

Знакомство с машинным обучением

1.1.1. Пример: понравится ли фильм пользователю

Для начала будет рассмотрен не очень сложный пример, на котором можно понять, в чем заключается суть обучения с учителем и машинного обучения. Пусть есть некоторый сайт, посвященный кино, на который можно зайти, найти страницу нужного фильма, прочитать информацию про него: когда он снят, кто в нем играет и какой бюджет у этого фильма, а также, возможно, купить его и посмотреть. Пусть есть некоторые пользователи, которые находят страницу нужного фильма, читают и задаются вопросом «смотреть или нет?».

Необходимо понять, понравится ли пользователю фильм, если выдать ему рекомендацию о фильме. Есть несколько подходов к решению:

- Подход первый, самый глупый — дать пользователю посмотреть этот фильм.
- Второй подход — дать случайный ответ и показать случайную рекомендацию. В обоих случаях пользователь может быть разочарован фильмом и он будет недоволен сайтом.
- Третий подход — пригласить психолога-киномана, чтобы разрешить ситуацию. Этот человек оценит пользователя, оценит фильм и поймет, понравится ли этот фильм этому пользователю, сопоставив информацию. Этот подход довольно сложный. Скорее всего таких специалистов не очень много, и будет сложно отмасштабировать это решение на миллионы пользователей сайта. Но на самом деле это не нужно.

Существует множество примеров — ситуаций, когда другие пользователи заходили на страницы фильмов, принимали решение посмотреть фильм и далее ставили оценку, по которой можно понять, понравился им фильм или нет. Задача машинного обучения состоит в восстановлении общей закономерности из информации в этих примерах.

1.1. Основные обозначения

Будем использовать следующие обозначения:

$\mathcal{X}$ — пространство объектов, x — объект, $\mathcal{Y}$ — пространство ответов, $y = y(x)$ — ответ на объекте x .

Объектом называется то, для чего нужно сделать предсказание. В данном примере объектом является пара (пользователь, фильм).

Пространство объектов — это множество всех возможных объектов, для которых может потребоваться делать предсказание. В данном примере это множество всех возможных пар (пользователь, фильм).

Ответом будет называться то, что нужно предсказать. В данном случае ответ — понравится пользователю фильм или нет.

Пространство ответов, то есть множество всех возможных ответов, состоит из двух возможных элементов: -1 (пользователю фильм не понравился) и +1 (понравился).

Признаковым описанием объекта называется совокупность всех признаков:

$$x = (x^1, x^2, \dots, x^d).$$

Признак - это число, характеризующее объект. Признаковое описание является d -мерным вектором.

Центральным понятием машинного обучения является **обучающая выборка**

$$X = (x_i, y_i)_{i=1}^l.$$

Это те самые примеры, на основе которых будет строиться общая закономерность. Отдельная задача — получение обучающей выборки (совокупность прецедентов). В вышеупомянутом случае y_i — это оценка фильма пользователем.

Предсказание будет делаться на основе некоторой модели (алгоритма) $a(x)$, которая представляет из себя функцию из пространства $\mathbb{X}$ в пространство $\mathbb{Y}$. Эта функция должна быть легко реализуема на компьютере, чтобы ее можно было использовать в системах машинного обучения. Примером такой модели является линейный алгоритм:

$$a(x) = \text{sign}(w_0 + w_1 x^1 + \dots + w_d x^d).$$

Операция взятия знака sign берется ввиду того, что пространство $\mathbb{Y}$ состоит из двух элементов.

Не все алгоритмы подходят для решения задачи. Например, константный алгоритм $a(x) = 1$ не подходит. Это довольно бесполезный алгоритм, который вряд ли принесет пользу сайту.

Для оценки качества работы алгоритма можно ввести некоторую характеристику — **функционал ошибки**. $Q(a, X)$ — ошибка алгоритма a на выборке X .

Например, функционал ошибки может быть *долей неправильных ответов*. Следует особо отметить, что Q называется функционалом ошибки, а не функцией. Это связано с тем, что первым его аргументом является функция.

Задача обучения состоит в подборе такого алгоритма a , для которого достигается минимум функционала ошибки Q .

Лучший в этом смысле алгоритм выбирается из некоторого семейства A алгоритмов.

Простейшим примером семейства алгоритмов являются решающие пни:

$$A = \{[x^j < t] \mid \forall j, t\}.$$

Здесь квадратные скобки соответствуют так называемой нотации Айверсона. Если логическое выражение внутри этих скобок — истина, то значение скобок равно 1, в ином случае — нулю.

Алгоритм работает следующим образом. Если значение определенного признака x_j меньше некоторого порогового значения t , то данный алгоритм возвращает ответ 0 (фильм не понравится), в ином случае — +1 (пользователю фильм понравится).

1.2 Обучение с учителем

Далее рассмотрим типы задач при обучении с учителем (или обучении на размеченных данных, как его ещё называют) — это когда есть и объекты, и истинные ответы на них.

Общая постановка задачи обучения с учителем (обучении на размеченных данных) следующая.

Для обучающей выборки $X = (x_i, y_i)_{i=1}^l$ нужно найти такой алгоритм $a \in A$, на котором будет достигаться минимум функционала ошибки:

$$Q(a, X) \rightarrow \min_{\{a \in A\}}$$

В зависимости от множества возможных ответов $\mathbb{Y}$, задачи делятся на несколько типов.

1.2.2. Задача бинарной классификации

В задаче бинарной классификации пространство ответов состоит из двух ответов $Y = \{0, 1\}$. Множество объектов, которые имеют один ответ, называется классом. Говорят, что нужно относить объекты к одному из двух классов, другими словами, классифицировать эти объекты.

1.2.3. Задача многоклассовой классификации

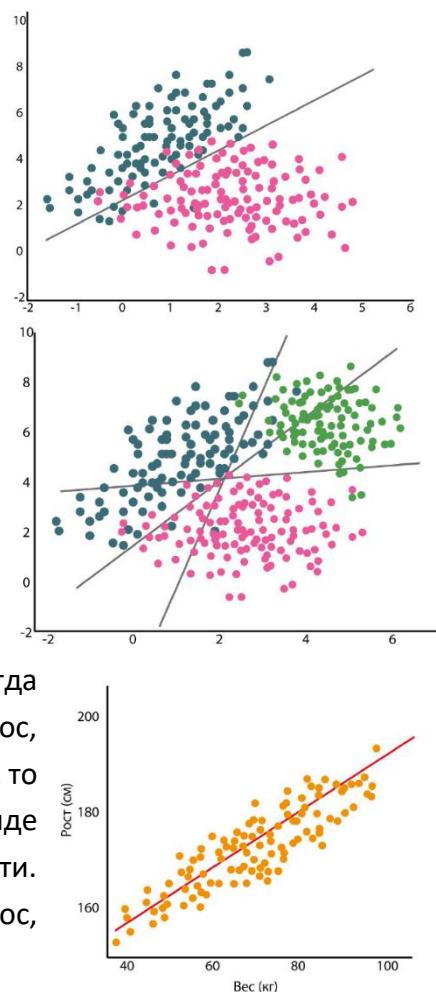
Классов может быть больше, чем два.

1.2.4. Задача регрессии

Когда y является вещественной переменной, говорят о задаче регрессии.

1.2.5. Задача ранжирования

Еще одним примером задачи обучения с учителем является задача ранжирования. Мы сталкиваемся с ней каждый день, когда ищем что-либо в интернете. После того, как мы ввели запрос, происходит ранжирование страниц по релевантности их запросу, то есть для каждой страницы оценивается ее релевантность в виде числа, а затем страницы сортируются по убыванию релевантности. Задача состоит в предсказании релевантности для пары (запрос, страница).



1.3. Обучение без учителя

Задача обучения без учителя — это такая задача, в которой есть только объекты, а ответов нет. Также бывают «промежуточные» постановки. В случае частичного обучения есть объекты, некоторые из которых с ответами. В случае активного обучения получение ответа обычно очень дорого, поэтому алгоритм должен сначала решить, для каких объектов нужно узнать ответ, чтобы лучше всего обучиться.

Рассмотрим несколько примеров постановки задач без учителя.

1.3.1. Задача кластеризации

Первый пример — задача кластеризации. Дано множество объектов. Необходимо найти группы похожих объектов. Есть две основные проблемы: не известно количество кластеров и не известны истинные кластеры, которые нужно выделять. Поэтому задача решается очень тяжело — здесь невозможно оценить качество решения. *Этим и отличается задача классификации* — там тоже нужно делить объекты на группы, но в классификации группы, а точнее классы, фиксированы, и известны примеры объектов из разных групп.



1.3.2. Поиск аномалий

Второй пример задачи обучения без учителя — поиск аномалий. Необходимо обнаружить, что данный объект не похож на все остальные, то есть является аномальным.

При обучении есть примеры только обычных, не аномальных, объектов. А примеров аномальных объектов либо нет вообще, либо настолько мало, что невозможно воспользоваться классическими методами обучения с учителем (методами бинарной классификации).

1.4. Признаки в машинном обучении

Существует несколько классов, или типов признаков. И у всех свои особенности — их нужно по-разному обрабатывать и по-разному учитывать в алгоритмах машинного обучения.

Признаки описывают объект в доступной и понятной для компьютера форме. Множество значений j -го признака будет обозначаться D_j .

1.4.1. Бинарные признаки

Первый тип признаков — **бинарные** признаки. Они принимают два значения: $D_j = \{0, 1\}$.

Если ответ на вопрос да — признак полагается равным 1, если ответ на вопрос нет — то равным 0.

1.4.2. Вещественные признаки

Более сложный класс признаков — **вещественные** признаки. В этом случае $D_j = \mathbb{R}$.

Множество значений последнего указанного признака, строго говоря, является множеством натуральных чисел $\mathbb{N}$, а не $\mathbb{R}$, но такие признаки тоже считают вещественными.

1.4.3. Категориальные признаки

Следующий класс признаков — **категориальные** признаки. В этом случае D_j — неупорядоченное множество.

Отличительная особенность категориальных признаков — невозможность сравнения «больше-меньше» значений признака.

Категориальные признаки очень трудны в обращении — до сих пор появляются способы учета этих признаков в тех или иных методах машинного обучения.

1.4.4. Порядковые признаки

Частным случаем категориальных признаков являются **порядковые** признаки. В этом случае D_j — упорядоченное множество.

Хотя и порядковые, и вещественные признаки упорядочены, они отличаются тем, что в случае порядковых признаков «расстояние» между двумя значениями признака не имеет смысла. Например, отличие значения 3 от значения 2 может быть не таким существенным, как отличие 1 от 0.

1.4.5. Смешанные признаки

Смешанный признак — это такой признак, область допустимых значений которого включает фрагменты из разных типов признаков.

При работе с признаками можно столкнуться с рядом проблем. Первая из них — существование выбросов.

Выбросом называется такой объект, значение признака на котором отличается от значения признака на большинстве объектов.

Наличие выбросов представляет сложность для алгоритмов машинного обучения, которые будут пытаться учесть и их тоже. Поскольку выбросы описываются совершенно другим законом, чем основное множество объектов, выбросы обычно исключают из данных, чтобы не мешать алгоритму машинного обучения искать закономерности в данных.

Проблема может быть и в том, как распределен признак. Не всегда признак имеет такое распределение, которое позволяет ответить на требуемый вопрос. Например, может быть слишком мало данных о клиентах из небольшого города, так как собрать достаточную статистику не представлялось возможным.

