

Last version

# Анализ данных

2024-2025 учебный год

Галицкий Борис\*

16 января 2025 г.

\*I never procrastinate



# Содержание

<b>1</b>	<b>Введение в анализ данных</b>	<b>5</b>
1.1	Основные понятия . . . . .	5



# 1 Введение в анализ данных

## § 1.1. Основные понятия

Введём основные понятия на конкретном примере. Пусть нам принадлежит большая сеть пунктов выдачи заказов (ПВЗ). Нами были найдены несколько помещений, которые мы можем приобрести и организовать там ещё один ПВЗ. Нам хочется открыть его в таком месте, что прибыль в нём окажется наибольшей (из представленных вариантов).

Множество всех рассматриваемых нами помещений для открытия нового ПВЗ называется **пространством объектов** и обозначается  $X$ . Величина, которую мы хотим определять (то есть, прибыль ПВЗ), называется **целевой переменной**, а множество её значений — **пространством ответов** (и обозначается  $Y$ ).

**Замечание 1.1.1.** Для приведённого нами примера  $Y = \mathbb{R}$ .

Поскольку мы владеем большой сетью, то у нас есть данные по достаточно большому числу ранее открытых ПВЗ и по их прибыли в течение нескольких лет. Каждый такой объект (точка размещения ПВЗ) называется **обучающим**, а множество всех таких объектов и ответов для них — **обучающей выборкой**, которая обозначается  $X = \{(x_1, y_1), \dots, (x_l, y_l)\}$ , где  $x_1, \dots, x_l$  — обучающие объекты,  $y_1, \dots, y_l$  — ответы для них,  $l$  — их количество.

**Замечание 1.1.2.** Объекты — это некоторые абстрактные сущности (в данном случае ПВЗ), которые компьютеры явно представлять не умеют. Для дальнейшего анализа их необходимо описать при помощи некоторого набора характеристик, которые называются **признаками (факторами)**. Вектор всех признаков объекта  $x$  называется **признаковым описанием** этого объекта.

**Замечание 1.1.3.** Объект и его признаковое описание будем считать эквивалентными.

Признаки могут быть:

- бинарными;
- вещественными;
- категориальными (принимают значения из неупорядоченного множества);
- ординальными (принимают значения из упорядоченного множества);
- множественными (set-valued — значения являются подмножествами некоторого универсального множества).

Признаки могут иметь сложную структуру: например, в качестве признака может быть фото. Его, безусловно, можно представить как некоторое количество бинарных или вещественных признаков, каждый из которых соответствует пикселю фото. Однако, есть много специфичных особенностей работы с изображениями (уменьшение размерности, цветовой гаммы и прочее). Специализируется на работе со сложными данными **глубинное обучение (deep learning)**.

Для приведённого нами примера могут быть полезными признаки, связанные с демографией (средний возраст жителей ближайших кварталов, динамика изменения их количества и состава) или недвижимостью (средняя стоимость квадратного метра в окрестности, количество школ, магазинов, банков, торговых центров и динамика их количества). Разработка признаков

(**feature engineering**) для любой задачи является одним из самых нетривиальных и самых важных этапов анализа данных.

Данная задача является примером **обучения с учителем (supervised learning)**, а если точнее — задачей **регрессии** — так называются задачи с действительной целевой переменной.

Другие примеры задачи обучения с учителем:

- $Y = \{0; 1\}$  — бинарная классификация: например, можно предсказать, является ли письмо спамом или нет.
- $Y = \{1; \dots; K\}$  — многоклассовая (multi-class classification) классификация: например, можно предсказать, по фотографии цифры определить, какая цифра написана.
- $Y = \{0; 1\}^K$  — многоклассовая классификация с пересекающимися классами (multi-label classification): например, можно определять часть речи слова (некоторые слова могут быть одновременно нескольких частей речи: течь — глагол (протекать) и существительное (протекание)).
- частичное обучение (semi-supervised learning) — задача, в которой для одной части объектов обучающей выборки известны и признаки, и ответы, а для другой только признаки: например, когда в задаче постановки диагноза требуется дорогостоящий анализ.

Также существует обучение без учителя (unsupervised learning) — класс задач, где ответы неизвестны или вообще не существуют. В этом случае требуется найти некоторые закономерности в данных на основе признаковых описаний:

- Кластеризация — задача разделения объектов на группы, обладающие некоторыми свойствами: например, определить локации, где люди чаще всего садятся в такси.
- Оценивание плотности — задача приближения распределения объектов: например, определение аномалий поведения игроков на бирже (зная лишь примеры «правильного» поведения).
- Визуализация.

### Основные этапы решения задачи анализа данных:

1. Постановка задачи.
2. Выделение признаков.
3. Формирование выборки.
4. Выбор метрики качества.
5. Предобработка данных.
6. Построение модели.
7. Оценивание качества модели.