность, — это *генеративно-состязательные сети* (Generative Adversarial Networks, GAN). Основная идея заключается в том, что у нас есть две соревнующиеся сети, первая из которых пытается сгенерировать поддельные данные с целью ввести в заблуждение вторую, в то время как вторая старается отличить искусственно сгенерированные данные от настоящих. С течением времени обе сети становятся все более и более искусными в выполнении своих заданий за счет того, что улавливают неочевидные характерные структуры в наборе данных.

Обучение с подкреплением является третьим подходом и находится где-то между полным контролем и совершенным отсутствием предопределенных меток. С одной стороны, в нем используются многие устоявшиеся методы обучения с учителем, такие как глубокие нейронные сети для аппроксимации функций, сто-

хастический градиентный спуск и метод обратного распространения для обучения представлению данных. С другой стороны, они чаще всего применяются несколько иным образом, чем в обучении с учителем.

В этой главе мы рассмотрим особенности обучения с подкреплением, включая формализмы и абстракции в более или менее четком виде. А пока, чтобы сравнить обучение с подкреплением, обучение с учителем и без учителя, обратимся к более наглядному примеру. Предположим, что у вас есть агент, которому нужно предпринимать действия, находясь в определенной среде. Робомышь в лабиринте на рис. 1.1 послужит хорошим примером, но вы также можете представить вертолет с автопилотом или программу для игры в шахматы. Для простоты остановимся на робомыши.

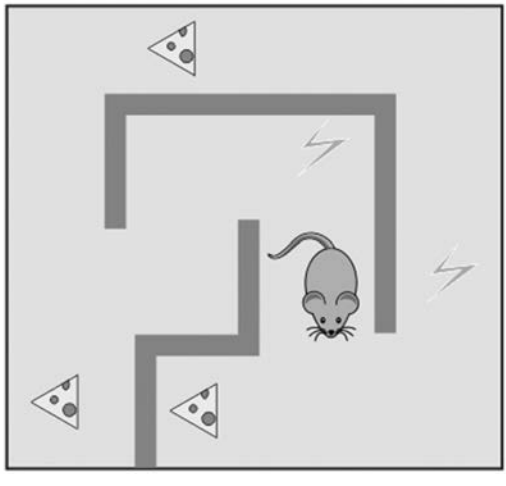

Рис. 1.1. Мир робомыши в лабиринте

Ее средой является лабиринт, в одних точках которого можно найти еду, а в других — получить удар электрическим током. Робомышь может совершать такие действия, как поворот налево или направо и движение вперед. Она может наблюдать полное состояние лабиринта для принятия решения о дальнейших действиях. Цель робота состоит в том, чтобы найти как можно больше еды, по возможности избегая ударов электрическим током. Эти сигналы о еде и электрическом токе являются вознаграждением, полученным агентом от среды, для дополнительной оценки его действий. Вознаграждение является весьма важной концепцией в обучении с подкреплением, и мы будем обсуждать его далее. На текущий момент достаточно понимать, что конечная цель агента заключается в том, чтобы получить как можно большее суммарное вознаграждение. В данном случае мышь может немного пострадать от удара электрическим током, чтобы добраться до еды, что будет лучше, чем если она будет просто стоять и ничего не получит.

В то же время нам нужно избегать жесткого прописывания в памяти робота информации о среде и лучших действиях в конкретной ситуации, так как это слишком трудоемко и может стать бесполезным даже при незначительном изменении

лабиринта. Чего бы нам хотелось, так это иметь некий магический набор методов, который позволит нашему роботу самостоятельно обучаться тому, как избегать ударов электрошоком и собирать как можно больше еды.

Обучение с подкреплением как раз и есть тот самый магический набор инструментов, который действует иным образом, чем методы обучения с учителем и без учителя. Он не работает с заранее определенными метками, как это делает обучение с учителем. Никто не помечает картинки, которые видит робот, как *плохие* или *хорошие*, и никто не задает для него наилучшее направление.

Тем не менее мы не действуем полностью вслепую, как на стадии оптимизации при обучении без учителя, — у нас есть система вознаграждений. Вознаграждения могут быть положительными при нахождении еды, отрицательными при получении ударов электрическим током и нейтральными, если ничего не происходит. Наблюдая подобные вознаграждения и связывая их с предпринятыми действиями, наш агент учится выполнять действия лучше, собирать большее количество еды и реже получать удары электрическим током.

Конечно же, за такую универсальность и гибкость обучения с подкреплением приходится платить. RL считается куда более сложным, чем обучение с учителем или без учителя. Вкратце рассмотрим, в чем заключается его сложность.

Первое, что следует отметить, — наблюдение в RL зависит от поведения агента и в некоторой мере является его *результатом*. Если ваш агент решает действовать неэффективно, то из наблюдений вы ничего не поймете о том, что было сделано неправильно и как следует поступить, чтобы улучшить результат (агент просто будет постоянно получать отрицательные вознаграждения). Если агент с завидным упрямством продолжает идти по неверному пути, то наблюдения могут дать ложное представление о том, что способа получить большее вознаграждение не существует (жизнь — это страдания), а это может оказаться абсолютно неверным. В терминах машинного обучения это может быть перефразировано как наличие не-i.i.d-данных. Аббревиатура *i.i.d* расшифровывается как *independent and identically distributed* («независимые и одинаково распределенные») — очень важное требование для большинства методов обучения с учителем.

Другие трудности в жизни нашего агента связаны с тем, что ему нужно не только использовать стратегию (политику), которой он обучился, но и активно исследовать среду, ведь, кто знает, может быть, если мы будем действовать по-другому, то сможем значительно улучшить полученный результат. Проблема заключается в том, что если исследований среды будет слишком много, то наше вознаграждение может значительно уменьшиться (не говоря уж о том, что агент может вообще забыть, чему он научился ранее). То есть нам нужно найти некоторый баланс между двумя этими видами деятельности. Проблема выбора между использованием стратегии и исследованием среды — одна из фундаментальных проблем обучения с подкреплением.

Люди постоянно сталкиваются с подобным выбором: пойти поужинать в уже известное место или заглянуть в новый модный ресторан? Как часто нужно менять работу? Что лучше: заняться изучением новой области или продолжить работу в прежней? На эти вопросы нет универсальных ответов.

Третьим усложняющим фактором является то, что вознаграждение и действия могут значительно отстоять друг от друга. В шахматах это может быть сильный ход

в середине игры, решивший ход партии. Во время обучения нам нужно выявлять подобные ситуации и делать выводы, что может оказаться трудновыполнимым.

Тем не менее, несмотря на все эти препятствия и сложности, интерес к обучению с подкреплением растет в области как теории, так и практического применения.

Итак, если вам это интересно, двинемся дальше и рассмотрим основные абстракции обучения с подкреплением, при этом немного углубимся в детали.

Зависимости и отношения в обучении с подкреплением

В каждой научной и инженерной сфере делаются свои предположения и вводятся ограничения. В предыдущем разделе мы рассмотрели обучение с учителем, в котором подобным предположением является знание пар входных и выходных данных (меток). У ваших данных нет меток? Простите, но вам необходимо придумать, как их получить, или использовать какой-нибудь другой метод обучения. Это не говорит о том, что обучение с учителем плохое или хорошее, это просто делает его неприменимым к вашей задаче. Важно знать и понимать правила игры для каждого метода, тогда можно сэкономить много времени. Да, известно множество теоретических и практических прорывов, совершенных, когда кто-то пытался бросить вызов правилам, подойдя к делу творчески. Но все-таки сначала вам нужно разобраться в этих ограничениях.

Конечно же, подобные формализмы существуют и для обучения с подкреплением, и сейчас лучшее время познакомиться с ними, раз уж оставшаяся часть книги посвящена их анализу с различных точек зрения.

На рис. 1.2 вы можете видеть две основные составляющие обучения с подкреплением — *агента* и *среду*, а также способы их взаимодействия — *действия*, *вознаграждения* и *наблюдения*.

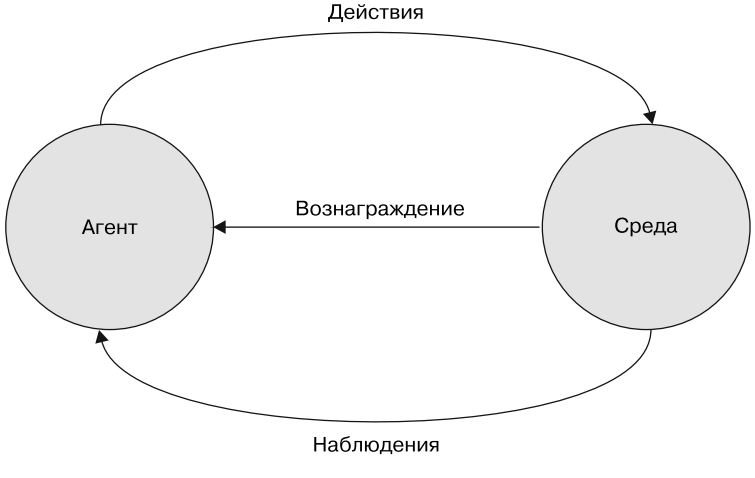

Рис. 1.2. Составляющие обучения с подкреплением и их взаимодействие

Вознаграждение

В первую очередь нужно рассмотреть понятие вознаграждения. В обучении с подкреплением это просто скалярное значение, периодически получаемое нами от среды. Оно может быть положительным или отрицательным, большим или маленьким — это просто число. Цель вознаграждения состоит в том, чтобы сообщить нашему агенту, насколько хорошим было его поведение. Частота, с которой агент получает вознаграждение, никак не задана — это может происходить ежесекундно либо один раз за всю жизнь, тем не менее распространенной практикой является получение вознаграждения через равные промежутки времени или при каждом взаимодействии со средой, просто для удобства. В случае, когда вознаграждение выдается однократно, все награды, за исключением последней, будут нулевыми.

Как уже упоминалось, цель вознаграждения в том, чтобы обеспечить агента обратной связью, информирующей о его успехах, и это важнейший принцип обучения с подкреплением. Сам термин «подкрепление» основан на том, что полученное агентом вознаграждение должно подкреплять его поведение положительным или отрицательным образом. Вознаграждение локально, это означает, что на его получение влияет только недавняя активность агента, а не успехи, достигнутые им за все время. Разумеется, получение значительного вознаграждения вовсе не означает, что секундой позже вы не встретитесь с катастрофическими последствиями ваших предыдущих решений. Это как ограбление банка, которое может показаться весьма неплохой идеей, пока не задумаешься о том, что за этим последует.

Основная цель агента — получить как можно большее вознаграждение за свои действия. Вот несколько конкретных примеров, поясняющих суть вознаграждения.

- **Торговля на финансовых рынках.** Итоговая прибыль является для участника торгов вознаграждением за покупку и продажу акций.
- **Шахматы.** В данном случае вознаграждение в конце игры принимает форму победы, проигрыша или ничьей и в значительной степени зависит от конкретной ситуации. Для меня, к примеру, сыграть вничью с гроссмейстером было бы серьезным достижением. На практике следует четко указывать фактическую ценность вознаграждения, притом что оценить ее может быть довольно сложно. Так, в шахматах вознаграждение может быть пропорциональным мастерству противника.
- **Дофаминовая система в головном мозге.** В мозге есть участок (лимбическая система), который вырабатывает дофамин, если нужно отправить положительный сигнал головному мозгу. Высокие концентрации дофамина вызывают удовольствие, что подкрепляет действия, которые данная система считает хорошими. К несчастью, лимбическая система очень древняя в том смысле, что она расценивает как хорошее еду, размножение и доминирование, поэтому то, что хорошо для выживания с точки зрения лимбической системы, может быть совершенно неприемлемо в социальном плане.