Знакомство с машинным обучением

Машинное обучение (Machine Learning) – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с проблемами *вычислительной эффективности* и *переобучения*. Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам. Многие методы тесно связаны с извлечением информации и интеллектуальным анализом данных.

Рассмотрим на примере:

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*.

Основные понятия

Объектом называется то, для чего нужно сделать предсказание (обозначается как x). В примере объектом является пара «пользователь-фильм».

Пространство объектов (X) – это множество всех возможных объектов, для которых может потребоваться сделать предсказание.

Ответом (y) будет называться то, что нужно предсказать. В данном случае ответ – понравится пользователю фильм или нет. Пространство ответов (Y), 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: $x = (x^1, x^2, \ldots, x^d)$.

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

Основные понятия

Центральным понятием машинного обучения является *обучающая выборка*:

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

Это те самые примеры, на основе которых будет строиться общая закономерность.

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

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

Операция взятия знака sign берется ввиду того, что пространство Y состоит из двух элементов. Не все алгоритмы подходят для решения задачи. Например константный алгоритм $\alpha(x)$ = 1 не подходит. Это довольно бесполезный алгоритм, который вряд ли принесет пользу сайту. Поэтому вводится некоторая характеристика качества работы алгоритма — *функционал ошибки*. $Q(\alpha, X)$ — ошибка алгоритма а 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 α, для которого достигается минимум функционала ошибки. Лучший в этом смысле алгоритм выбирается из некоторого семейства A алгоритмов.

Обучение на размеченных данных

Общая постановка задачи обучения с учителем:

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

В зависимости от множества возможных ответов Y, задачи делятся на несколько типов:

1. *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.

2. *Задача многоклассовой классификации*

В задаче многоклассовой классификации пространство ответов состоит более чем из двух ответов $Y = \{0, 1, 2, \ldots, n\}$.

Обучение на размеченных данных

Общая постановка задачи обучения с учителем:

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

В зависимости от множества возможных ответов Y, задачи делятся на несколько типов:

1. *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.

2. *Задача многоклассовой классификации*

В задаче многоклассовой классификации пространство ответов состоит более чем из двух ответов $Y = \{0, 1, 2, \ldots, n\}$.

3. *Задача регрессии*

Когда у является вещественной переменной, говорят о задаче регрессии.

Другие типы задач машинного обучения

Обучением с учителем называются такие задачи, в которых есть и объекты, и истинные ответы на них. И нужно по этим парам восстановить общую зависимость. *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* алгоритм сам определяет для каких объектов нужно узнать ответ, чтобы лучше всего обучиться.

К задачам обучения без учителя относятся:

- кластеризация,
- визуализация,
- поиск аномалий

Признаки в машинном обучении

- Бинарные признаки: принимают два значения: $D_j = \{0, 1\}$ (например, понравился ли человеку фильм?);
- Вещественные признаки: $D_j = \mathbf{R}$ (например, возраст);
- Категориальные признаки: D_j — неупорядоченное множество. Отличительная особенность категориальных признаков — невозможность сравнения «больше-меньше» значений признака (например, цвет глаз);
- Порядковые признаки: D_j — упорядоченное множество (например, роль в фильме: первый план, второй план, массовка);
- Множествозначеные признаки: значением множествозначного признака на объекте является подмножество некоторого множества (например, какие фильмы посмотрел пользователь).