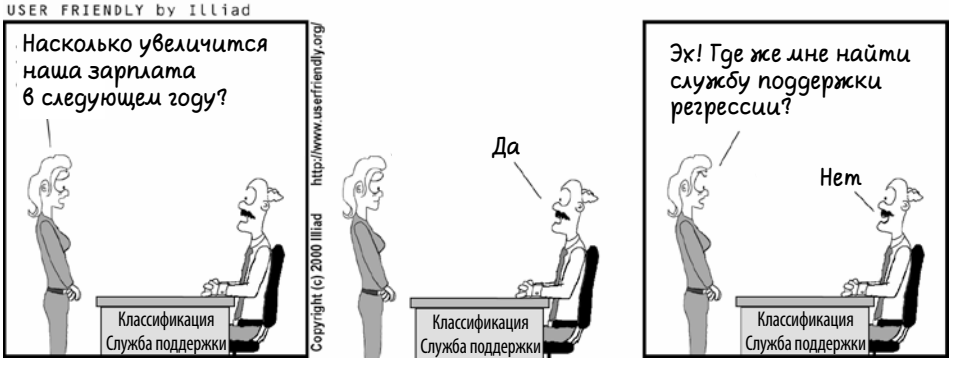
2 Типы машинного обучения

В этой главе

- ✓ Три различных типа машинного обучения: контролируемое, неконтролируемое и обучение с подкреплением.
- ✓ Разница между размеченными и неразмеченными данными.
- ✓ Разница между регрессией и классификацией, а также их использованием.



Как мы узнали из главы 1, машинное обучение — это здравый смысл для компьютера. Принимая решения на основе предыдущих данных, он примерно имитирует процесс, посредством которого люди принимают решения на основе опыта. Естественно, программирование компьютеров для имитации процесса человеческого мышления — сложная задача, потому что компьютеры созданы для хранения и обработки чисел, а не для принятия решений. Ее и стремится решить машинное обучение. Оно делится на несколько ветвей в зависимости от типа принимаемого решения. В этой главе мы рассмотрим некоторые наиболее важные из них.

Машинное обучение нашло применение во многих областях, например таких, как:

- прогнозирование цен на жилье на основе размера дома, количества комнат и местоположения;
- прогнозирование сегодняшних цен на фондовом рынке на основе вчерашних цен и других рыночных факторов;
- обнаружение спама на основе слов, употребленных в электронном письме, и данных отправителя;
- распознавание изображений, например лиц или животных, на основе составляющих их пикселов;
- обработка длинных текстовых документов и создание их резюме;
- рекомендации пользователю видеороликов или фильмов, например, на YouTube или Netflix;
- создание чат-ботов, которые взаимодействуют с людьми и отвечают на вопросы;
- обучение автомобилей с автопилотом самостоятельно двигаться по городу;
- постановка диагноза людям и разделение их на больных и здоровых;
- сегментация рынка на аналогичные группы в зависимости от местоположения, покупательской способности и интересов;
- игры, подобные шахматам или го.

Попробуйте представить, как можно использовать машинное обучение в каждой из этих областей. Обратите внимание на то, что некоторые из этих применений отличаются друг от друга, но задачи могут решать аналогичные. Например, прогнозировать цены на жилье и на акции можно одинаковыми методами. Прогноз того, является ли электронное письмо спамом или является ли транзакция по кредитной карте мошеннической, также можно сделать с помощью сходных методов. А как насчет группировки пользователей приложения на основе их сходства? Это звучит иначе, чем прогнозирование цен на жилье, но выполнить это можно аналогично группировке газетных статей по темам. А как насчет игры в шахматы? Не похоже на все предыдущие приложения, но зато может иметь сходство с игрой в го.

Модели машинного обучения подразделяются на типы в зависимости от того, как они работают.

Итак, три основных семейства моделей машинного обучения:

- контролируемое обучение;
- неконтролируемое обучение;
- обучение с подкреплением.

В этой главе рассмотрим все три. Однако в рамках книги сосредоточимся только на контролируемом обучении, так как это самый естественный способ начать обучение и, возможно, наиболее широко используемый в настоящее время. О других типах можно самостоятельно узнать из литературы, потому что все они интересны и полезны! Среди ресурсов, собранных в приложении B, имеется несколько интересных ссылок, в том числе на несколько видеороликов, созданных автором.

В чем разница между размеченными и неразмеченными данными

Что такое данные

Мы говорили о данных в главе 1, но прежде чем двигаться дальше, дадим четкое определение того, что в этой книге подразумевается под *данными*. Данные — это всего лишь информация. Всегда, когда у нас есть таблица с информацией, у нас есть данные. Обычно каждая строка в таблице представляет собой точку данных. Допустим, у нас есть набор данных о домашних животных. В этом случае каждая строка представляет отдельного четвероногого. Каждое домашнее животное в таблице описывается определенными признаками.

Что же это за признаки

В главе 1 мы определили признаки как свойства или характеристики данных. Если наши данные содержатся в таблице, то признаками являются столбцы таблицы. В примере с домашним животным признаками могут быть размер, имя, тип или вес. Признаками могут служить даже цвета пикселов на изображении домашнего животного. Это то, что описывает наши данные. Однако некоторые признаки — особенные, мы называем их *метками*.

Метки?

Этот вариант чуть менее точен, так как зависит от контекста проблемы, которую мы пытаемся решить. Как правило, если мы хотим спрогнозировать определенный признак на основе других, этот признак служит меткой. Если пытаемся спрогнозировать тип домашнего животного (например, кошки или собаки) на основе информации о нем, то метка — это тип домашнего животного (кошка/собака). Если стремимся спрогнозировать, болен питомец или здоров, основываясь на симптомах и другой информации, то метка — это состояние здоровья (болен/здоров). Если пытаемся предсказать возраст питомца, то метка — это возраст (число).

Прогнозы

Мы уже использовали концепцию свободного прогнозирования, теперь давайте определим ее. Цель прогностической модели машинного обучения состоит в том, чтобы угадать метки в данных. Предположение, которое делает модель, называется *прогнозом*.

Теперь, зная, что такое метки, мы можем допустить, что существуют два основных типа данных: *размеченные* и *неразмеченные*.

Размеченные и неразмеченные данные

Размеченные данные — такие, которые поставляются с тегом или меткой (метка может быть как типом, так и числом). Неразмеченные данные — те, что поставляются без меток. Примером размеченных данных может служить набор данных, относящихся к электронным письмам, в котором есть столбец, где записано, является ли письмо спамом, или столбец, где указано, связано ли оно с работой. Пример неразмеченных данных — набор электронных писем, где нет конкретного столбца, который мы хотели бы спрогнозировать.

На рис. 2.1 мы видим три набора данных, содержащих изображения домашних животных. В первом наборе есть столбец с указанием типа домашнего животного, во втором — столбец, где приведен его вес. Это два примера размеченных данных. Третий набор состоит из изображений без меток, что делает его неразмеченными данными.

Рис. 2.1. Наборы данных: *слева* — размечен, меткой служит тип домашнего животного (собака/кошка); *посередине* — размечен, меткой служит вес (в килограммах); *справа* — не размечен

Конечно, это определение содержит некоторую двусмысленность, потому что в зависимости от проблемы мы определяем, подходит ли конкретный признак в качестве метки. Таким образом, определение того, размечены данные или нет, во многих случаях зависит от решаемой проблемы.

Размеченные и неразмеченные данные дают две ветви машинного обучения — *контролируемое* и *неконтролируемое*. Их определение мы найдем в следующих трех разделах.

Контролируемое обучение — раздел машинного обучения, который работает с размеченными данными

Мы можем обнаружить контролируемое обучение в некоторых наиболее распространенных современных приложениях, включая распознавание изображений, различные формы обработки текста и системы рекомендаций. Контролируемое обучение — это тип машинного обучения, при котором используются размеченные данные. Короче говоря, цель модели контролируемого обучения состоит в том, чтобы спрогнозировать (угадать) метки.

В примере на рис. 2.1 набор данных слева содержит изображения собак и кошек, а метки — «собака» и «кошка». Для этого набора данных модель машинного обучения будет использовать предыдущие данные и прогнозировать метки новых точек данных. Это означает, что, если мы введем новое изображение *без* метки, модель будет угадывать, кто это — собака или кошка, прогнозируя таким образом метку точки данных (рис. 2.2). Здесь точка данных соответствует собаке, и алгоритм контролируемого обучения обучается прогнозировать, что она действительно соответствует собаке.