

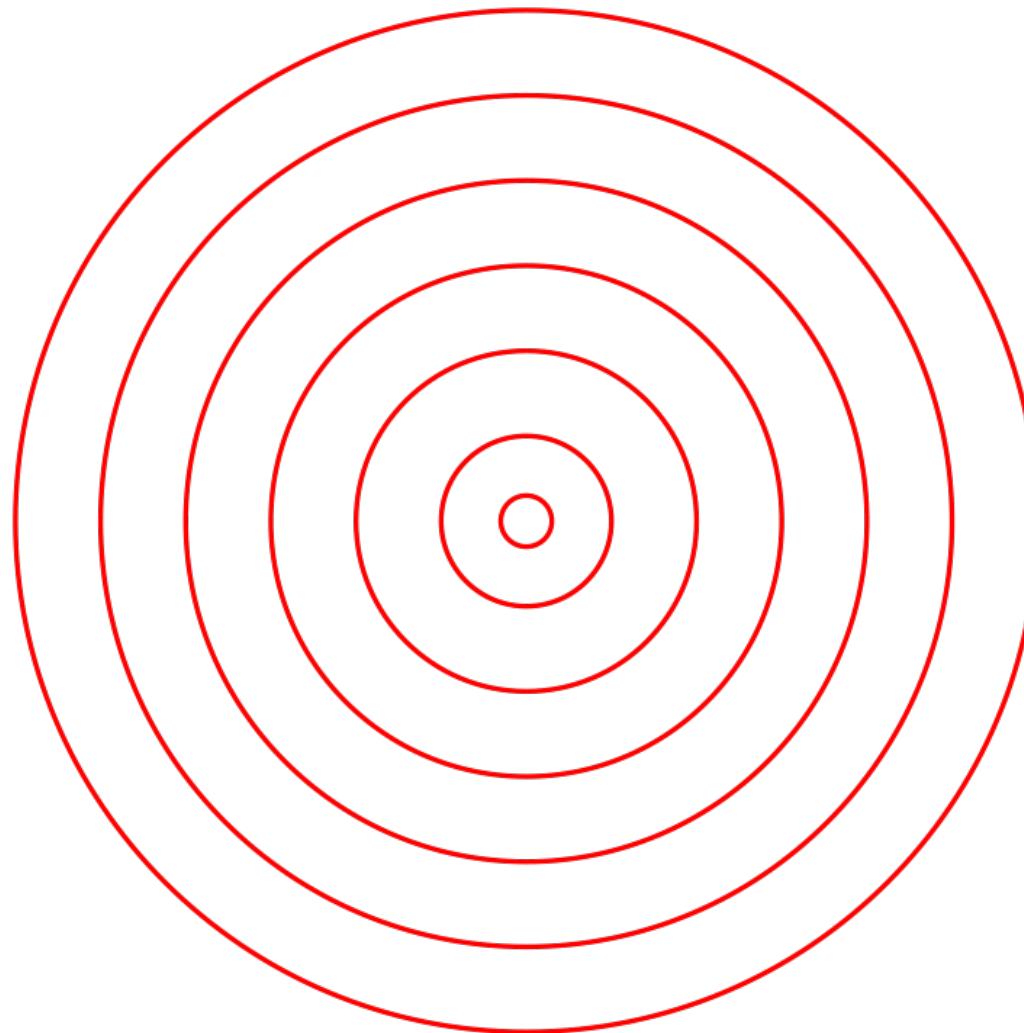
МАСТАБИРОВАНИЕ ПРИЗНАКОВ

МАСШТАБ

- › Две неизвестных величины
- › Задача:

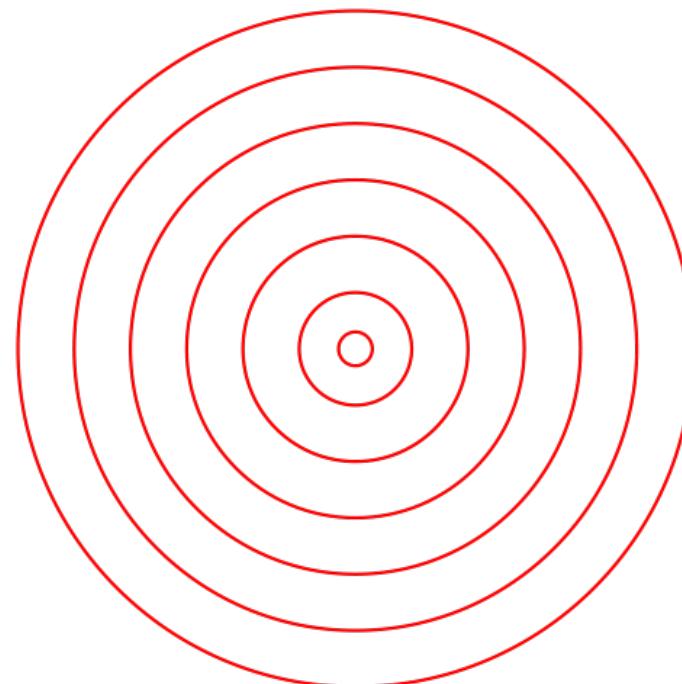
$$w_1^2 + w_2^2 \rightarrow \min_w$$

ЛИНИИ УРОВНЯ



ГРАДИЕНТНЫЙ СПУСК

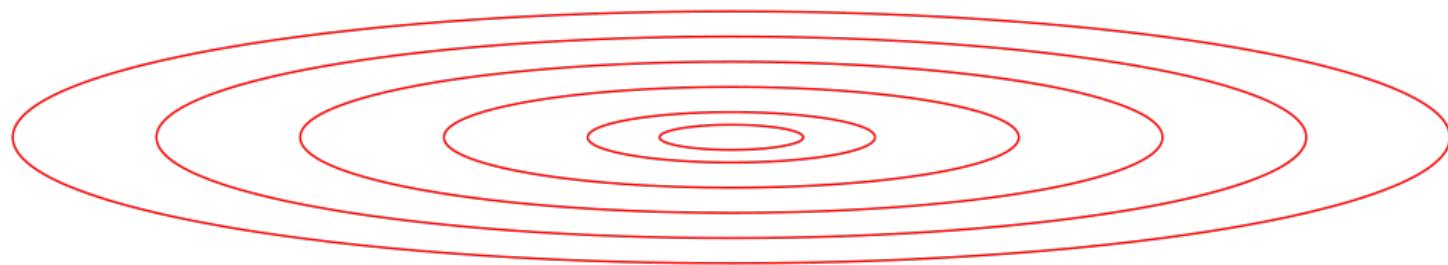
- › Инициализация: $w = (1, 1)$
- › Антиградиент: $(-2, -2)$



МАСШТАБ

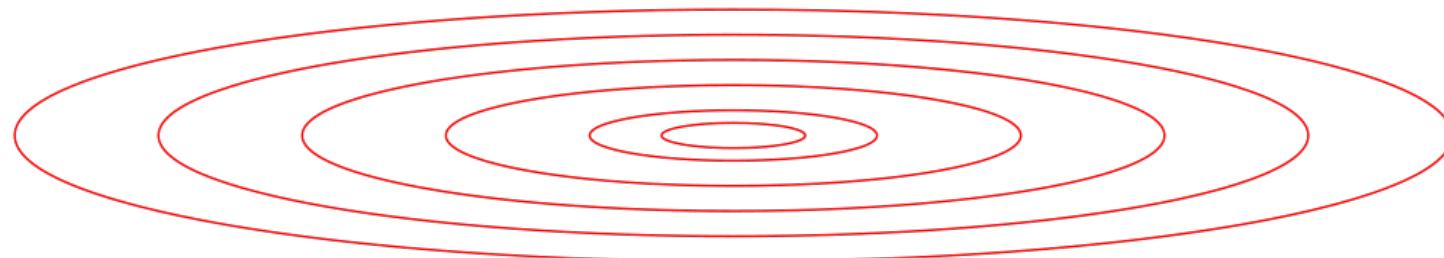
- › Две неизвестных величины
- › Задача: $w_1^2 + 100w_2^2 \rightarrow \min_w$

ЛИНИИ УРОВНЯ

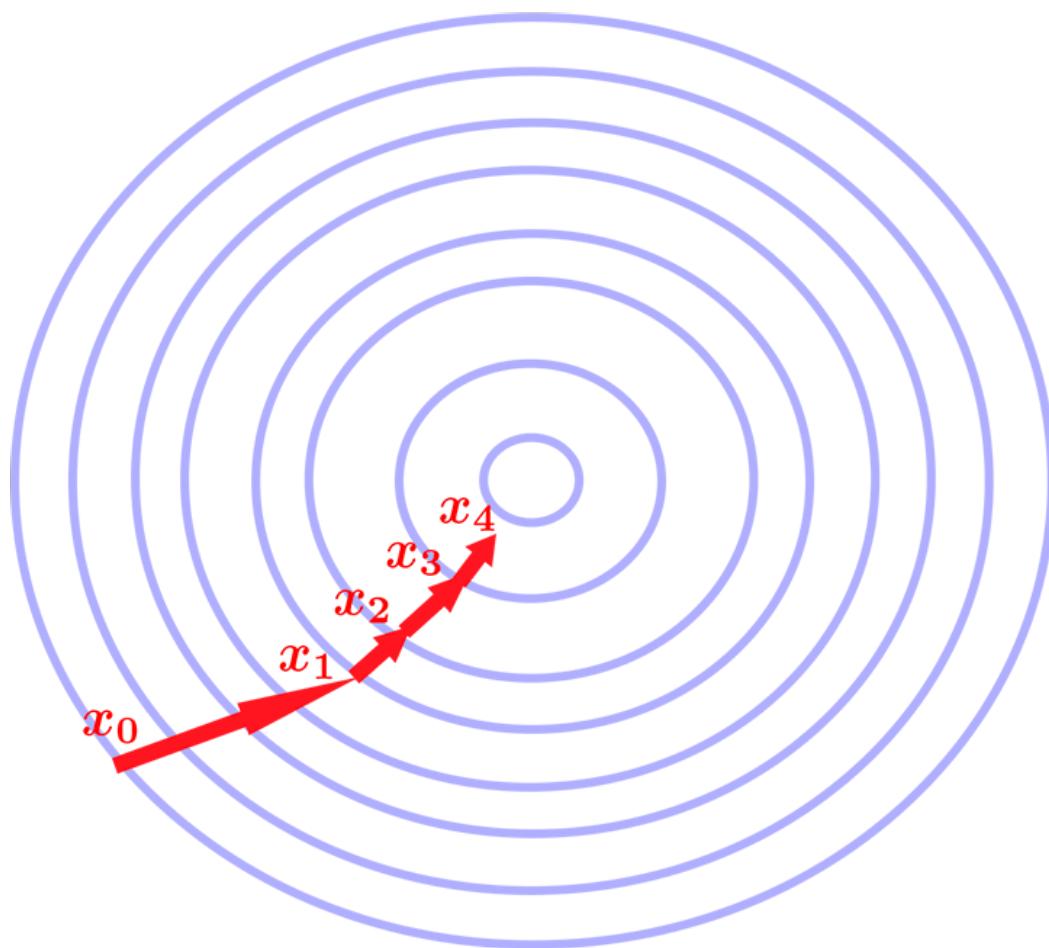


ГРАДИЕНТНЫЙ СПУСК

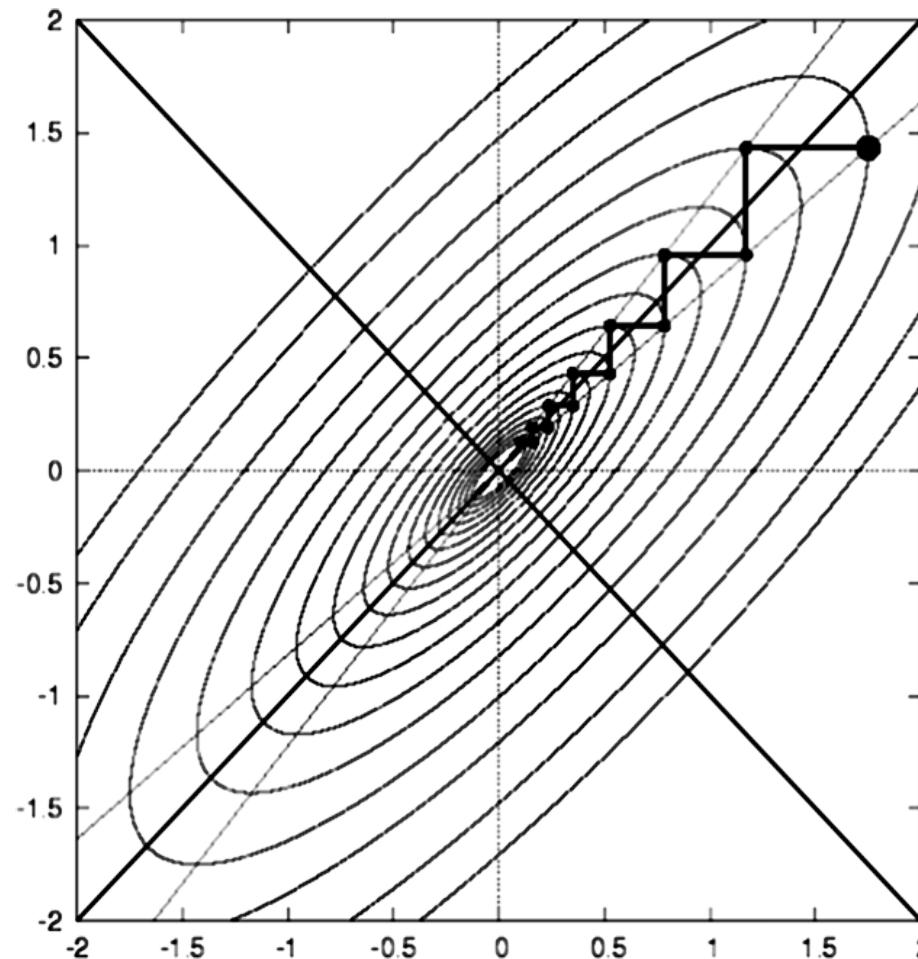
- › Инициализация: $w = (1, 1)$
- › Антиградиент: $(-2, -200)$



МАСШТАБИРОВАНИЕ ВЫБОРКИ



МАСШТАБИРОВАНИЕ ВЫБОРКИ



МАСШТАБИРОВАНИЕ ВЫБОРКИ

- › Задача: одобрят ли заявку на грант?
- › Признак 1: сколько успешных заявок было до этого у заявителя
- › Признак 2: год рождения заявителя

МАСШТАБИРОВАНИЕ ВЫБОРКИ

- › Среднее значение и стандартное отклонение признака на выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (x_i^j - \mu_j)^2}$$

МАСШТАБИРОВАНИЕ ВЫБОРКИ

› Стандартизация:

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j}$$

МАСШТАБИРОВАНИЕ ВЫБОРКИ

- Минимальное и максимальное значение признака на выборке:

$$m_j = \min(x_1^j, \dots, x_\ell^j)$$

$$M_j = \max(x_1^j, \dots, x_\ell^j)$$

МАСШТАБИРОВАНИЕ ВЫБОРКИ

› Масштабирование на [0, 1]:

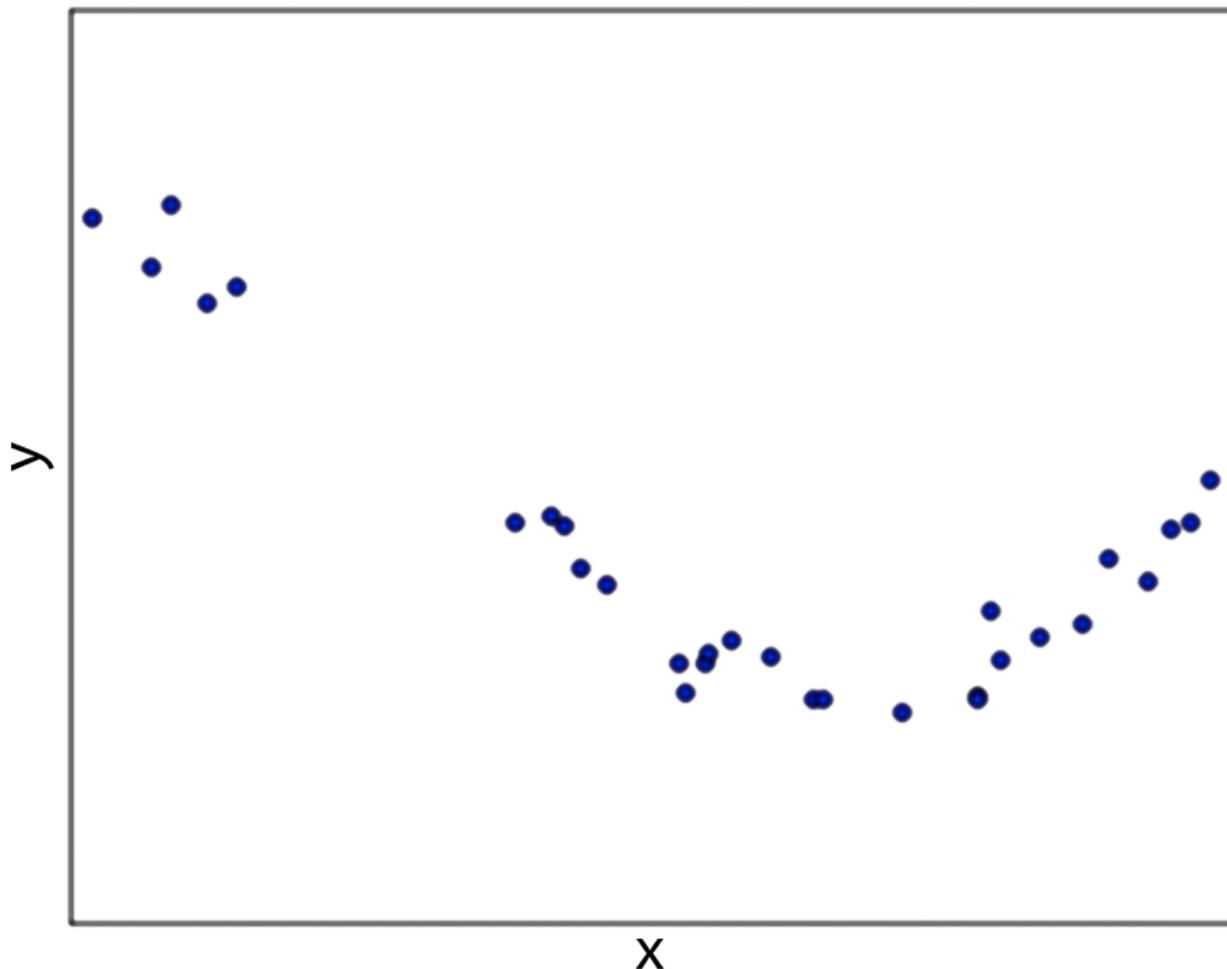
$$x_i^j := \frac{x_i^j - m_j}{M_j - m_j}$$

РЕЗЮМЕ

- › Разный масштаб признаков плохо влияет на сходимость градиентных методов
- › Два способа масштабирования:
 - ▶ Нормализация
 - ▶ Масштабирование на отрезок [0, 1]

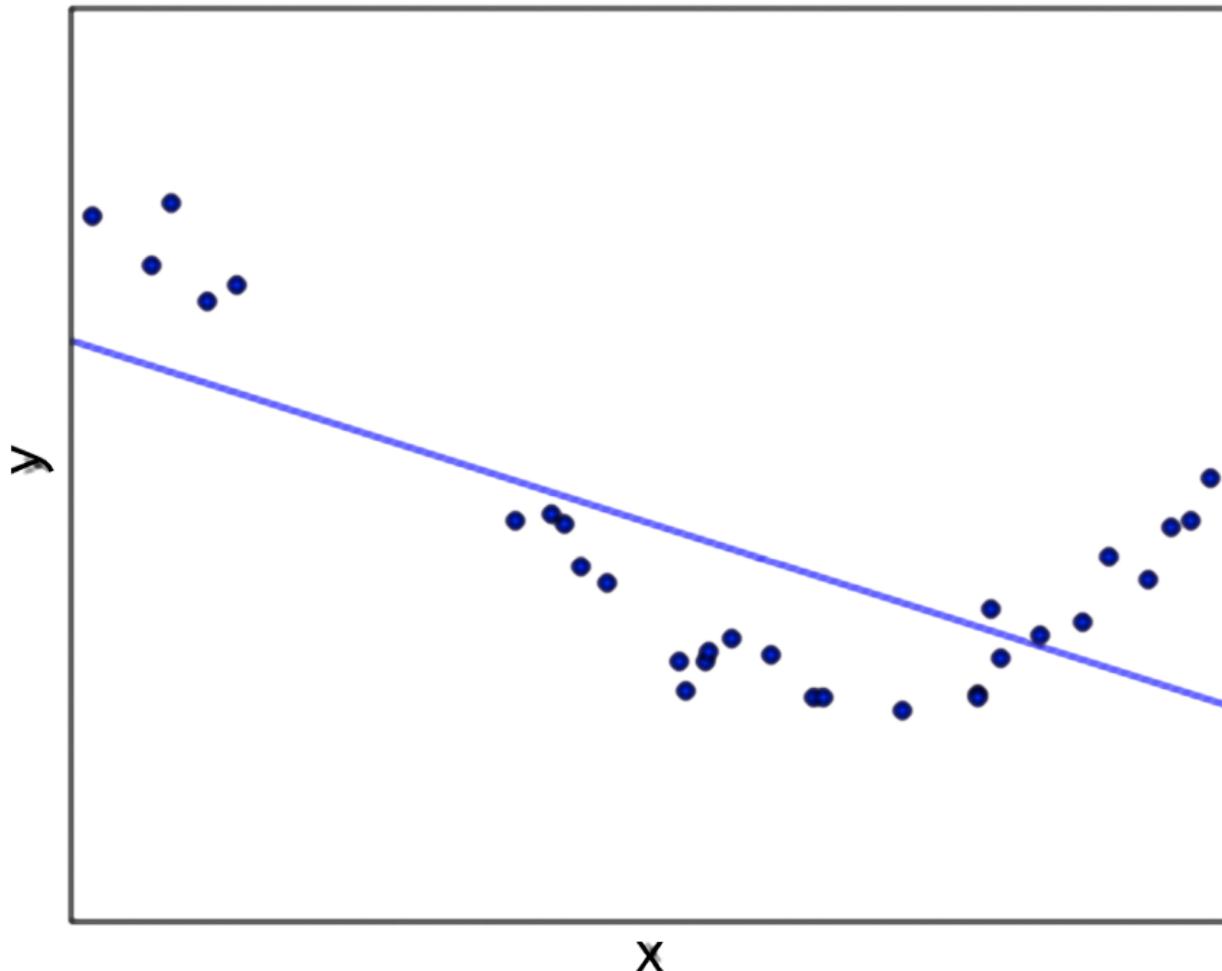
СПРЯМЛЯЮЩИЕ ПРОСТРАНСТВА

НЕЛИНЕЙНЫЕ ЗАВИСИМОСТИ



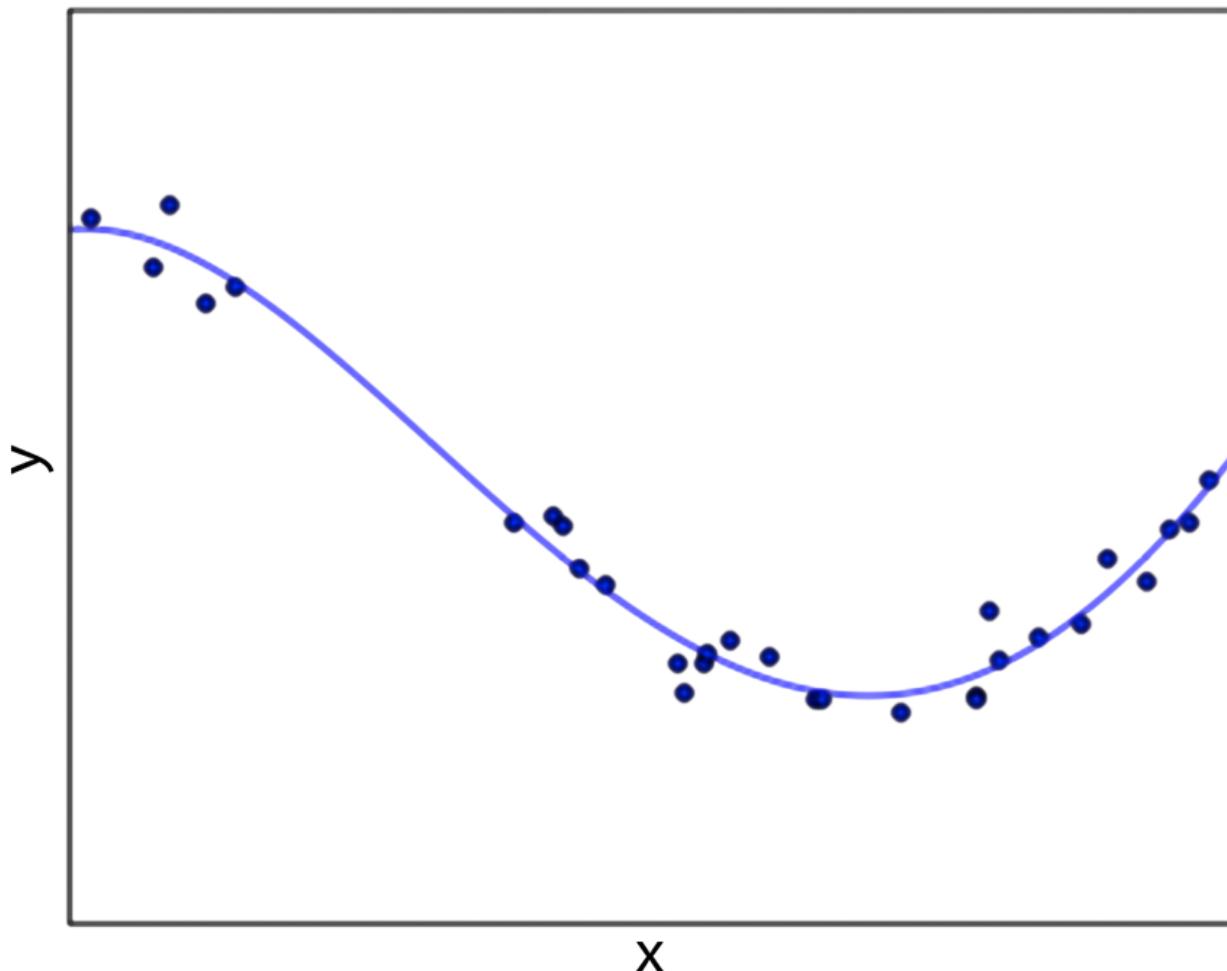
НЕЛИНЕЙНЫЕ ЗАВИСИМОСТИ

» $a(x) = w_0 + w_1 x$

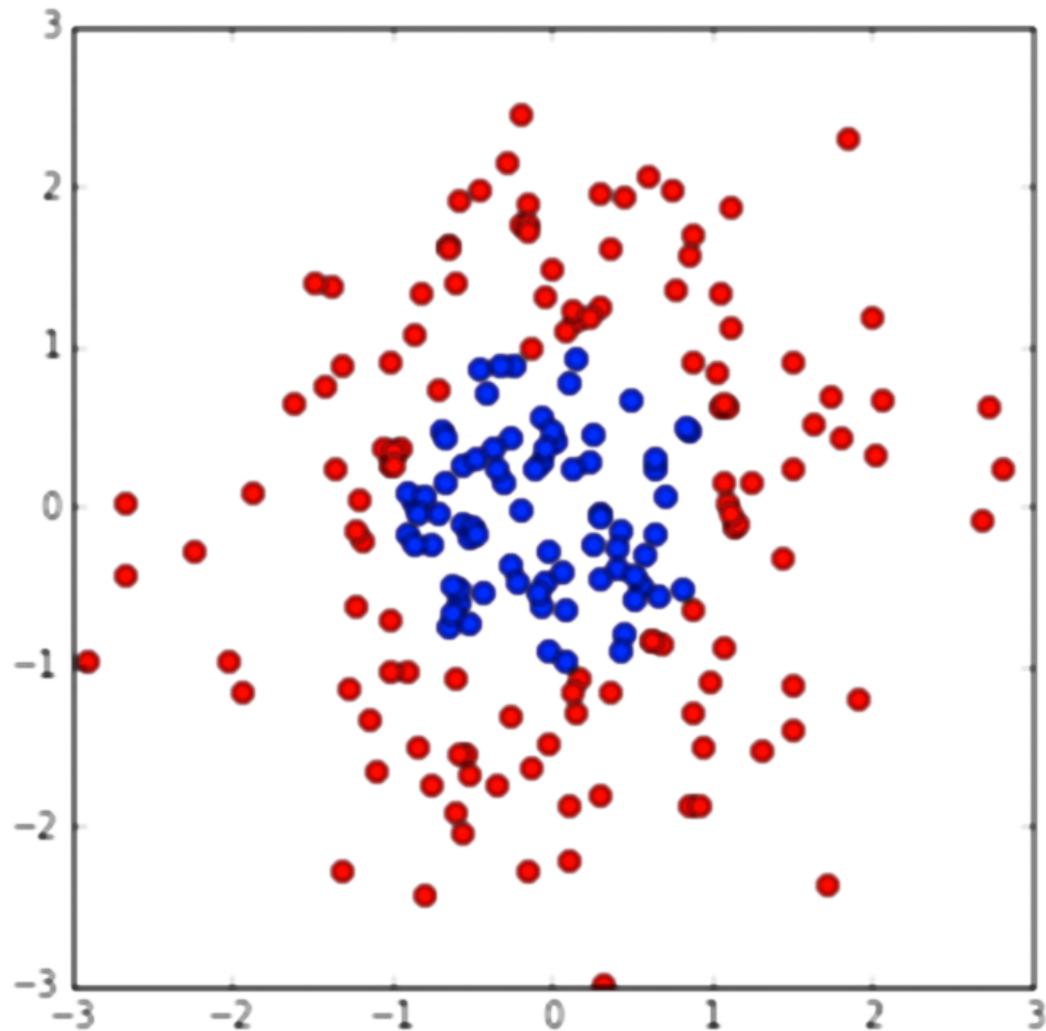


НЕЛИНЕЙНЫЕ ЗАВИСИМОСТИ

» $a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$

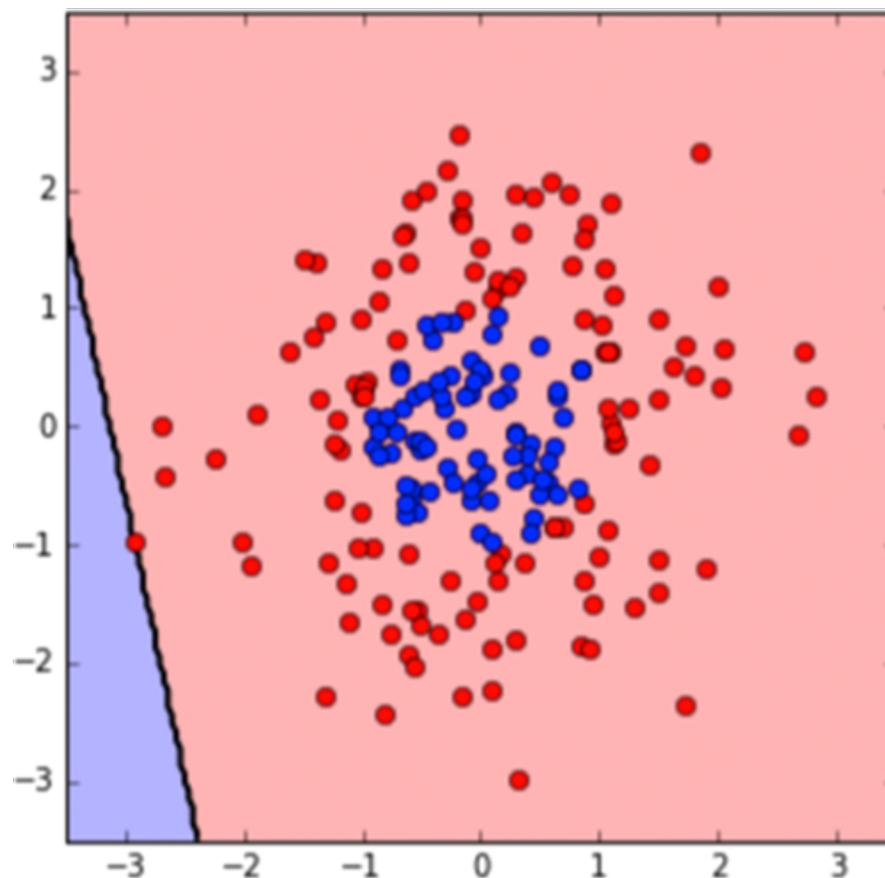


НЕЛИНЕЙНЫЕ ЗАВИСИМОСТИ



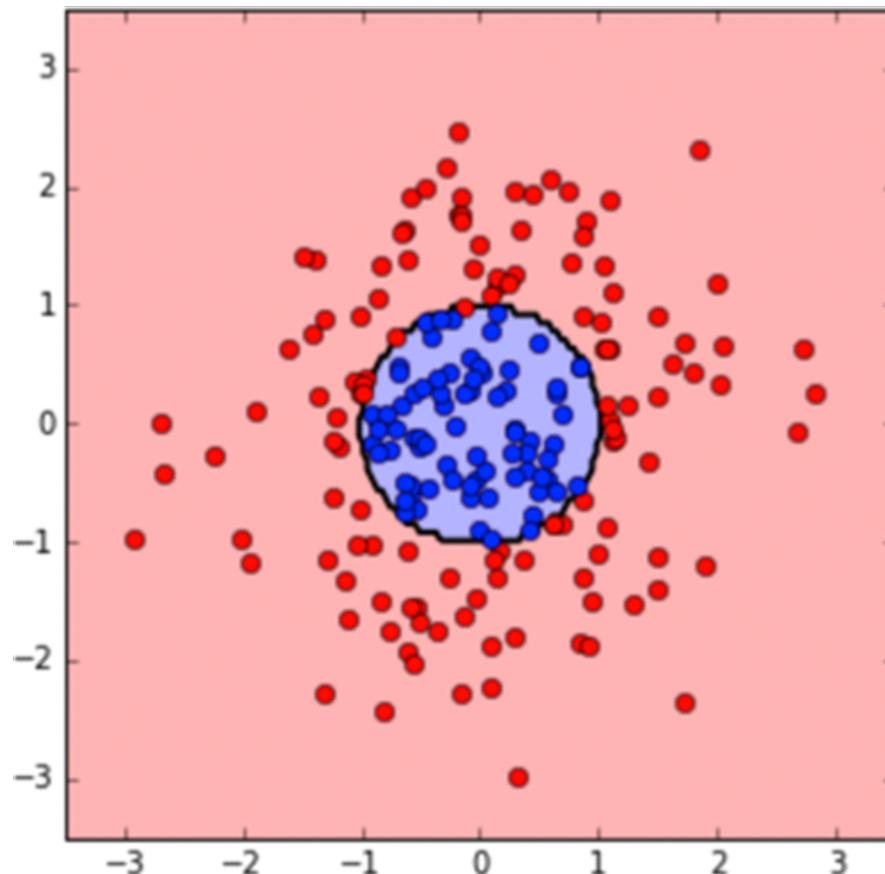
НЕЛИНЕЙНЫЕ ЗАВИСИМОСТИ

» $a(x) = \text{sign}(w_0 + w_1 x_1 + w_2 x_2)$



НЕЛИНЕЙНЫЕ ЗАВИСИМОСТИ

» $a(x) = \text{sign}(w_0 + w_1 x_1 + w_2 x_2 +$
 $+ w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2)$



СПРЯМЛЯЮЩЕЕ ПРОСТРАНСТВО

- › Пространство, в котором задача хорошо решается линейной моделью

СПРЯМЛЯЮЩЕЕ ПРОСТРАНСТВО

» Пример: квадратичные признаки

$$(x_1, \dots, x_d) \rightarrow (x_1, \dots, x_d, x_1^2, \dots, x_d^2, x_1 x_2, \dots, x_{d-1} x_d)$$

СПРЯМЛЯЮЩЕЕ ПРОСТРАНСТВО

» Пример: полиномиальные признаки

$$(x_1, \dots, x_d) \rightarrow (x_1, \dots, x_d, \dots, x_i x_j, \dots, x_i x_j x_k, \dots)$$

СПРЯМЛЯЮЩЕЕ ПРОСТРАНСТВО

› Пример: логарифмирование

$$x_i \rightarrow \ln(x_i + 1)$$

$$x_i \rightarrow \ln(|x_i| + 1)$$

СПРЯМЛЯЮЩЕЕ ПРОСТРАНСТВО

- › Не все задачи решаются линейными моделями
- › Но это можно исправить переходом в спрямляющее пространство
- › Переход — порождение новых признаков по исходным
- › Примеры: полиномы, логарифмы, другие нелинейные функции

РАБОТА С КАТЕГОРИАЛЬНЫМИ ПРИЗНАКАМИ

КАТЕГОРИАЛЬНЫЕ ПРИЗНАКИ

- › Город
- › Цвет
- › Тарифный план
- › Марка автомобиля
- › ...

КАТЕГОРИАЛЬНЫЕ ПРИЗНАКИ

- › Элементы неупорядоченного множества
- › Нельзя умножать на число и складывать

БИНАРНОЕ КОДИРОВАНИЕ

- › Категориальный признак $f_j(x)$
- › n возможных значений
- › Пронумеруем их: c_1, c_2, \dots, c_n
- › Заведем n бинарных признаков:

$$b_1(x), \dots, b_n(x)$$

- › $b_i(x) = [f_j(x) = c_i]$

ПРИМЕР

- › Признак принимает значения:
{синий, зелёный, красный}

- › Три объекта:

$$f_j(x_1) = \text{синий}$$

$$f_j(x_2) = \text{красный}$$

$$f_j(x_3) = \text{синий}$$

ПРИМЕР

- › Признак принимает значения:
{синий, зелёный, красный}
- › Кодирование:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

НОВЫЕ ЗНАЧЕНИЯ

- › Что, если признак принял новые значения на тестовой выборке?
- › Игнорируем
- › $b_1(x) = \dots = b_n(x) = 0$

РЕЗЮМЕ

- › Категориальные признаки нельзя непосредственно использовать в линейных моделях
- › Бинарное кодирование

НЕСБАЛАНСИРОВАННЫЕ ДАННЫЕ

НЕСБАЛАНСИРОВАННАЯ ВЫБОРКА

- › Задача классификации
- › Объектов одного из классов существенно меньше, чем объекты остальных классов
- › Бинарная классификация: объектов одного из классов менее 10%

ПРИМЕР

- › Предсказание резких скачков курса доллара
- › Единицы примеров за всю историю
- › Практически каждый день — отрицательный пример

ПРИМЕРЫ

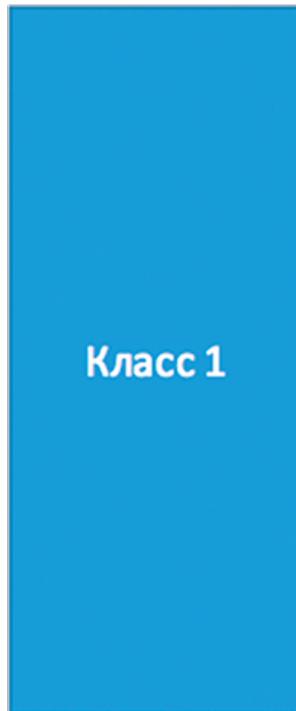
- › Медицинская диагностика
- › Обнаружение мошеннических транзакций
- › Классификация текстов

ПРОБЛЕМА

- › Классификаторы минимизируют число неправильных ответов
- › Цена ошибки на каждом классе одинаковая
- › Может быть выгоднее предсказывать самый популярный класс

UNDERSAMPLING

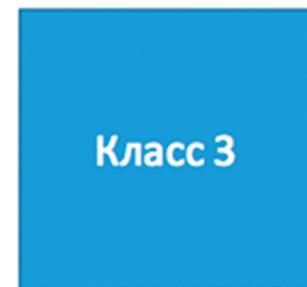
- › Зачем нам так много объектов одного класса?



Класс 1



Класс 2



Класс 3

UNDERSAMPLING

- › Зачем нам так много объектов одного класса?



Класс 2



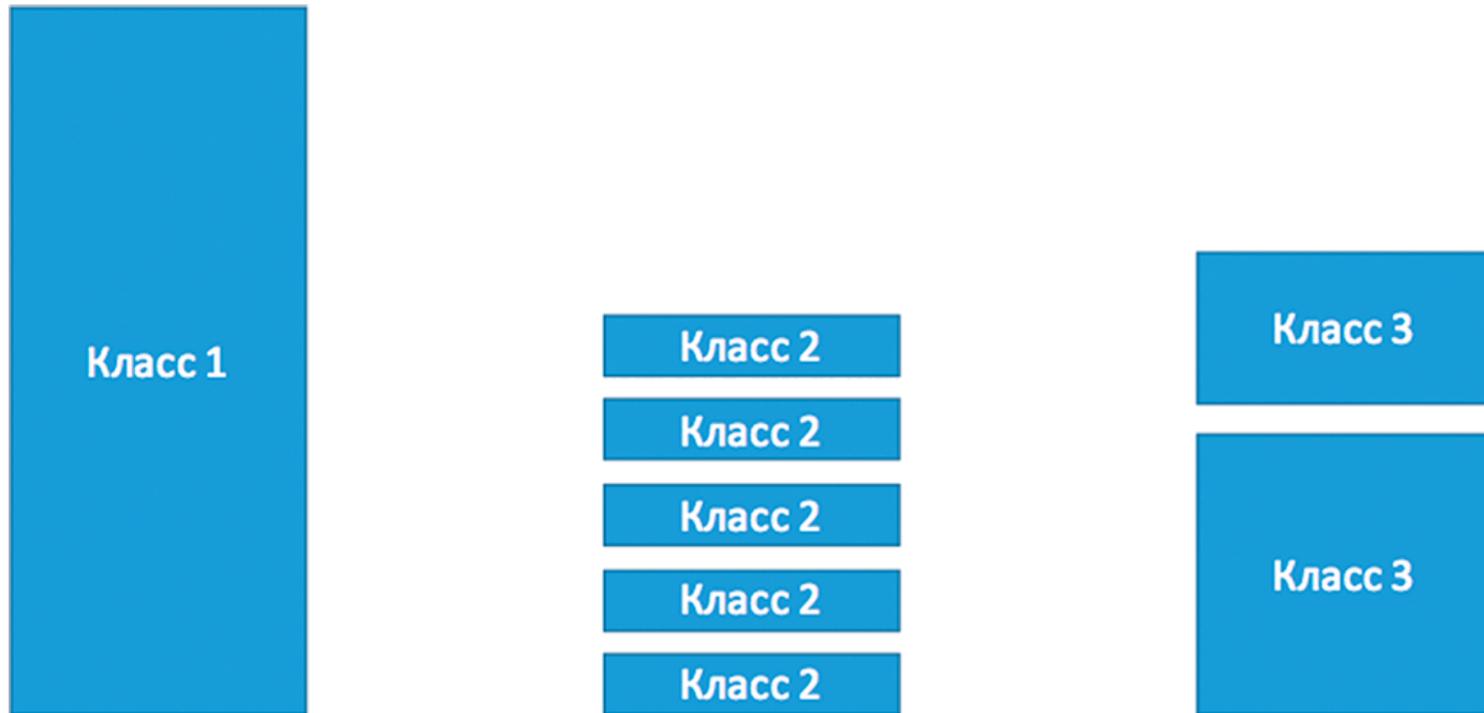
OVERSAMPLING

- › Дублируем объекты маленьких классов



OVERSAMPLING

- › Дублируем объекты маленьких классов



OVERSAMPLING

$$\text{MSE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

OVERSAMPLING

- › Дублирование соответствует выставлению весов

$$\text{MSE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \nu_i (a(x_i) - y_i)^2$$

СТРАТИФИКАЦИЯ

- › Кросс-валидация разбивает выборку на блоки
- › Нужно разбивать так, чтобы в каждом блоке сохранялось соотношение классов



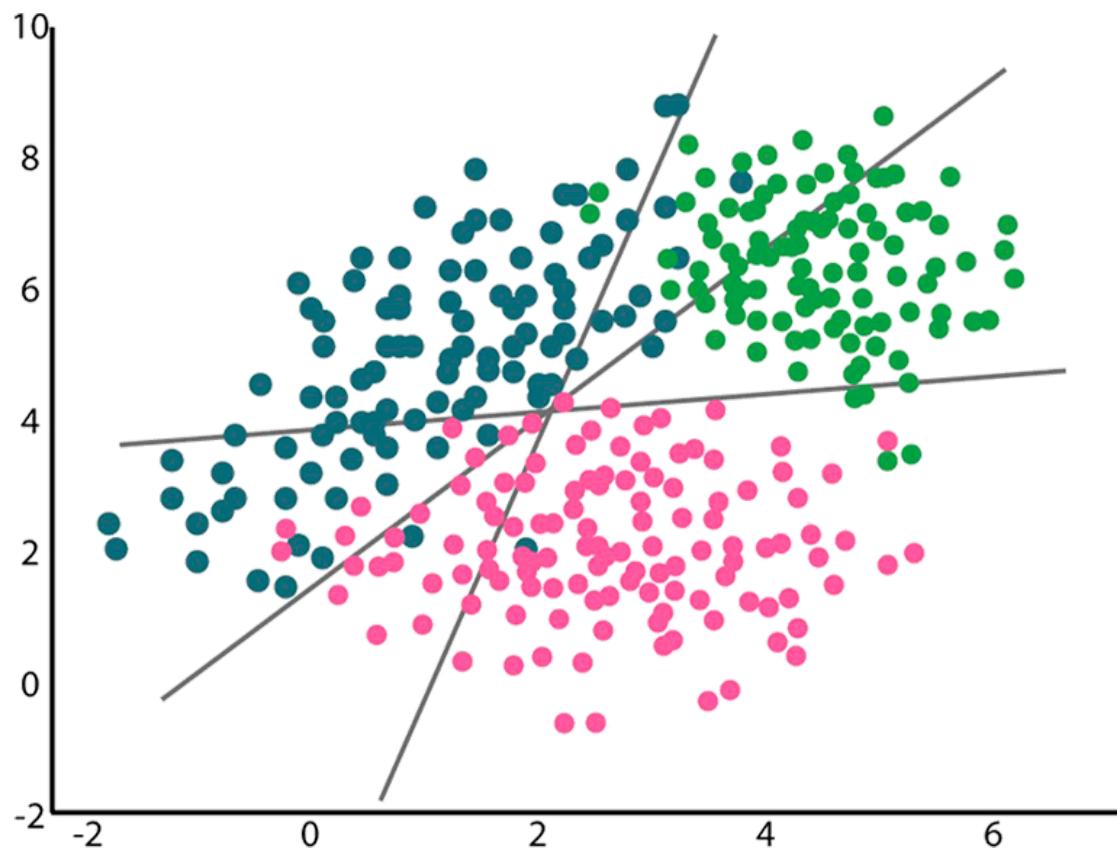
РЕЗЮМЕ

- › На несбалансированных выборках могут быть проблемы при классификации
- › Oversampling и undersampling
- › Стратификация

МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ

МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ

» $\mathbb{Y} = \{1, 2, \dots, K\}$



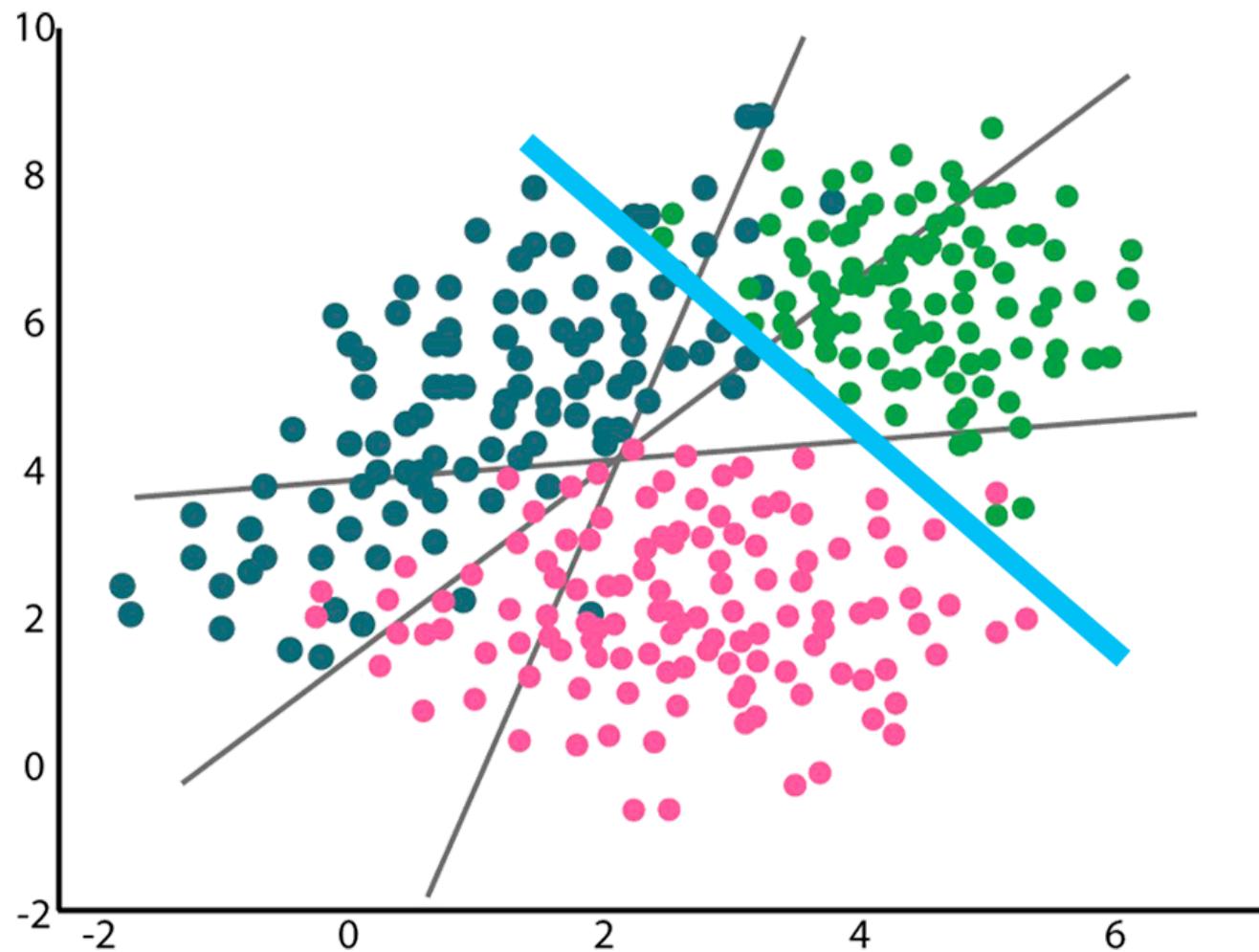
БИНАРНАЯ КЛАССИФИКАЦИЯ

» $a(x) = \text{sign}\langle w, x \rangle$

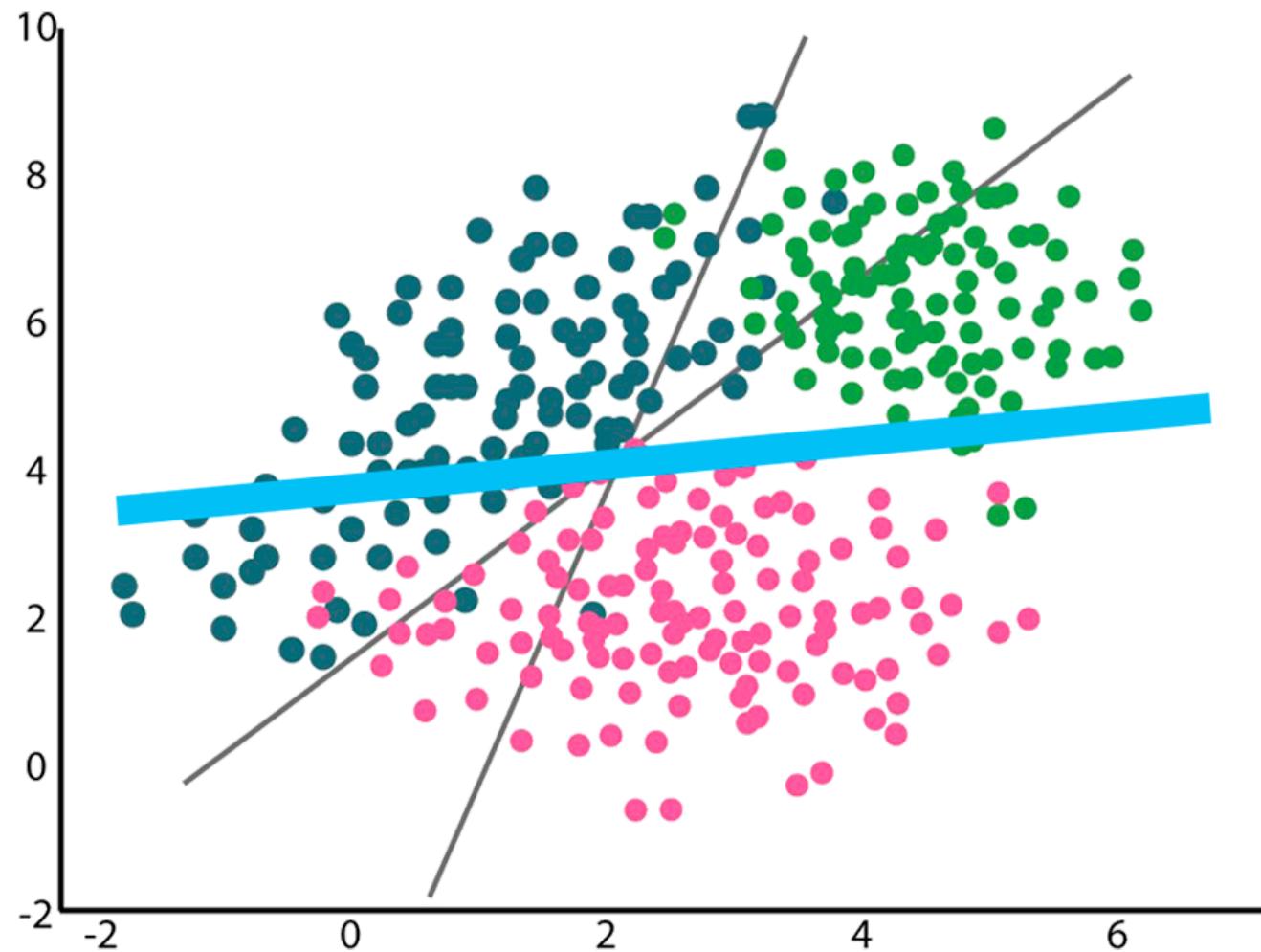
ONE-VS-ALL

- › Обучаем свой классификатор для каждого класса
- › Задача: отделение класса от всех остальных

ONE-VS-ALL



ONE-VS-ALL



ONE-VS-ALL

- » K задач бинарной классификации
- » k -я задача:
 - $X = (x_i, [y_i = k])_{i=1}^\ell$
 - Классификатор $a_k(x) = \text{sign}\langle w_k, x \rangle$
- » Алгоритм: $a(x) = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} \langle w_k, x \rangle$

МАТРИЦА ОШИБОК

	$y = 1$	$y = 2$	\cdots	$y = K$
$a(x) = 1$	q_{11}	q_{12}	\cdots	q_{1K}
$a(x) = 2$	q_{21}	q_{22}	\cdots	q_{2K}
\cdots	\cdots	\cdots	\cdots	\cdots
$a(x) = K$	q_{K1}	q_{K2}	\cdots	q_{KK}

ДОЛЯ ПРАВИЛЬНЫХ ОТВЕТОВ

$$\text{accuracy}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

ТОЧНОСТЬ И ПОЛНОТА

- › Относительно каждого класса
- › Можно усреднить точность и полноту по всем классам
- › Можно усреднить F -меру

РЕЗЮМЕ

- › От многоклассовой задачи можно перейти к нескольким бинарным
- › Матрица ошибок
- › Точность, полнота, F -мера