**Учебный материал 1**

Машинное обучение

Машинное обучение является быстро развивающейся областью изучения, основной задачей которой является разработка и анализ алгоритмов, которые позволяют компьютерам учиться. Пока еще молодая дисциплина, с гораздо большим ожиданием открытия, чем в настоящее время известно, сегодня машинное обучение можно использовать для обучения компьютеров выполнению широкого спектра полезных задач. Это включает в себя такие задачи, как автоматическое обнаружение объектов на изображениях (важнейший компонент автомобилей с автоматическим управлением и водителем), распознавание речи (которая обеспечивает технологию голосовых команд), обнаружение знаний в области медицинских наук (используется для улучшения нашего понимания комплексных заболеваний) и прогнозной аналитики (с использованием для продаж и экономического прогнозирования).
Чтобы получить общее представление о том, как работает машинное обучение, мы начнем с обсуждения простой проблемы машинного обучения игрушкам: обучения компьютера тому, как отличить изображения кошек от изображений с собаками. Это позволит нам неформально описать процедуры, используемые для решения проблем машинного обучения в целом.
Чтобы научить ребенка отличать «кошку» от «собаки», родители (почти!) Никогда не дают своим детям какое-либо формальное научное определение, чтобы различать их; то есть, что собака является представителем вида Canis Familiaris из более широкого класса млекопитающих, и что кошка, принадлежащая к тому же классу, принадлежит к другому виду, известному как Felis Catus. Нет, вместо этого ребенку, естественно, предоставляют много изображений того, что ему говорят, «собаки» или «кошки», пока они полностью не поймут эти два понятия. Как мы узнаем, когда ребенок может успешно различать кошек и собак? Интуитивно, когда они сталкиваются с новыми кошками и собаками, и могут правильно определить каждый новый пример. Как и люди, компьютеры можно научить, как выполнять подобные задачи подобным образом. Задача такого рода, когда мы стремимся научить компьютер различать разные типы вещей, называется проблемой классификации в машинном обучении.

**1.** **Сбор данных.** Как и люди, компьютер должен быть обучен распознавать разницу между этими двумя типами животных, изучая серию примеров, обычно называемых обучающим набором данных. На рисунке 1.1 показан такой тренировочный набор, состоящий из нескольких изображений разных кошек и собак. Интуитивно понятно, что чем больше и разнообразнее учебный набор, тем лучше компьютер (или человек) может выполнять учебное задание, поскольку более широкий спектр примеров дает учащемуся больше опыта. 

**2. Создание особенностей**. Подумайте на минуту о том, как вы сами отличаете изображения, содержащие кошек, от изображений, содержащих собак. Что вы ищете, чтобы отличить эти две части? Скорее всего, вы используете цвет, размер, форму ушей или носа и / или некоторую комбинацию этих функций, чтобы различать их. Другими словами, вы не просто рассматриваете изображение как просто набор из множества маленьких квадратных пикселей. Вы выбираете детали или функции из изображений, подобных этим, чтобы определить, что именно вы смотрите. Это верно и для компьютеров. Для успешного обучения компьютера выполнению этой задачи (и любой задачи машинного обучения в целом) нам необходимо
чтобы предоставить ему должным образом разработанные функции или, в идеале, чтобы он сам нашел такие функции. Обычно это не тривиальная задача, так как разработка качественных функций может зависеть от приложения. Например, такая функция, как «количество ног», была бы бесполезной при различении между кошками и собаками (поскольку у них обоих по четыре!), но весьма полезна при разграничении кошек и змей. Кроме того, извлечение функций из учебного набора данных также может быть сложной задачей. Например, если некоторые из наших тренировочных изображений были размытыми или взяты с точки зрения, где мы не могли видеть голову животного, разработанные нами функции могут быть неправильно извлечены. Однако, ради простоты нашей проблемы с игрушками, предположим, что мы можем легко извлечь следующие две особенности из каждого изображения в обучающем наборе:
1. размер носа относительно размера головы (от маленького до большого);
2. форма ушей (от круглых до острых).

Изучая тренировочные изображения, показанные на рис. 1.1, мы можем видеть, что у всех кошек маленькие носы и заостренные уши, а у собак - большие и круглые. Обратите внимание, что при текущем выборе функций каждое изображение теперь может быть представлено всего двумя числами: число, обозначающее относительный размер носа, и другое число, фиксирующее точечность или округлость ушей. Поэтому теперь мы представляем каждое изображение в нашем тренировочном наборе в двумерном пространстве признаков, где элементы «размер носа» и «форма уха» являются горизонтальной и вертикальной осями координат соответственно, как показано на рис. 1.2. Поскольку наши разработанные элементы отличают кошек от собак в нашем тренировочном наборе, они так хорошо представляют друг друга в одной части пространства, где образы кошек объединяются в одну часть пространства, а изображения собак - вместе в другой части пространства.



**3. Тренировка модели**. Теперь, когда у нас есть хорошее представление характеристик наших тренировочных данных, заключительный акт обучения компьютера тому, как различать кошек и собак, представляет собой простую геометрическую проблему: пусть компьютер найдет линию или линейную модель, которая четко отделяет кошек от собак в нашем тщательно проработанном пространстве признаков1. Поскольку линия (в двумерном пространстве) имеет два параметра: наклон и точку пересечения, это означает нахождение правильных значений для обоих. Поскольку параметры этой линии должны быть определены на основе (представления признаков) обучающих данных, процесс определения надлежащих параметров, который основан на наборе инструментов, известных как числовая оптимизация, называется обучением модели. На рисунке 1.3 показана обученная линейная модель (черного цвета), которая делит пространство признаков на области кошек и собак. Как только эта линия будет определена, любое будущее изображение, изображение элемента которого находится над ним (в синей области), будет считаться компьютером кошкой, а также любое изображение, которое опускается ниже линии (в красной области), будет считается собакой.



**4. Тестирование модели**. Чтобы проверить эффективность работы нашего ученика, мы теперь показываем компьютеру пакет ранее невиданных изображений кошек и собак (обычно называемый набором данных для тестирования) и видим, насколько хорошо он может идентифицировать животное в каждое изображение. На рис. 1.4 показан пример тестового набора для рассматриваемой проблемы, состоящий из трех новых изображений кошек и собак. Чтобы сделать это, мы берем каждое новое изображение, извлекаем наши разработанные элементы (размер носа и форму ушей) и просто проверяем, на какую сторону нашей линии попадает представление объекта. В этом случае, как видно на рис. 1.5, все новые кошки и все собаки, кроме одной, из набора для тестирования были идентифицированы правильно.



Неправильная идентификация одной собаки (бостонского терьера) полностью обусловлена нашим выбором характеристик, которые мы разработали на основе тренировочного набора на рис. 1.1. Эта собака была опознана просто потому, что ее особенности, маленький нос и острые уши, совпадают с характеристиками кошек из нашего тренировочного набора. Итак, хотя сначала казалось, что сочетание размера носа и формы ушей действительно может отличить кошек от собак, мы теперь видим, что наш тренировочный набор был слишком маленьким и недостаточно разнообразным, чтобы этот набор функций был полностью эффективным. Чтобы улучшить нашего ученика, мы должны начать снова. Во-первых, мы должны собрать больше данных, формируя более широкий и разнообразный учебный набор. Затем нам нужно будет рассмотреть вопрос о разработке более отличительных признаков (возможно, цвета глаз, формы хвоста и т. Д.), Которые еще больше помогут отличить кошек от собак. Наконец, мы должны обучить новую модель, используя разработанные функции, и протестировать ее таким же образом, чтобы увидеть, является ли наша новая обученная модель улучшением по сравнению со старой.

**Стадии типичной проблемы машинного обучения**
Давайте теперь кратко рассмотрим ранее описанный процесс, с помощью которого была создана обученная модель для игрушечной задачи по дифференциации кошек от собак. Этот же процесс используется, по существу, для выполнения всех задач машинного обучения, и поэтому стоит остановиться на мгновение и рассмотреть шаги, предпринятые при решении типичных проблем машинного обучения.
Мы перечислим эти шаги ниже, чтобы подчеркнуть их важность, которую мы все вместе называем общим конвейером для решения проблем машинного обучения, и предоставим картину, которая кратко суммирует весь конвейер на рис. 1.6.

