Last version

Анализ данных

2024-2025 учебный год

Галицкий Борис*

12 января 2025 г.

*I never procrastinate

Содержание

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 $\mathbb{X}$. Величина, которую мы хотим определять (то есть, прибыль ПВЗ), называется **целевой переменной**, а множество её значений — **пространством ответов** (и обозначается $\mathbb{Y}$).

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

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 $X = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, где x_1, …, x_l — обучающие объекты, y_1, …, 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.

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.