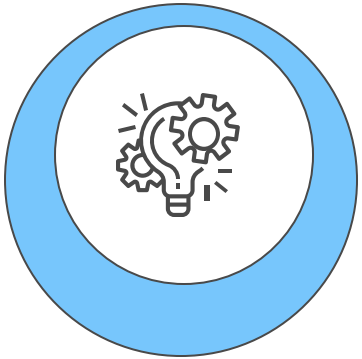
Логистическая регрессия

Авторы: Солодова София
Штыхно Илья
Мазепа Илья

01
Введение



Логистическая регрессия — это метод классификации, который предсказывает вероятность принадлежности объекта к одному из двух классов.
Она широко используется в медицине, экономике, биоинформатике и других областях.

Цель:
Разобраться в принципах работы логистической регрессии, её математике, методах обучения и реализовать модель самостоятельно.

02

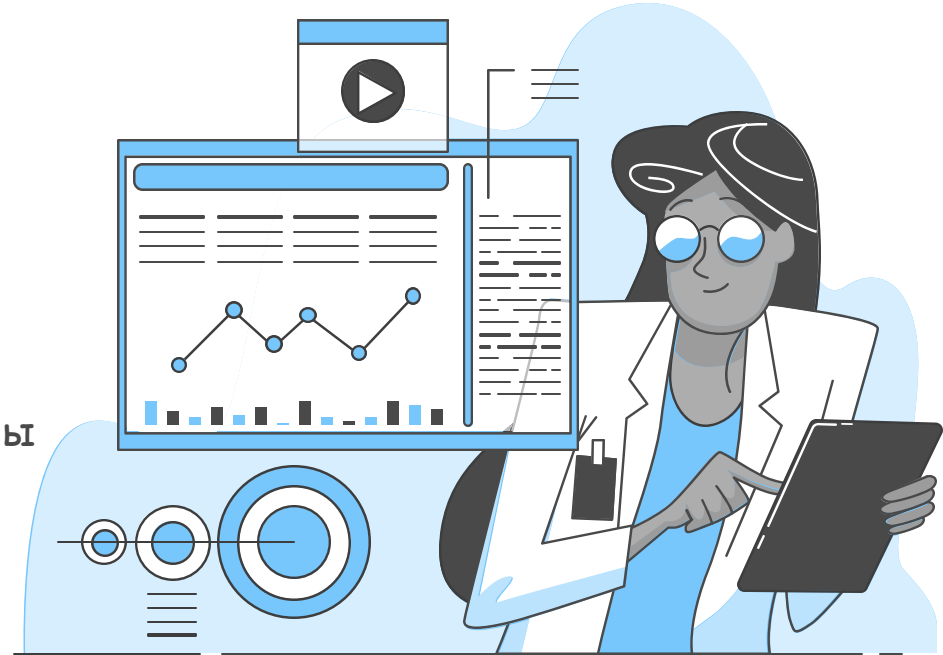
Задача логистической регрессии

Основная цель модели – определить вероятность того, что объект относится к классу 1 (а не 0)

Примеры задач:

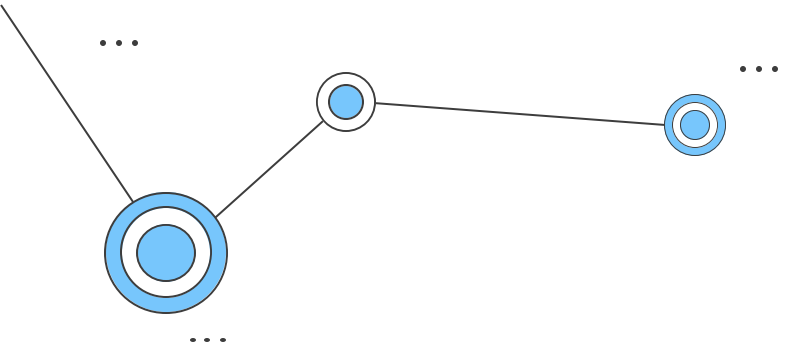
- Спам или не спам
- Болен или здоров
- Покупка или отказ

Почему не линейная регрессия?
Потому что её прогнозы не ограничены диапазоном [0, 1], и она не подходит для вероятностных интерпретаций.



03

Идея модели

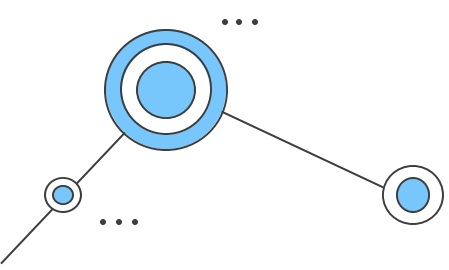


Модель вычисляет линейную комбинацию признаков:

$$z = \theta^T x$$

и преобразует её в вероятность с помощью сигмоиды:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



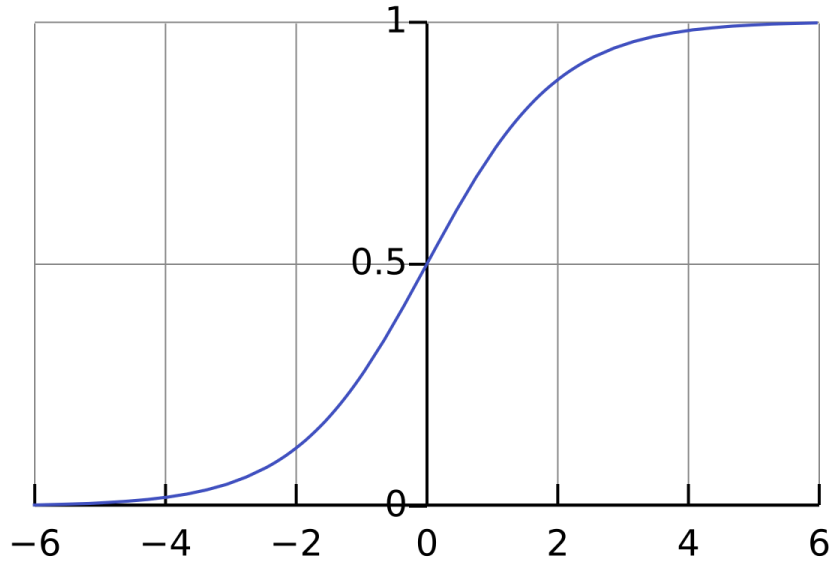
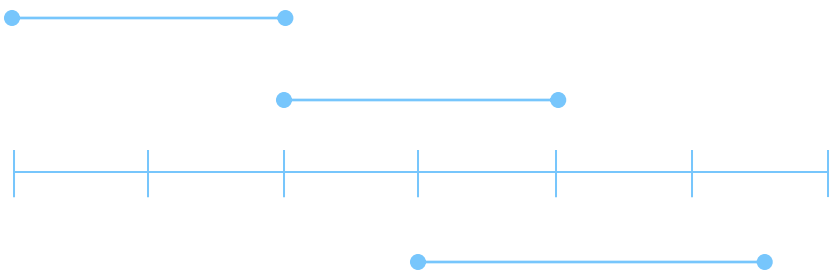


Рис. 1 Функция активации — сигмоида



04

Математическая формулировка

Модель предсказывает:

$$h_\theta(x) = \sigma(\theta^T x)$$

Логарифм отношения шансов (log-odds) выражается как:

$$\log \frac{P(y=1)}{1 - P((y=1))} = \theta^T x$$

Интерпретация:
Каждый коэффициент показывает, как изменение признака влияет на вероятность перехода в класс 1.

05

Функция потерь и обучение

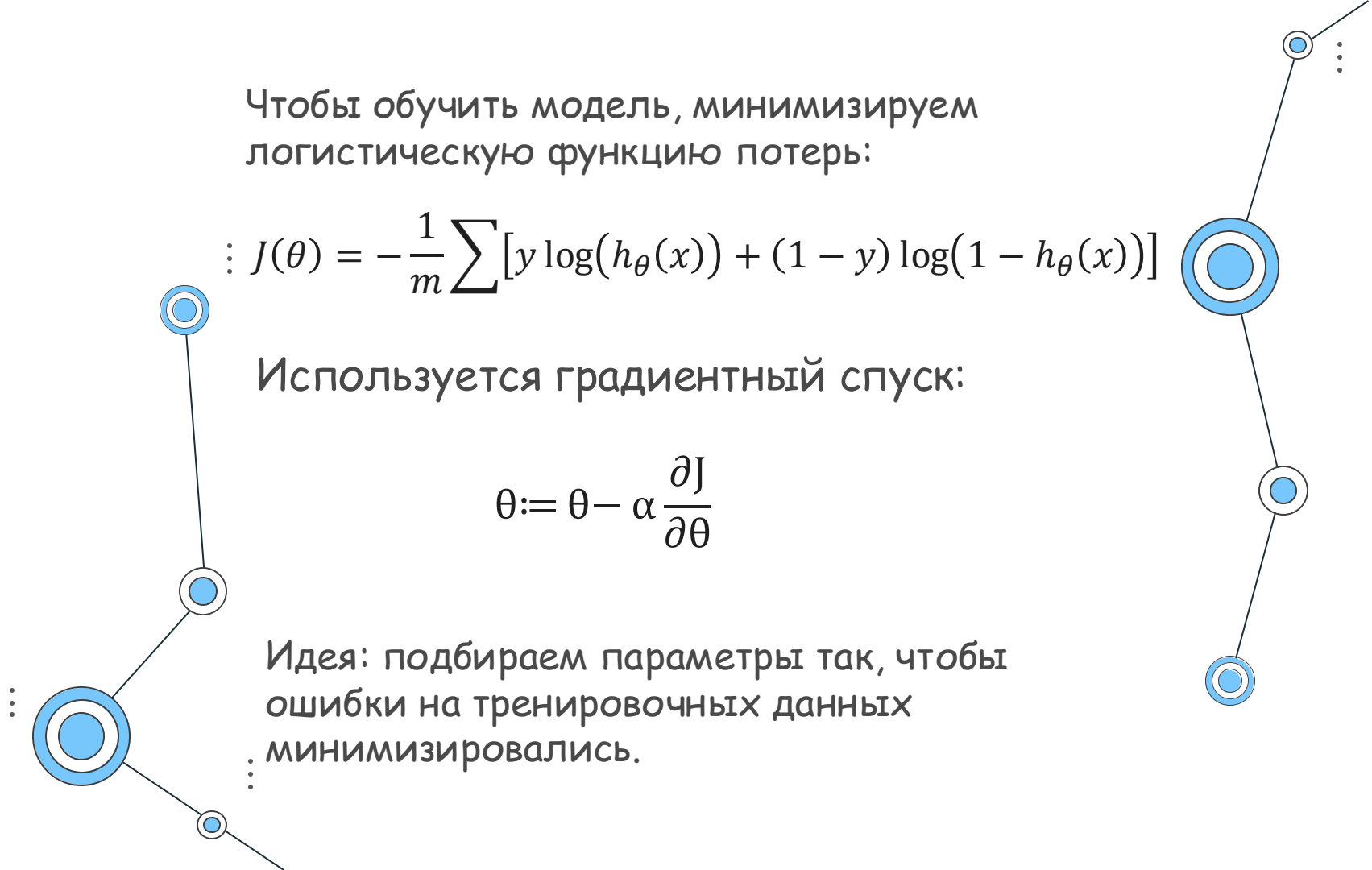
Чтобы обучить модель, минимизируем логистическую функцию потерь:

$$J(\theta) = -\frac{1}{m}\sum\left[y\log\left(h_\theta(x)\right) + (1-y)\log\left(1-h_\theta(x)\right)\right]$$

Используется градиентный спуск:

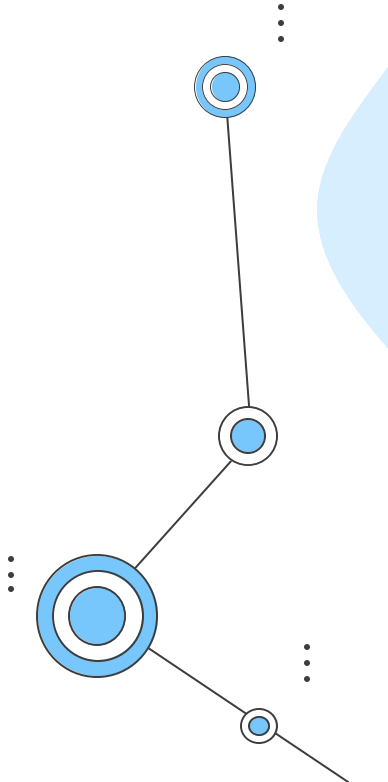
$$\theta := \theta - \alpha\frac{\partial J}{\partial \theta}$$

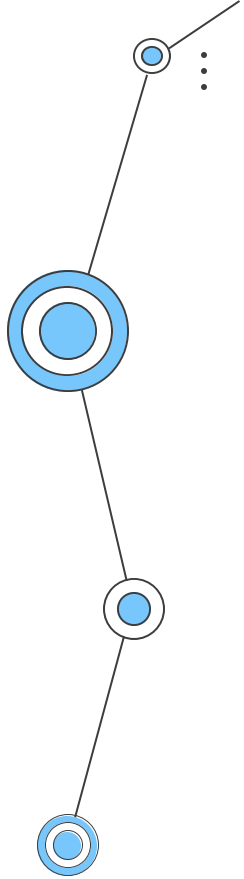
Идея: подбираем параметры так, чтобы ошибки на тренировочных данных минимизировались.



06

Регуляризация





Чтобы избежать переобучения, добавляется штраф за большие веса:

$$J_{reg}(\theta) = J(\theta) + \lambda|\theta|^2$$

- L1-регуляризация (Lasso): зануляет лишние параметры
- L2-регуляризация (Ridge): сглаживает веса

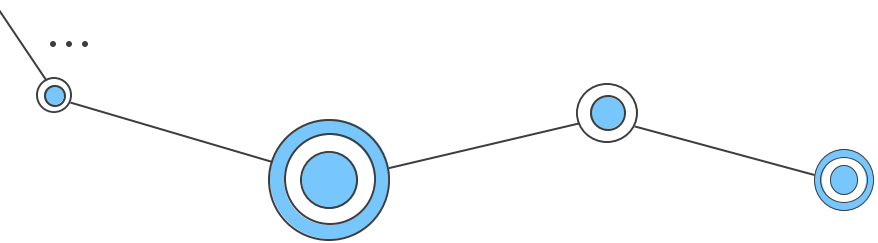
Помогает модели быть более устойчивой и обобщающей.

07

Интерпретация результатов

После обучения модель выдаёт вероятности.
Чтобы получить метку, задаётся порог классификации (обычно 0.5).

Оценка качества:
- Accuracy — доля правильных предсказаний
- Precision, Recall, F1 — более точные метрики для несбалансированных данных
- ROC и AUC — показывают качество различения классов



08

Сравнение с другими моделями

Модель	Преимущества	Недостатки
Логистическая регрессия	Простая, интерпретируемая	Только линейные границы
Дерево решений	Учитывает нелинейности	Может переобучаться
SVM	Работает при сложных разделениях	Трудно интерпретировать
Random Forest	Высокая точность	Меньше интерпретируемости

Логистическая регрессия часто используется как базовая модель для сравнения с более сложными алгоритмами.

09

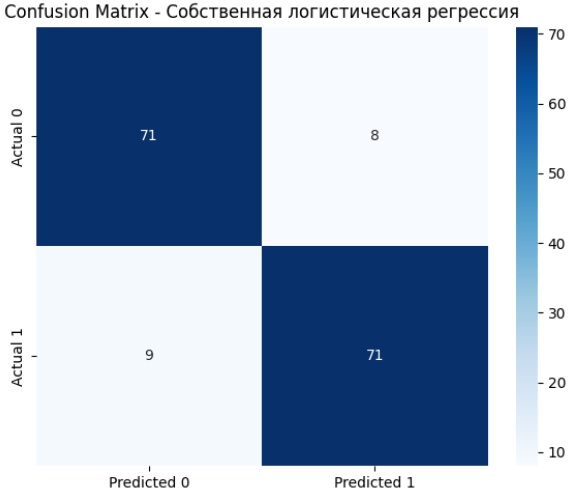
Реализация собственной логистической регрессии

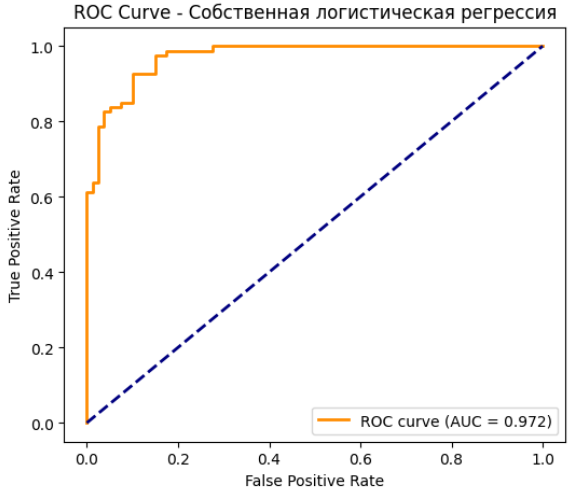
Основные шаги

1. Инициализация параметров (все нули или случайные значения)
2. Прямой проход — вычисление предсказаний через сигмоиду
3. Расчёт функции потерь
4. Градиентный спуск — обновление параметров
5. Проверка качества модели

Этот цикл повторяется до сходимости, пока потери не перестанут уменьшаться.

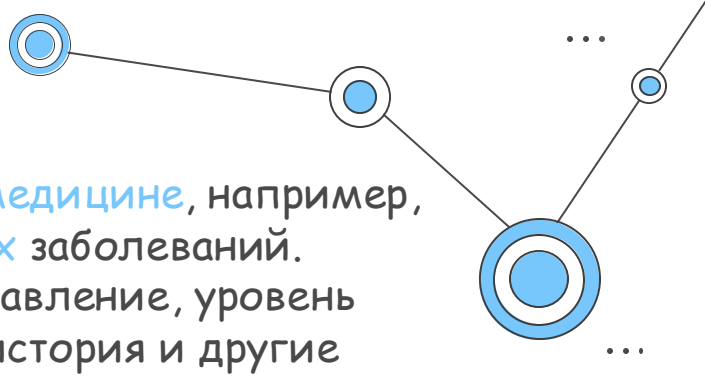
```python
import numpy as np

class LogisticRegression:
    def __init__(self, lr=0.01, epochs=1000):
        self.lr = lr
        self.epochs = epochs

    def sigmoid(self, z):
        return 1 / (1 + np.exp(-z))

    def fit(self, X, y):
        self.theta = np.zeros(X.shape[1])
        for _ in range(self.epochs):
            z = np.dot(X, self.theta)
            h = self.sigmoid(z)
            grad = np.dot(X.T, (h - y)) / len(y)
            self.theta -= self.lr * grad

    def predict(self, X):
        return (self.sigmoid(np.dot(X, self.theta)) >= 0.5).astype(int)
```





10

Пример применения



Логистическую регрессию часто используют в медицине, например, для прогнозирования риска сердечно-сосудистых заболеваний.
В качестве признаков берутся возраст, кровяное давление, уровень холестерина, наличие болей в груди, семейная история и другие параметры.
Модель позволяет:

- ✓ Рассчитывать индивидуальную вероятность: пациенту можно сказать, что риск перенести заболевание — X%.
- ✓ Понять, какие факторы наиболее существенно влияют на прогноз.
- ✓ Упростить принятие решений для врача благодаря понятной интерпретации коэффициентов.

Благодаря способности интерпретировать причины прогноза, логистическая регрессия завоевала доверие в клинической практике

11

Ограничения модели

Ограничения модели

Логистическая регрессия справляется только с линейными границами между классами — если взаимосвязи в данных сложные, точность будет низкой.

Перед обучением признаки обязательно нужно масштабировать и стандартизировать.

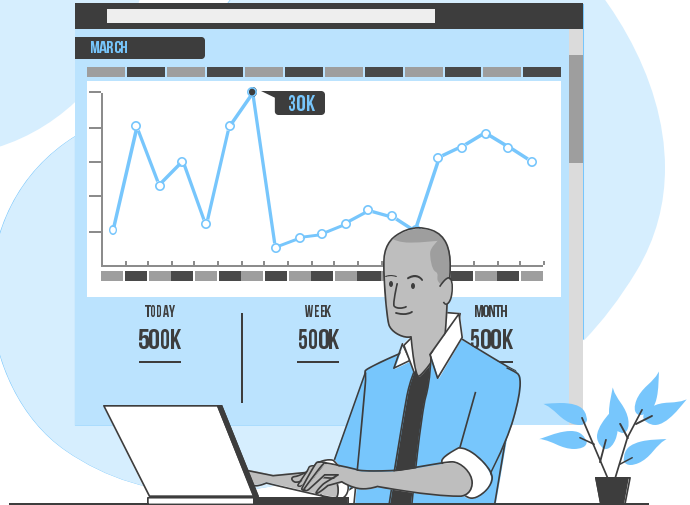
Для несбалансированных выборок результаты могут быть неадекватны, приходится дополнительно учитывать метрики (Recall, Precision).

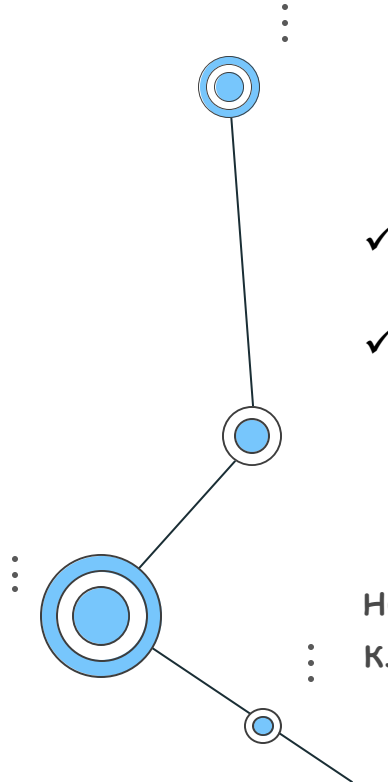
Модель предполагает, что каждый признак влияет на логит (логарифм шансов) линейно.

По этим причинам её часто используют как отправную точку, а затем пробуют более сложные алгоритмы.

12

Итоги





- ✓ Логистическая регрессия — простая, быстрая и прозрачная модель бинарной классификации.
- ✓ Она использует сигмоиду для преобразования линейной комбинации признаков в вероятность, а учится с помощью логарифмической функции потерь.
- ✓ Модель легко анализировать, проверять и применять на практике.
- ✓ Является отличной основой для изучения оптимизации, нейросетей, и построения сложных ансамблей.

«Логистическая регрессия — это фундамент, на котором строится вся современная машинная классификация!»