СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 23. Введение в машинное обучение. Часть 1

Машинное обучение

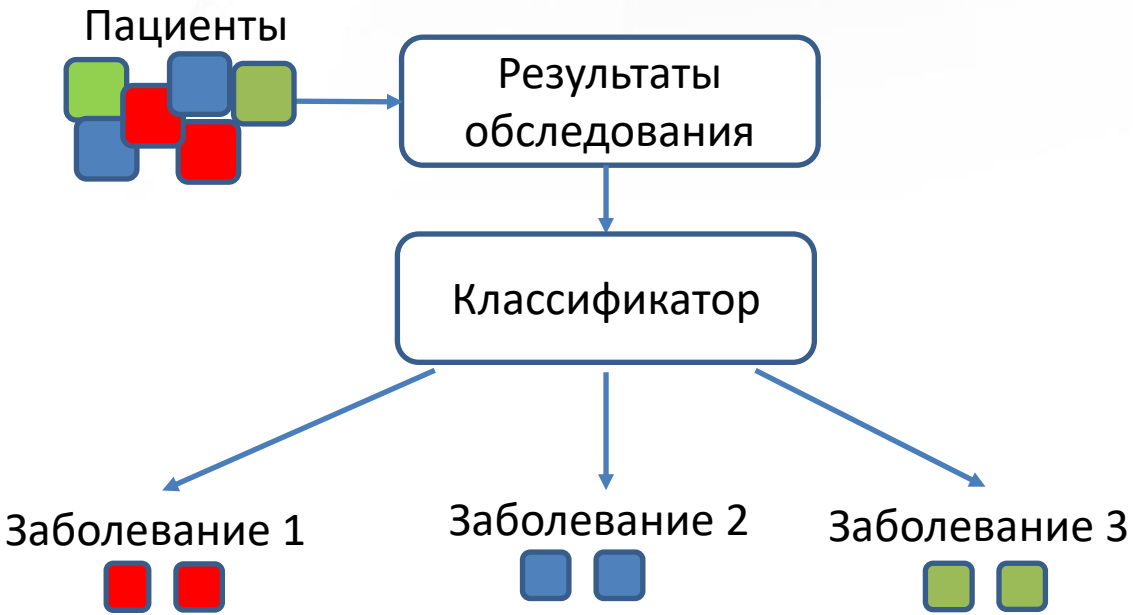
Большую часть задач машинного обучения можно разделить на:

- обучение с учителем (supervised learning)
- обучение без учителя (unsupervised learning).

Пример задачи обучения с учителем. Имеется набор данных о 10 000 квартирах в Москве. Известны параметры (признаки), расположение и прочие данные о каждой квартире. Известна стоимость каждой квартиры (ответ, отклик).
Задача состоит в построение модели, которая на основе данных признаков будет предсказывать стоимость квартиры. Эта задача обучения с учителем относится к задачам регрессии.

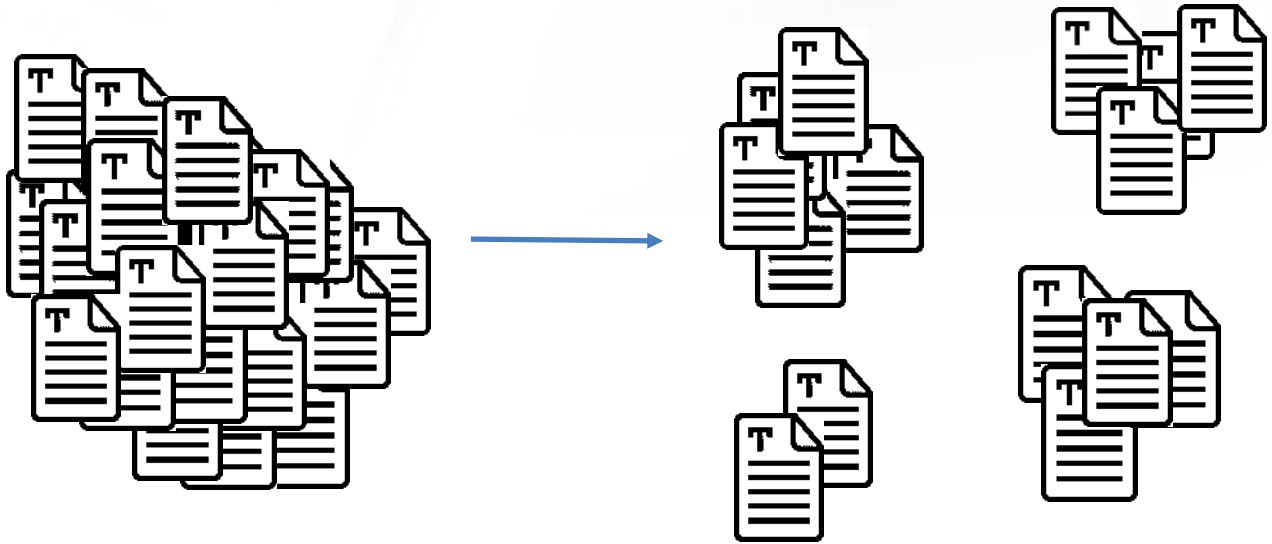
Пример задачи МО

Пример задачи классификации. Медицинская диагностика



Пример задач МО

Пример задачи кластеризации. Распределение новостей по рубрикам



Базовые понятия и обозначения

Машинное обучение (Machine Learning) — наука об алгоритмах, которые сами настраиваются на данные.
Задача **обучения по прецедентам** основана на выявлении общих закономерностей по известным частным данным.
Самая распространенная задача обучения по прецедентам – это задача **обучения с учителем**. Также рассматриваются задачи **обучения без учителя**.

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

Постановка задачи обучения по прецедентам

Дано конечное **множество прецедентов** (объектов, ситуаций), по каждому из которых собраны некоторые данные. Совокупность всех имеющихся описаний прецедентов называется **обучающей выборкой** или **учителем**. Требуется по этим частным данным выявить общие зависимости.

Этапы решение задачи обучения по прецедентам:

- фиксируется **модель восстанавливаемой зависимости** – семейство функций с параметрами;
- вводится **функционал качества**, значение которого показывает, насколько хорошо модель описывает наблюдаемые данные;
- алгоритм обучения (learning algorithm) ищет набор параметров модели, при котором функционал качества на обучающей выборке принимает оптимальное значение;
- процесс настройки (fitting) модели по выборке данных в большинстве случаев сводится к применению численных методов оптимизации.

Задача обучения с учителем

Обучение с учителем (supervised learning) — наиболее распространённый случай обучения по прецедентам. Каждый прецедент рассматривается как пара «объект, ответ». Требуется построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ. Функционал качества обычно определяется как средняя ошибка ответов, выданных алгоритмом, по всем объектам выборки.

Рассматриваемые **объекты** описываются набором признаков - вектор *x* заданной размерности. Рассматриваются:

- **Множество объектов** X, $x \in X$;
- **Множество ответов (откликов, меток, выходов)** Y;
- **Обучающая выборка** $X^l = \{x_1, \dots, x_l\} \subset X$ для которой задано **множество известных ответов** $y_i = a^*(x_i)$, где функция $a^*: X \to Y$ — неизвестная зависимость, генерирующая правильный ответ.

Задача обучения с учителем

Совокупность всех упорядоченных пар "объект-ответ" (x_i, y_i) называется **обучающей выборкой.**

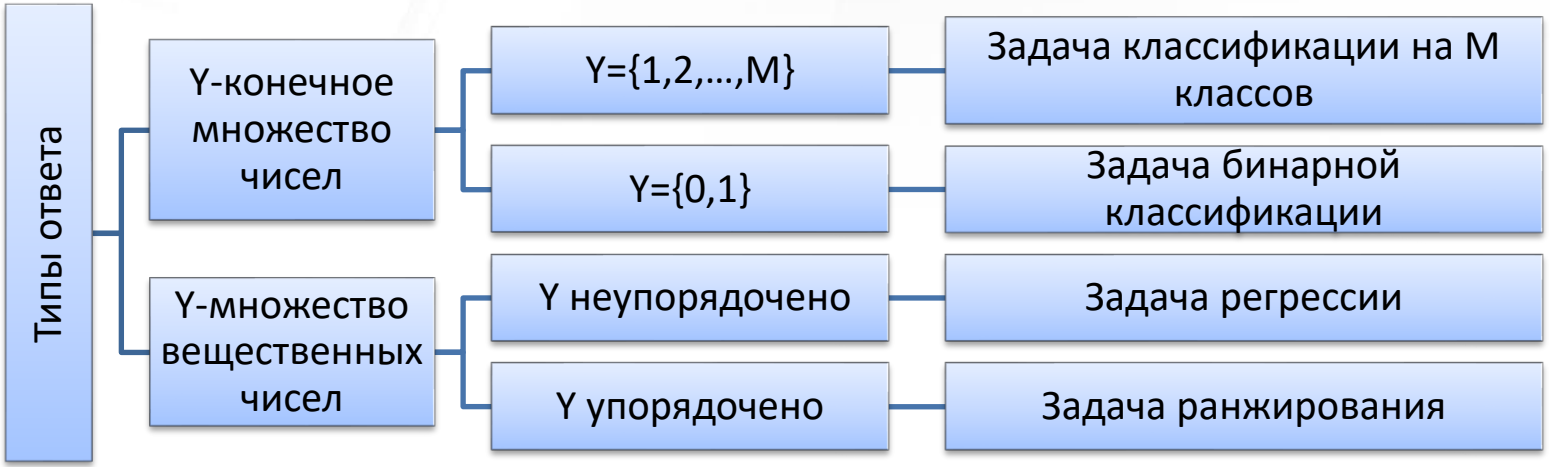
Задача. Построить **решающее правило (решающую функцию, алгоритм)** (decision function) $a: X \to Y$, которая приближала бы функцию $a^*(x)$ на всём множестве X.

Типы ответов

Ответы (отклики, метки) $y \in Y$ по типу их значения подразделяются на:

1. булевые $Y = \{0,1\}$;
2. номинальные, Y - конечное подмножество чисел *N;*
3. порядковые, - номинальные признаки, для которых определен линейный порядок;
4. количественные, значение которых - вещественное число,$Y \subset R$.

Задача обучения с учителем



Задача обучения с учителем

- **Задача классификации (classification):** множество *Y* - конечно. Ответы *y* - метки классов (class label). Класс — это множество всех объектов с данным значением метки. Если Y={0,1}, то это задача бинарной классификации;
- **Задача регрессии (regression):** ответ *y* является действительное число или числовой вектор.
- **Задача ранжирования (learning to rank):** ответы *y* получаем сразу на множестве объектов, после чего объекты сортируются по значениям ответов.
- **Задача прогнозирования (forecasting):** объекты являются временными рядами и требуется прогноз на будущее.

Функция потерь

Функция потерь — неотрицательная функция $\mathbb{L}\big(a(x), a^*(x)\big)$, характеризующая величину ошибки алгоритма a на объекте x.

Примеры функций потерь:

1. $\mathbb{L} = |a(x) - a^*(x)|$ — отклонение от правильного ответа;
2. $\mathbb{L} = (a(x) - a^*(x))^2$ — квадратичная функция потерь. Обычно применяется в задаче регрессии.
3. $\mathbb{L} = I(a(x) \neq a^*(x))$ — индикатор ошибки, обычно применяется в задачах классификации. Индикаторная функция $I(\cdot)$ равна 1, если условие выполнено и нулю в противном случае.

Задача обучения с учителем

Математическое ожидание функции потерь $R(a) = E\mathbb{L}\big(\boldsymbol{a}(x), \boldsymbol{a}^*(x)\big)$, где x – случайный вектор, распределенный на множестве X, называется **средней ошибкой, или средним риском.**

Закон распределения случайных величин x и $\boldsymbol{a}^*(x)$ как правило не известен. Вместо среднего риска рассматривают эмпирический средний риск (эмпирическую ошибку):

$$Q(a) = \frac{1}{l}\sum_{i=1}^{l} \mathbb{L}\big(a(x_i), a^*(x_i)\big)$$

Задача обучения с учителем

Задача обучения с учителем. По обучающей выборке построить **решающую функцию (алгоритм)** a: $X \to Y$, которая минимизирует эмпирический риск $Q(a)$.

Как правило **решающую функцию** ищут среди функций из некоторого параметрического семейства $a = a(x, \beta)$, где β – вектор параметров и задача сводится к определению оптимального значения вектора параметров:

$$\beta^* = \arg\min_{\beta} Q(a(x, \beta))$$

Пример. Задача регрессии

Исследуется зависимость отклика y от значений факторов $x_1, \dots, x_p$.

Наблюдаемое значение отклика y задается моделью $y = f(x) + e$

Имеется n наблюдений факторов x и отклика y – множество пар $\left(x^{(i)}, y_i\right)$.

Строится оценка неизвестной функции $f(x)$ в виде линейной функции $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$

$$\beta^* = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 .$$

Здесь X – p-мерное векторное пространство; Y – множество вещественных чисел; обучающая выборка – множество пар $\left(x^{(i)}, y_i\right)$; $a^*(x) = f(x)$, решающая функция $a(x, \beta)$ – линейная функция с параметрами β; $\mathbb{L} = (a(x) - a^*(x))^2$.

Пример. Задача распознавания спама

Задача бинарной классификации: распознавание спама.

Здесь X есть множество документов: $X = \{d\}$, $Y = \{0,1\}$, т.е. ответ может быть только «да» и «нет». Классификатор состоит из двух классов «спам» – «не спам».

Имеется некоторое обучающее множество документов $X^l = \{d_1, d_2, \dots, d_l\}$, класс которых заранее известен: $a^*(d_i) = y_i$ здесь $y_i -$ булевская переменная. Задача состоит в построении решающей функции $a(d): X \to Y$, которая сопоставляет каждый документ с одним из двух классов.