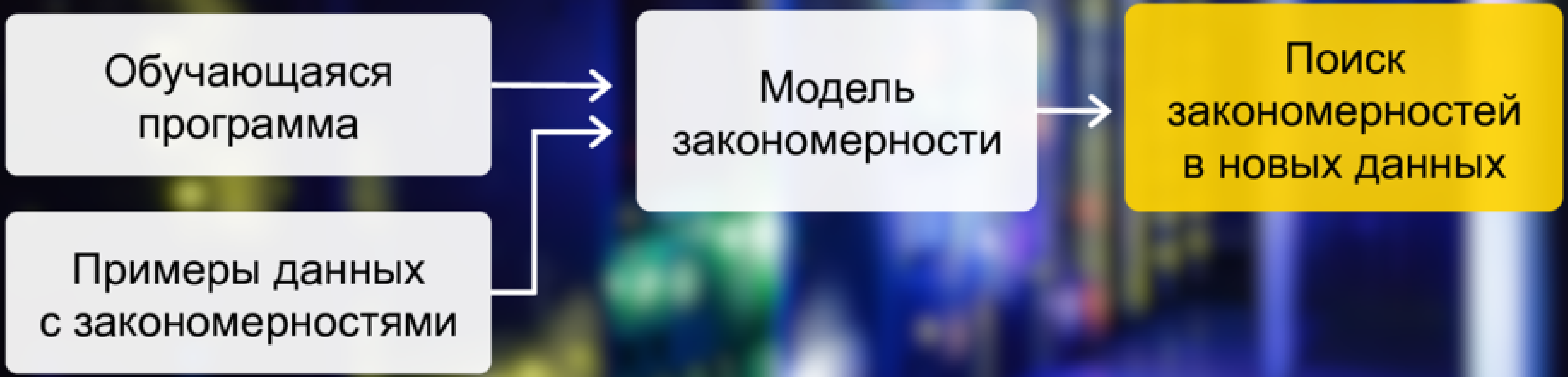
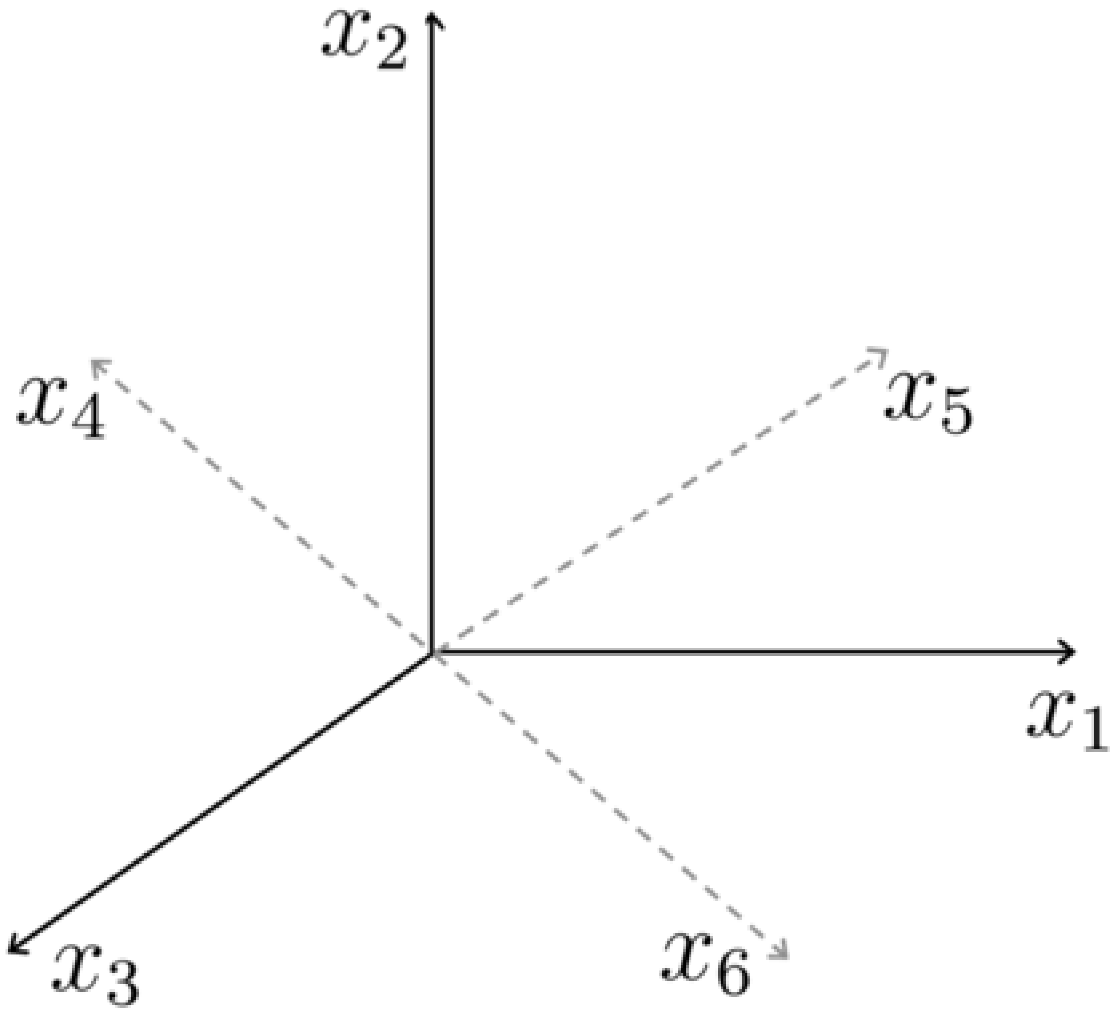
Что такое машинное обучение

Лена Бунина

Главная идея машинного обучения



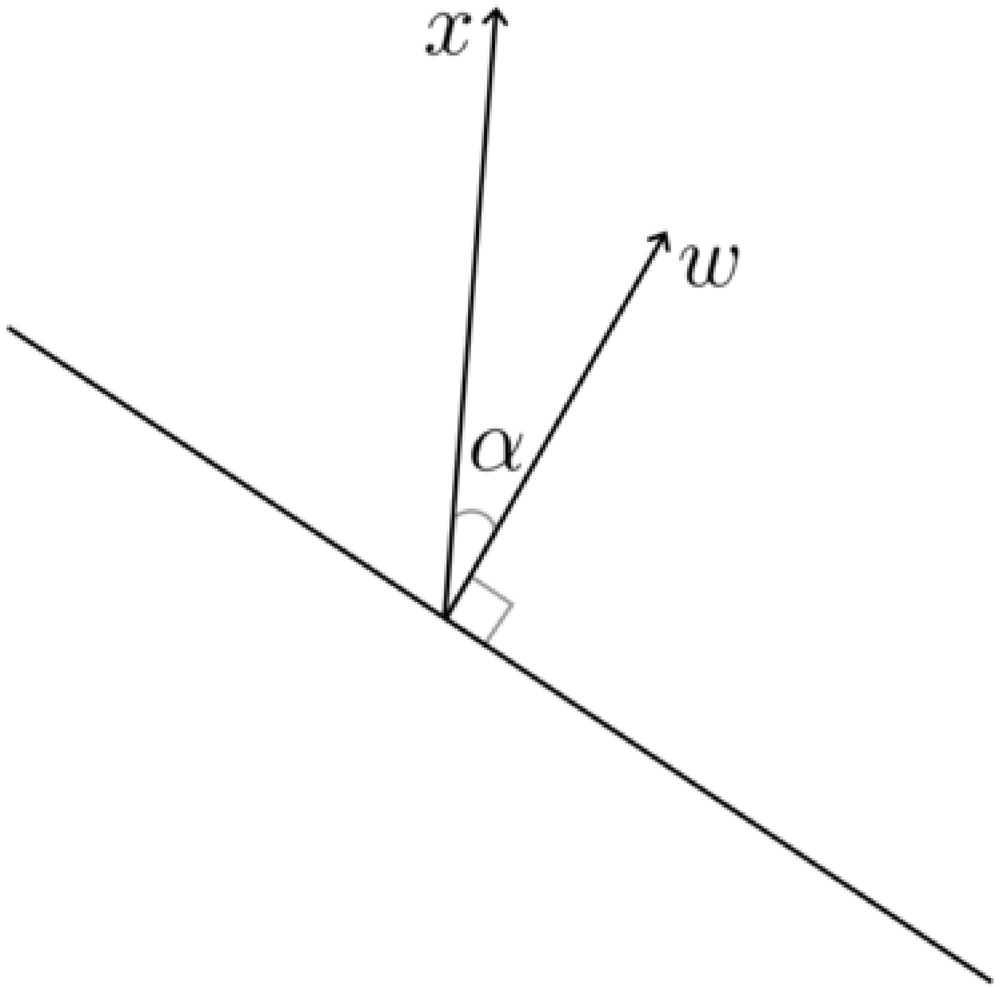
n-мерное пространство



$$x = (x_1, \ldots, x_n)$$

n-мерный вектор

Скалярное произведение на плоскости



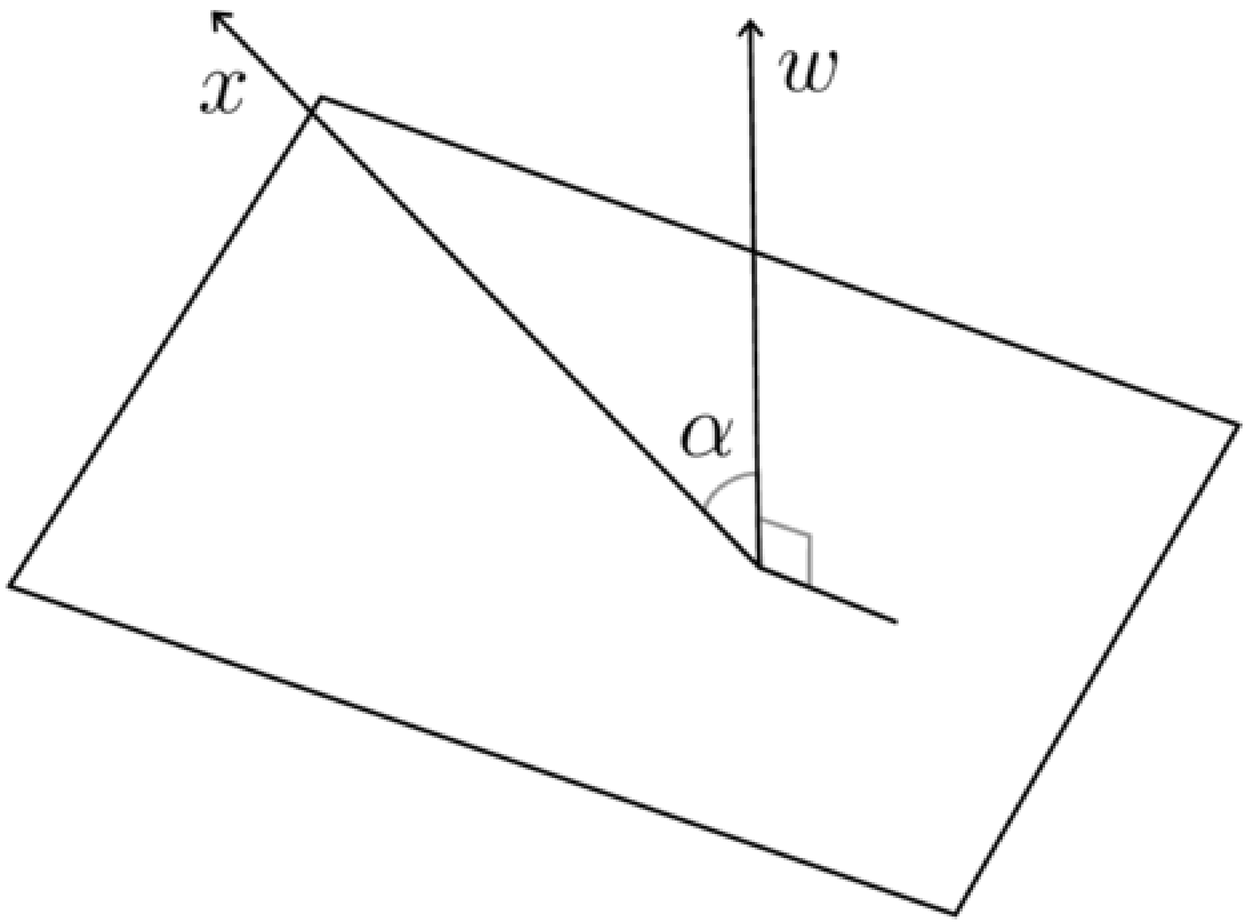
$$x = (x_1, x_2)$$

$$w = (w_1, w_2)$$

$$\langle x, w \rangle = x_1 w_1 + x_2 w_2$$

$$\langle x, w \rangle = \|x\| \cdot \|w\| \cdot \cos\alpha$$

Скалярное произведение в пространстве



$$x = (x_1, \dots, x_n)$$

$$w = (w_1, \dots, w_n)$$

$$\langle x, w \rangle = x_1 w_1 + \dots + x_n w_n$$

$$\langle x, w \rangle = \|x\| \cdot \|w\| \cdot \cos \alpha$$

Основные понятия машинного обучения

Дано:

X множество объектов (заданное своими признаками, чаще всего как точки n-мерного пространства)

Y множество ответов (чаще всего $\{0, 1\}$ или $\{-1, 1\}$, или $\{1, 2, \ldots, n\}$ или $\mathbb{R}$)

множество $\{x_1, x_2, \ldots, x_l\} \subset X$ – обучающая выборка

$y_i = y(x_i) \in Y$, $i = 1, \ldots, l$, – известные ответы

Основные понятия машинного обучения

Найти: $a : X \to Y$ – алгоритм, решающую функцию, приближающую неизвестную функцию y на всём множестве X

Типы задач машинного обучения

Задачи классификации

Классификация на два класса: $Y = \{0, 1\}$ или $Y = \{-1, 1\}$

Классификация на M непересекающихся классов: $Y = \{1, 2, \dots, M\}$

Классификация на M классов, которые могут пересекаться

Задачи регрессии

Задачи ранжирования

Типы задач машинного обучения

Задачи классификации

Классификация на два класса: $Y = \{0, 1\}$ или $Y = \{-1, 1\}$

Классификация на M непересекающихся классов: $Y = \{1, 2, \ldots, M\}$

Классификация на M классов, которые могут пересекаться

Задачи регрессии

$Y = \mathbb{R}$ или $Y = \mathbb{R}^m$

Задачи ранжирования

Типы задач машинного обучения

Задачи классификации

Классификация на два класса: $Y = \{0, 1\}$ или $Y = \{-1, 1\}$

Классификация на M непересекающихся классов: $Y = \{1, 2, \ldots, M\}$

Классификация на M классов, которые могут пересекаться

Задачи регрессии

$Y = \mathbb{R}$ или $Y = \mathbb{R}^m$

Задачи ранжирования

Y – конечное упорядоченное множество

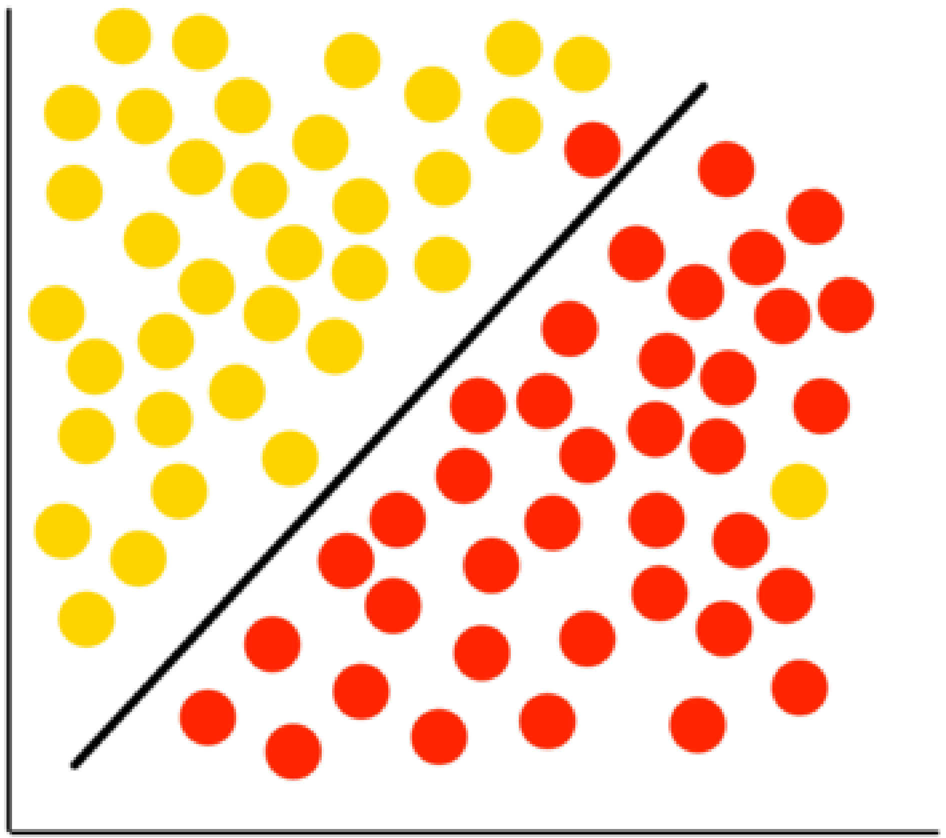
Модель машинного обучения

Модель
(предсказательная модель)
это параметрическое семейство функций

$$A = \{g(x, \theta) \mid \theta \in \Theta\}$$

Где: $g : X \times \Theta \to Y$ – фиксированная функция
Θ – множество допустимых значений параметра θ

Пример: линейная модель



Пусть вектор параметров – это $\theta = (\theta_1, \ldots, \theta_n)$, $\Theta = \mathbb{R}^n$.

Тогда для задач регрессии и ранжирования можно положить

$$g(x, \theta) = \sum_{j=1}^{n} \theta_j f_j(x), \quad Y = \mathbb{R}$$

Для задачи классификации обычно полагают

$$g(x, \theta) = \operatorname{sgn} \sum_{j=1}^{n} \theta_j f_j(x), \quad Y = \{-1, 1\}$$

Этапы обучения

Метод обучения

это отображение вида $\mu : (X \times Y)^l \to A$, которое произвольной выборке $\{(x_i, y_i) \mid i = 1, \ldots, l\}$ ставит в соответствие алгоритм $a \in A$

В задачах обучения всегда есть два этапа:

1. Этап обучения

метод μ по выборке $\overline{x} \in X^l$
строит алгоритм обучения $a = \mu(\overline{x})$

2. Этап применения

алгоритм a для новых объектов $x \in X$
выдаёт ответы $a(x)$

Функция потерь

Функция потерь $\mathcal{L}(a,x)$ – это величина ошибки алгоритма $a \in A$ на объекте $x \in X$

Функция потерь для задач классификации:

$\mathcal{L}(a,x) = [a(x) \neq y(x)]$ – индикатор ошибки

Функции потерь для задач регрессии:

$\mathcal{L}(a,x) = |a(x) - y(x)|$ – абсолютное значение ошибки

или

$\mathcal{L}(a,x) = (a(x) - y(x))^2$ – квадратичная ошибка

Минимизация эмпирического риска

Функционал качества – это эмпирический риск алгоритма a на X^l:

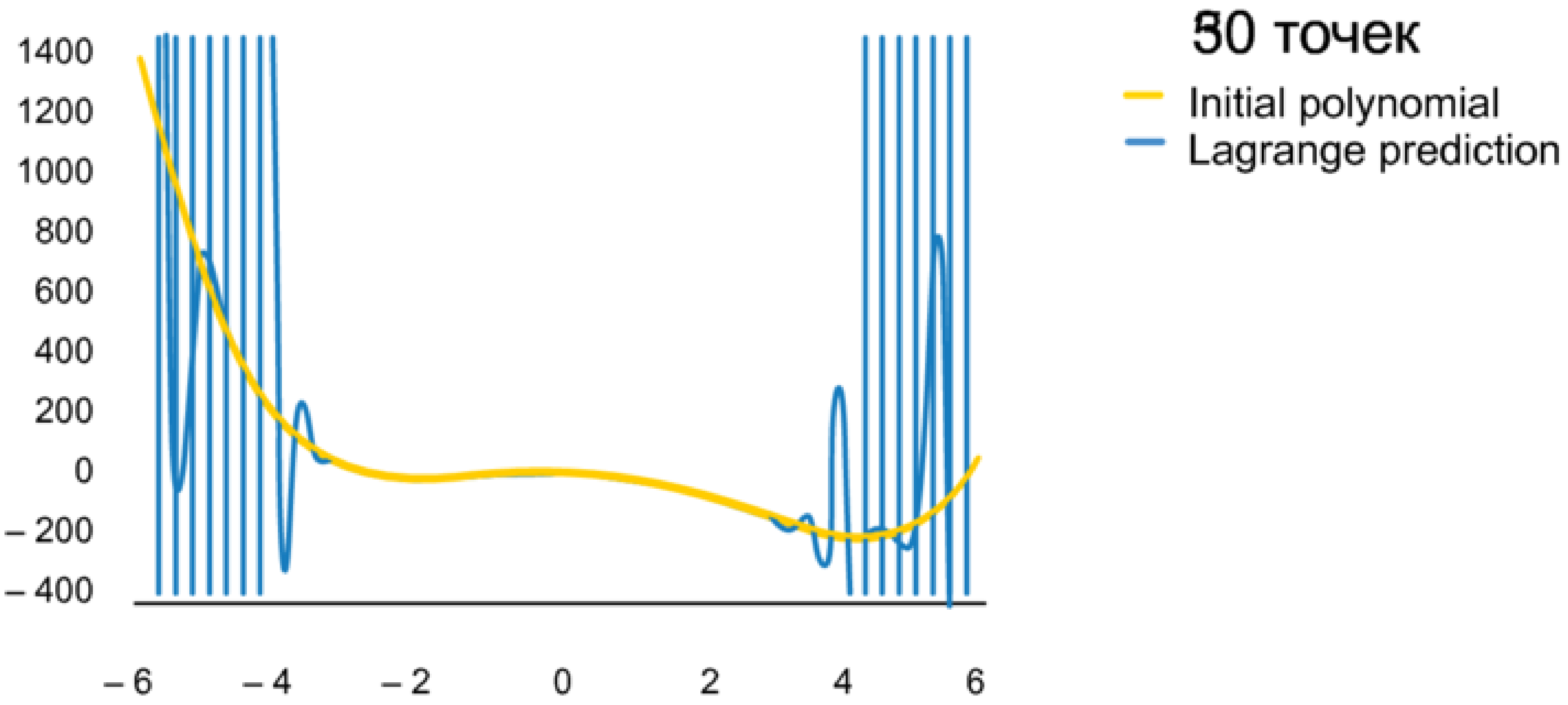
$$Q(a, X^l) = \frac{1}{l}\sum_{j=1}^{l}\mathcal{L}(a, x_j)$$

Вот каким образом мы сводим задачу обучения к обычной задаче оптимизации:

$$\mu(X^l) = \arg\min_{a\in A} Q(a, X^l)$$

Проблема переобучения

Многочлен $x^4 - 3x^3 - 17x^2 + 4$ на отрезке $[-6; 6]$



Проверка качества обучения

Как проверить, сильно ли вы переобучились?

разделить выборку на две части, обучаться
на одной части, проверять на другой:

$$Q(\mu(X^l), X^k) \to \min$$

Скользящий контроль (leave-one-out)

$$\frac{1}{L}\sum_{i=1}^{L}\mathcal{L}(\mu(X^L \setminus \{x_i\}), x_i) \to \min$$

Кросс-проверка

$$X^L = X_n^l \cup X_n^k, \quad n = 1, \ldots, N$$

$$\frac{1}{N}\sum_{n=1}^{N} Q(\mu(X_n^l), X_n^k) \to \min$$

Цикл решения задачи

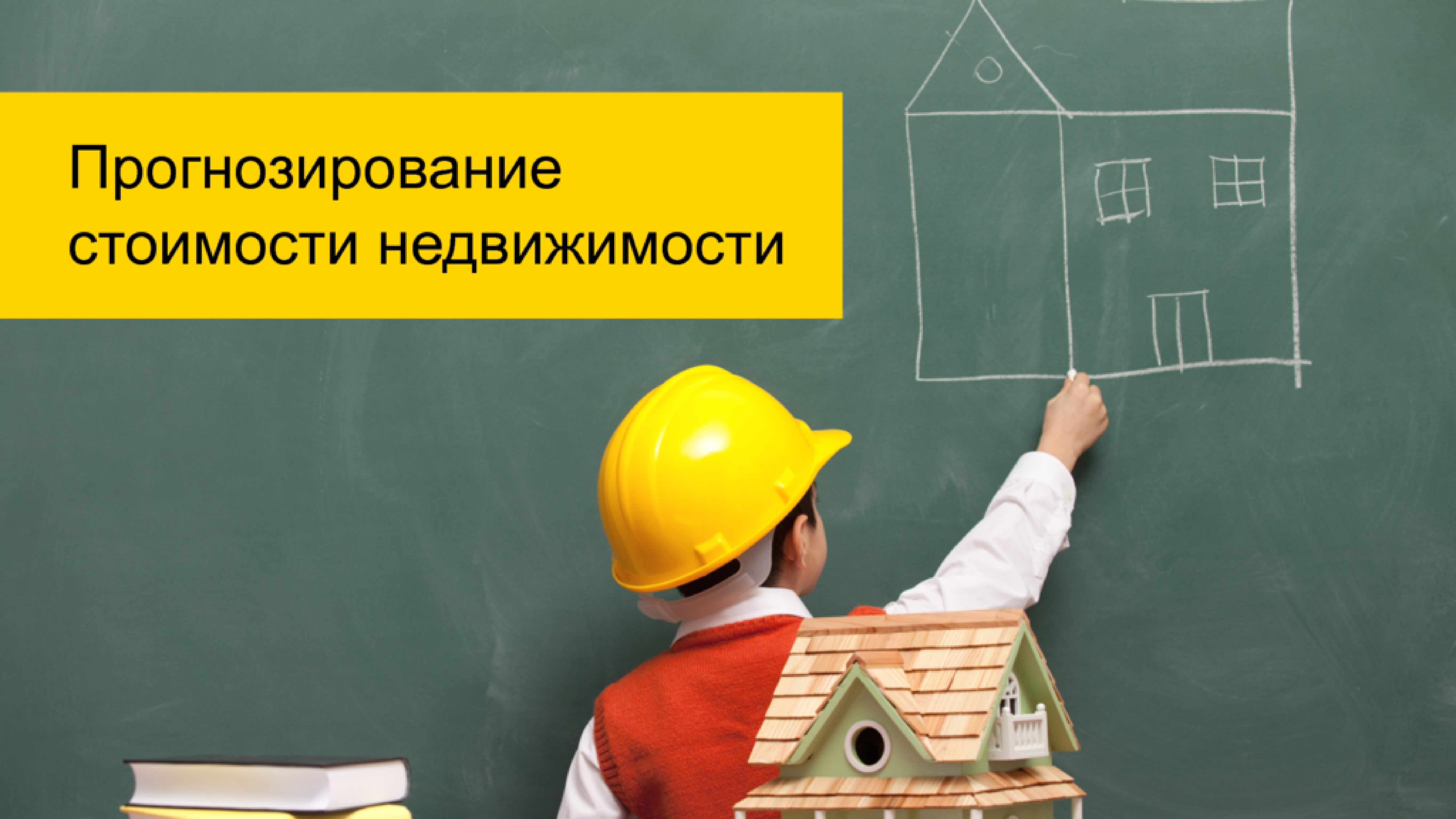
Как применяется Машинное обучение в жизни

Медицина

Банковская сфера

Категоризация текстовых документов

Прогнозирование стоимости недвижимости



Ранжирование поисковой выдачи

Поиск изображений

Рекомендательная система

Метрические методы

Гипотезы компактности и непрерывности

Гипотеза компактности (для задач классификации):

Близкие объекты, как правило, лежат в одном классе

Гипотеза непрерывности (для задач регрессии):

Близким объектам соответствуют близкие ответы

Формализация понятия близости

Задана функция расстояния $\rho : X \times X \to [0, \infty)$

Например, любая метрика:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$ – евклидова

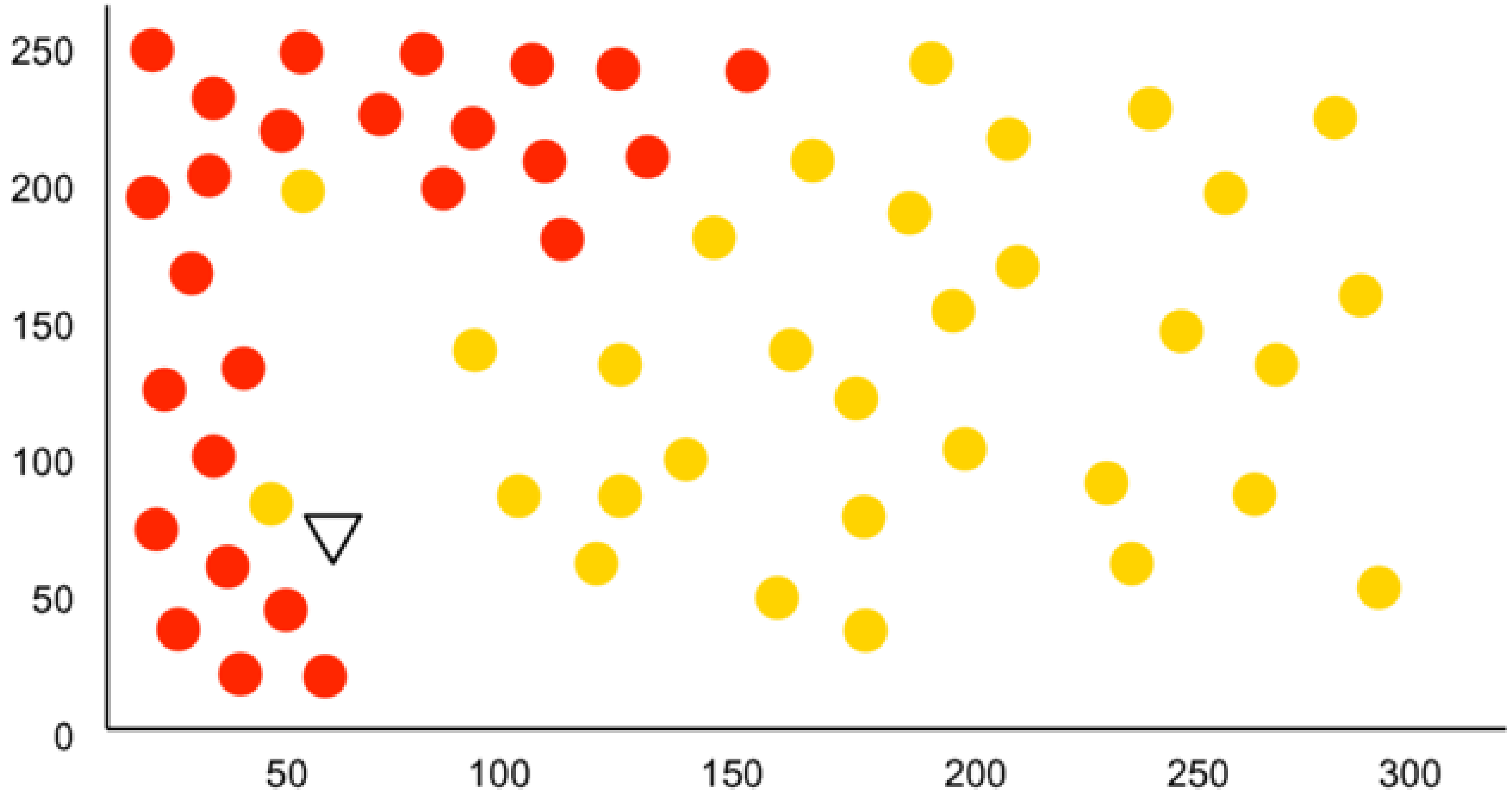
$$\rho(x, y) = \sum_{i=1}^{n} |x_i - y_i| \text{ — метрика } L_1$$

$$\rho(x, y) = \max_{i=1,\dots,n} |x_i - y_i| \text{ — метрика } L_\infty$$

Ещё можно вводить веса у признаков

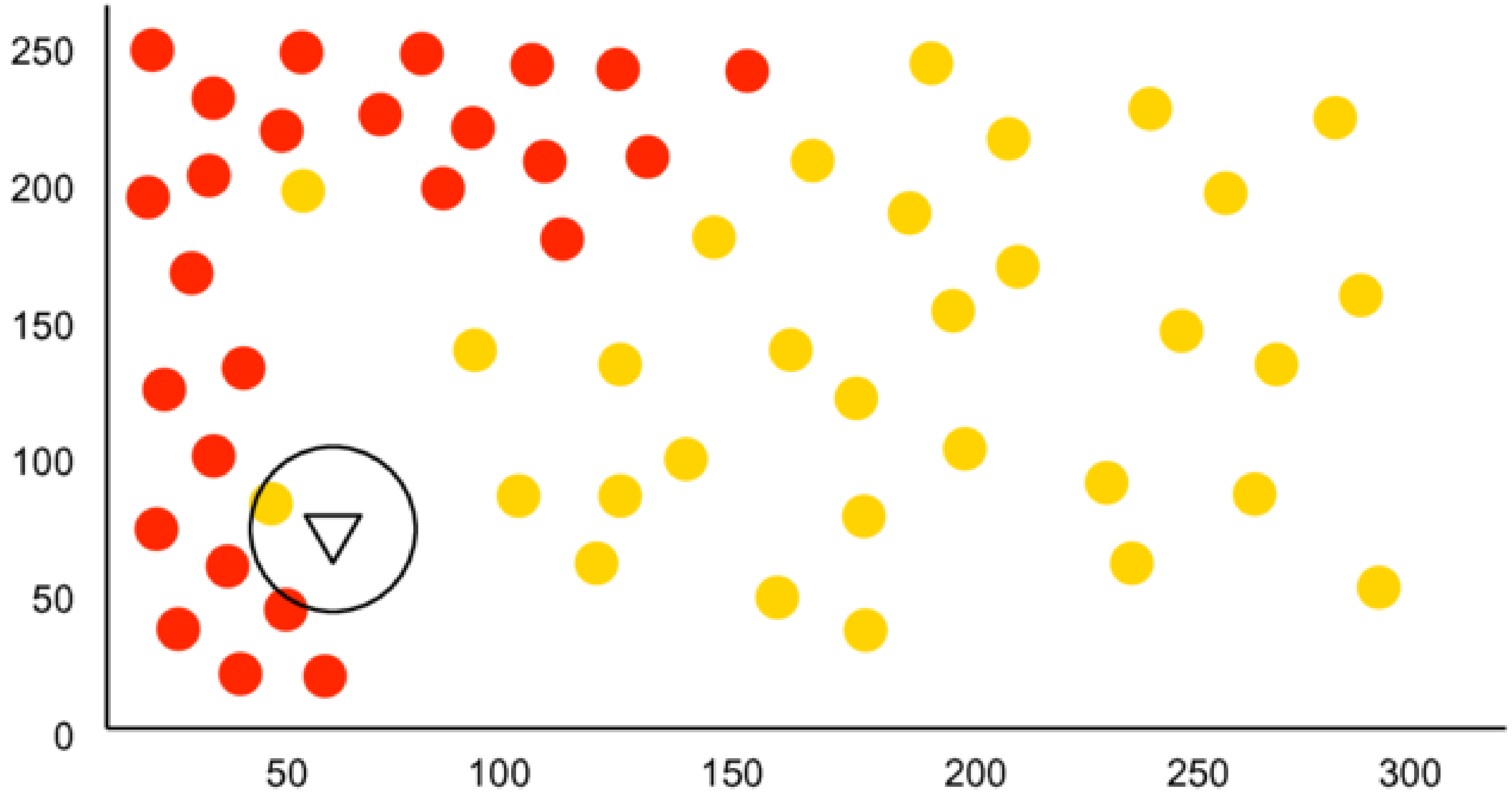
Метод одного ближайшего соседа

Красный или жёлтый новый объект?

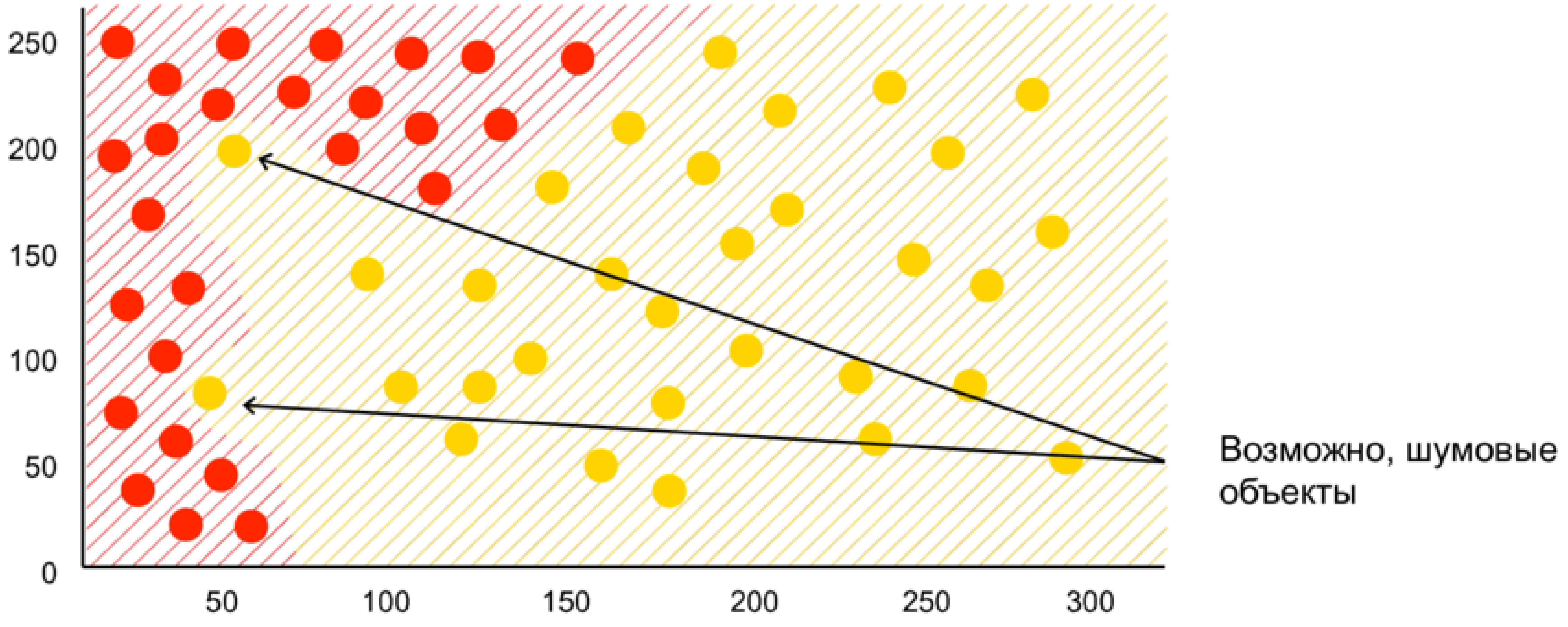


Метод одного ближайшего соседа

Пусть новый объект принадлежит к тому же классу, что и его ближайший сосед

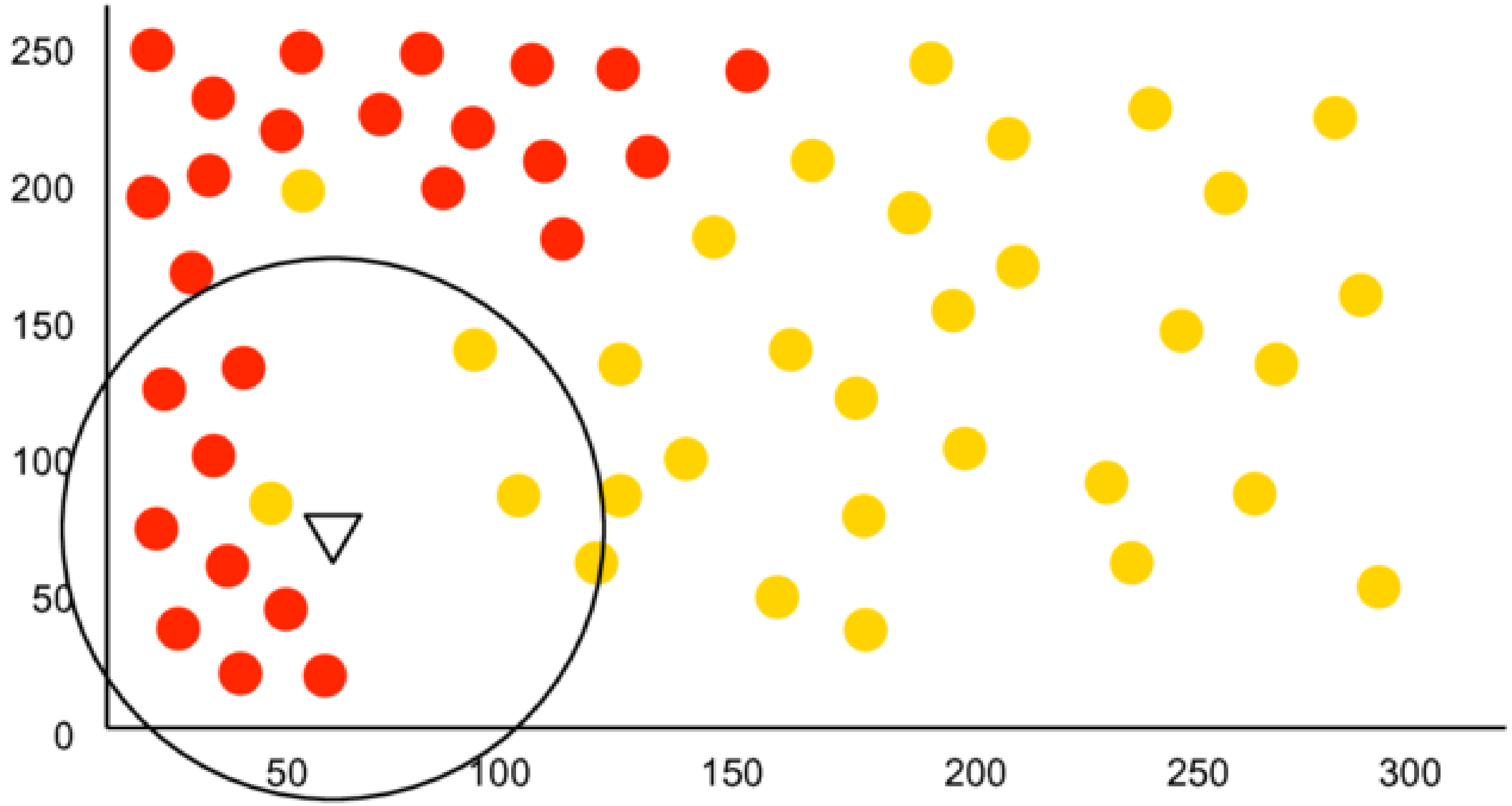


Граница разделения классов



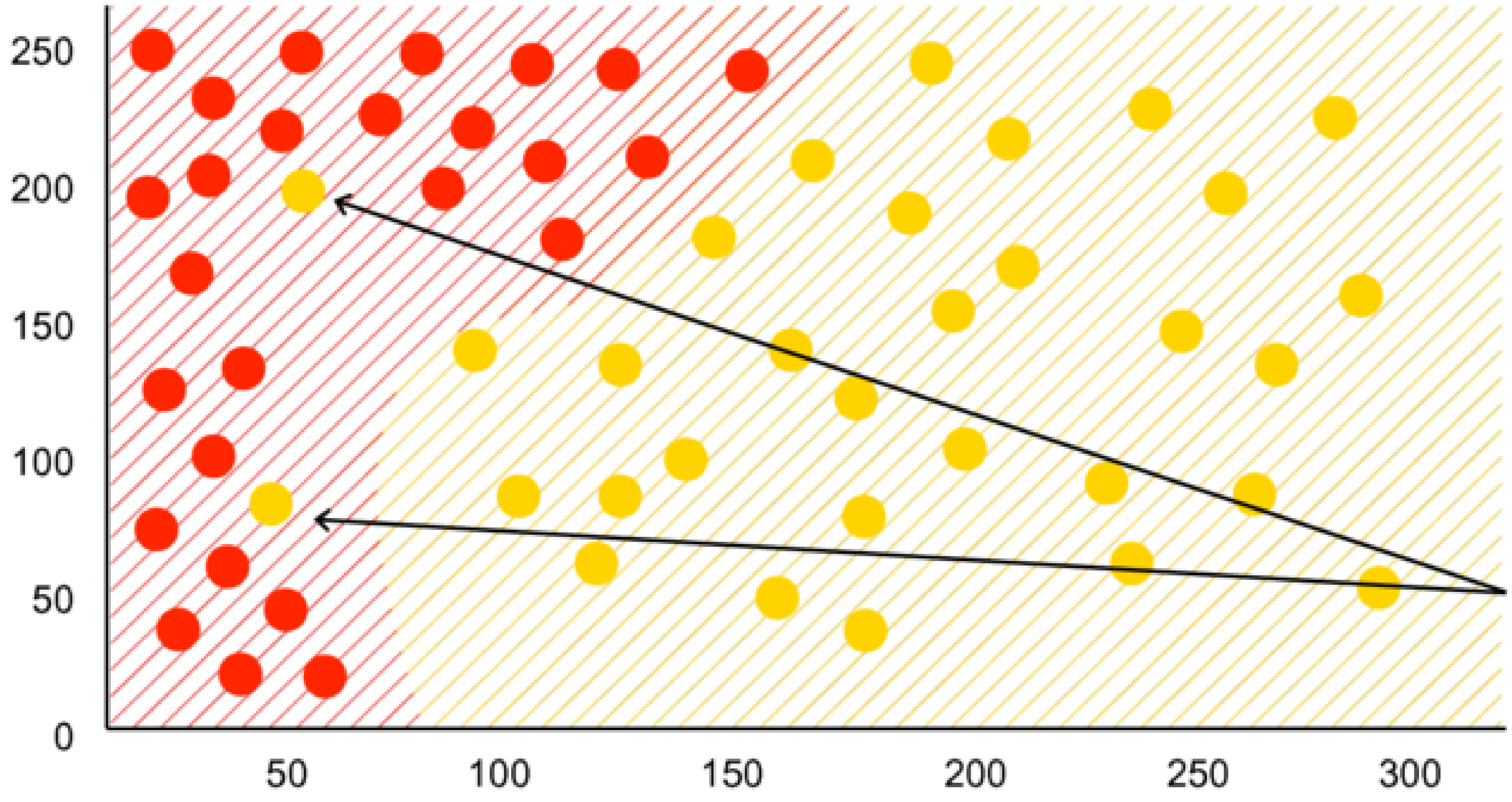
Метод k ближайших соседей

Новый объект принадлежит тому же классу,
что и большинство из k его соседей

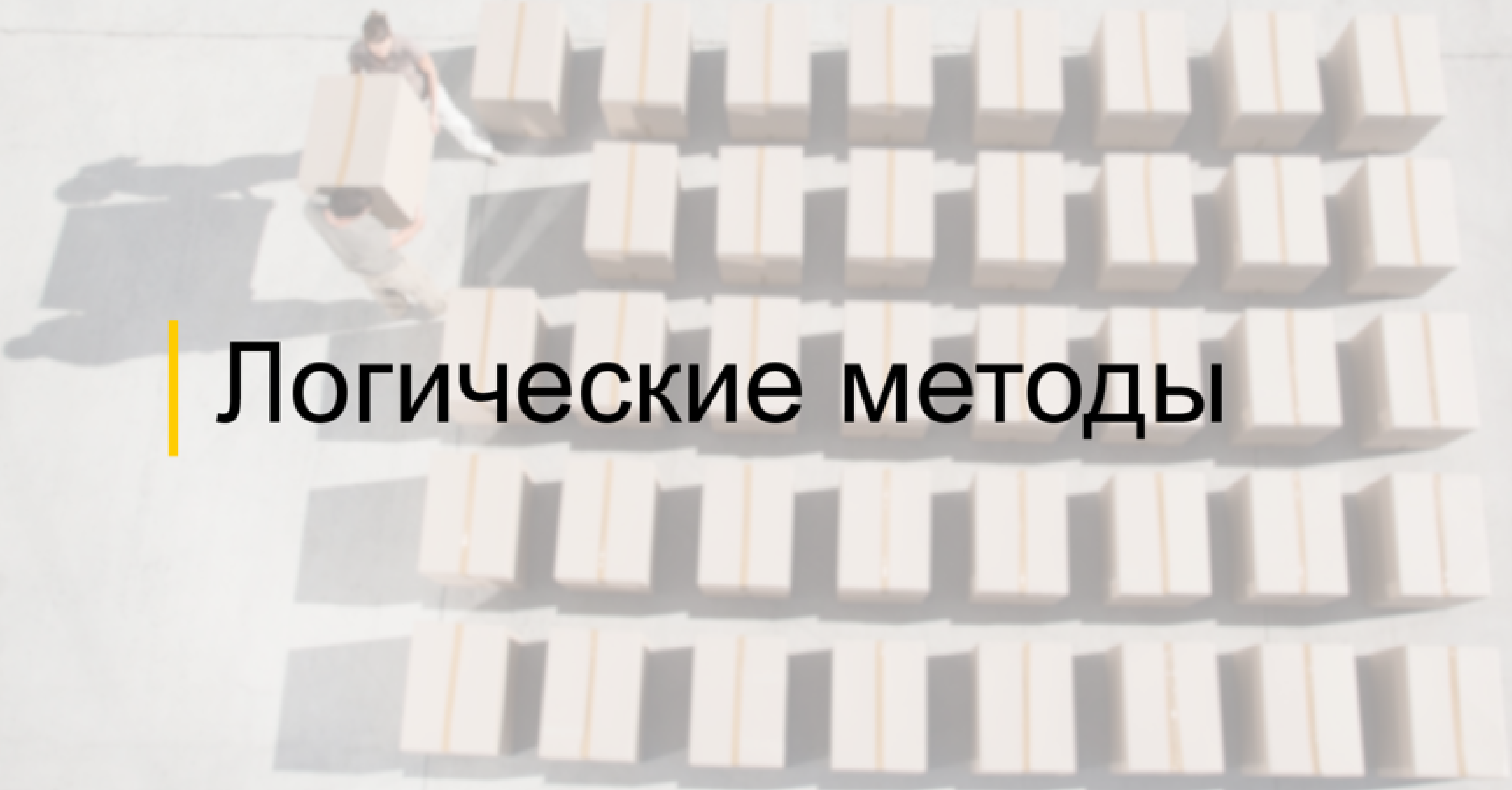


Метод k ближайших соседей

Граница разделения классов k=5



Оказывается, алгоритм дает ошибку на обучающей выборке! А это и не плохо



Логические методы

Логические закономерности

Логическая закономерность – это предикат (функция) $R : X \to \{0, 1\}$ удовлетворяющий двум требованиям:

1. Интерпретируемость:

   R записывается на естественном языке

   R зависит от небольшого числа признаков $(1\text{–}7)$

2. Информативность относительно одного из классов: $c \in Y$

$$\#\{x_i \mid R(x_i) = 1 \text{ и } y_i = c\} \to \max$$

$$\#\{x_i \mid R(x_i) = 1 \text{ и } y_i \neq c\} \to \min$$

Часто используемые виды закономерностей

Пороговое условие (решающий пень):

$$R(x) = [f_j(x) \leqslant a_j] \text{ или } [a_j \leqslant f_j(x) \leqslant b_j]$$

Конъюнкция пороговых условий:

$$R(x) = \bigwedge_{j \in J} [a_j \leqslant f_j(x) \leqslant b_j]$$

Синдром – выполнение не менее d условий из J:

$$R(x) = \left[\sum_{j \in J} [a_j \leqslant f_j(x) \leqslant b_j] \geqslant d \right]$$

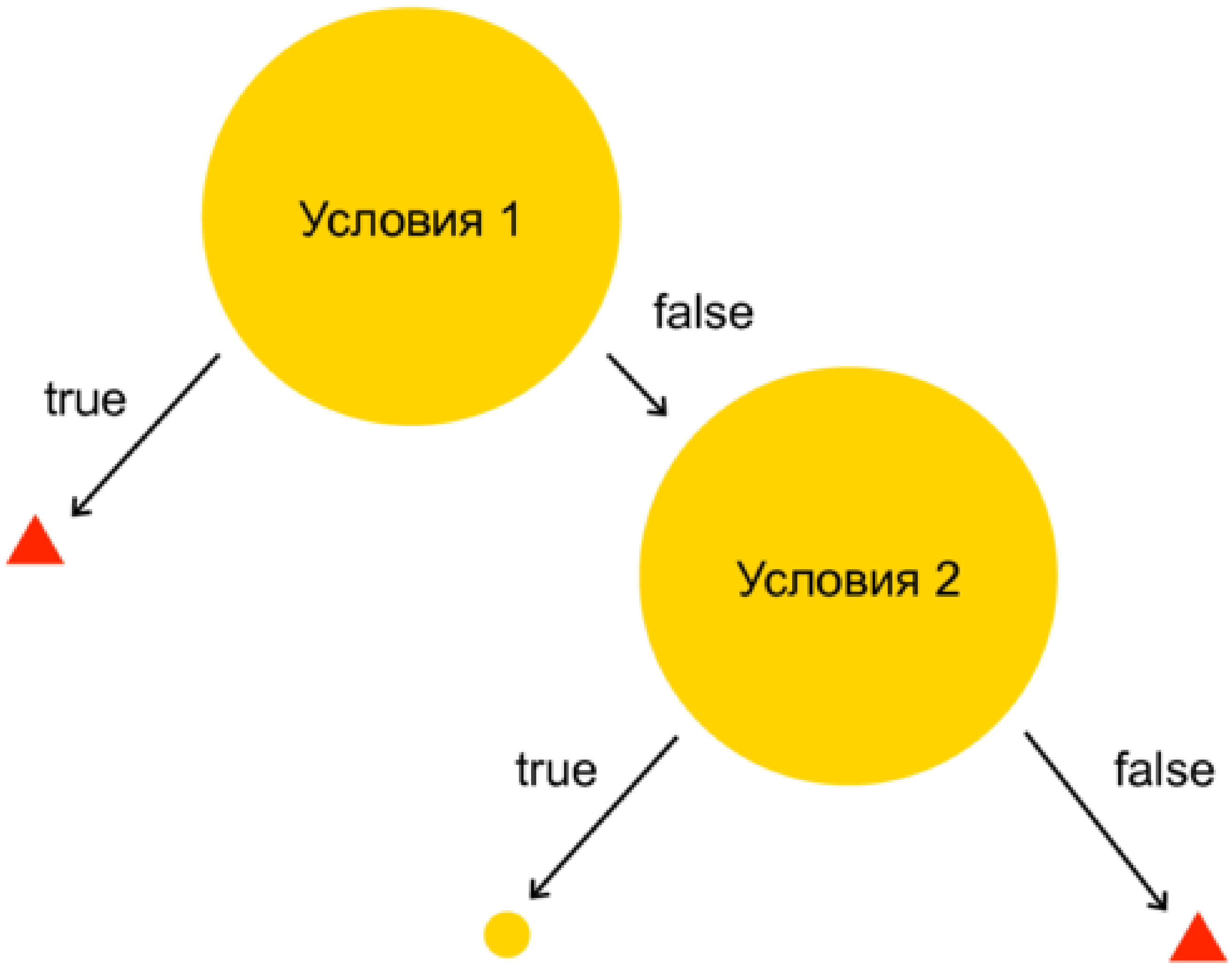
Полуплоскость – линейная пороговая функция:

$$R(x) = \left[\sum_{j \in J} w_j f_j(x) \geqslant w_0 \right]$$

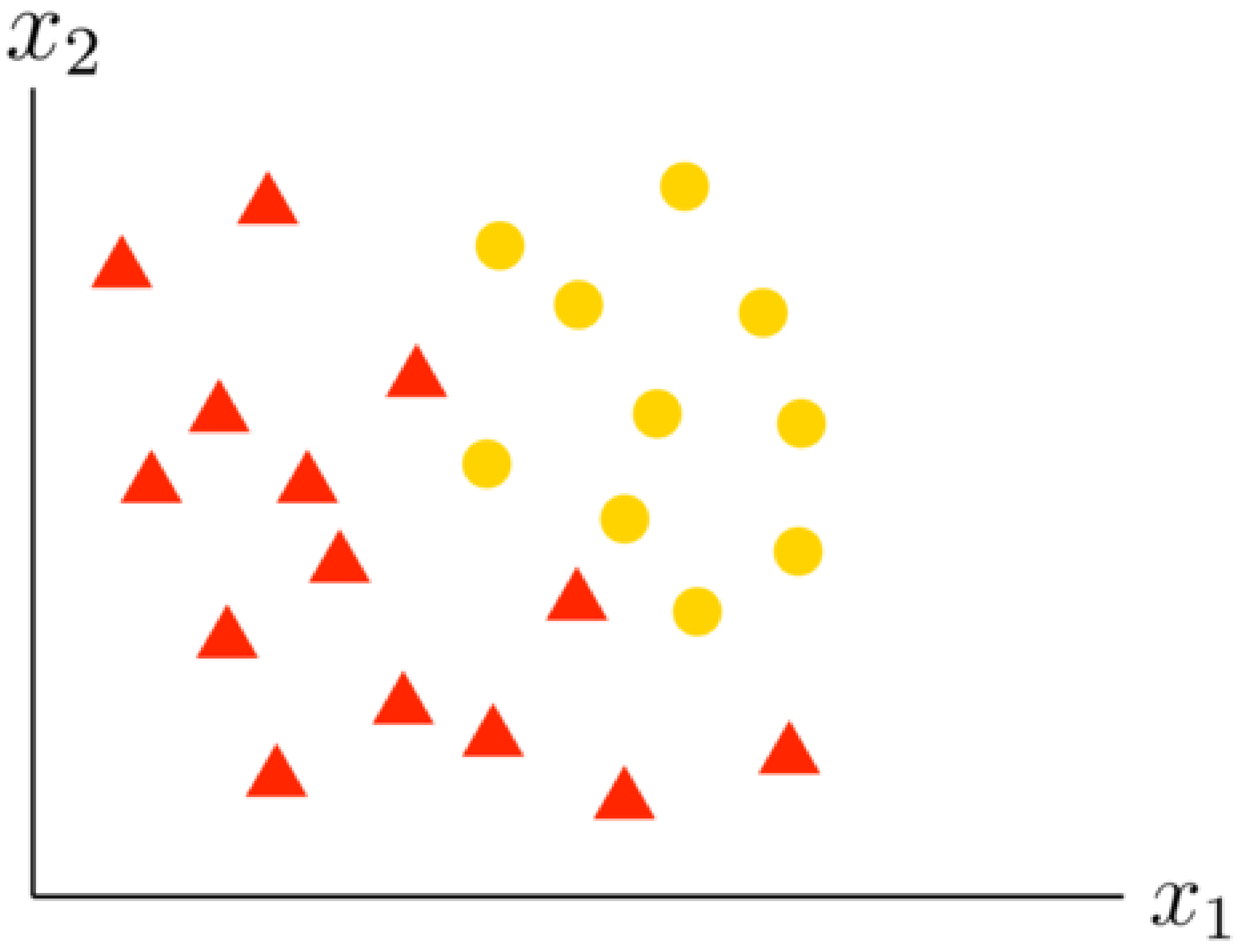
Шар – пороговая функция близости:

$$R(x) = [\rho(x, x_0) \leqslant w_0]$$

Решающие деревья



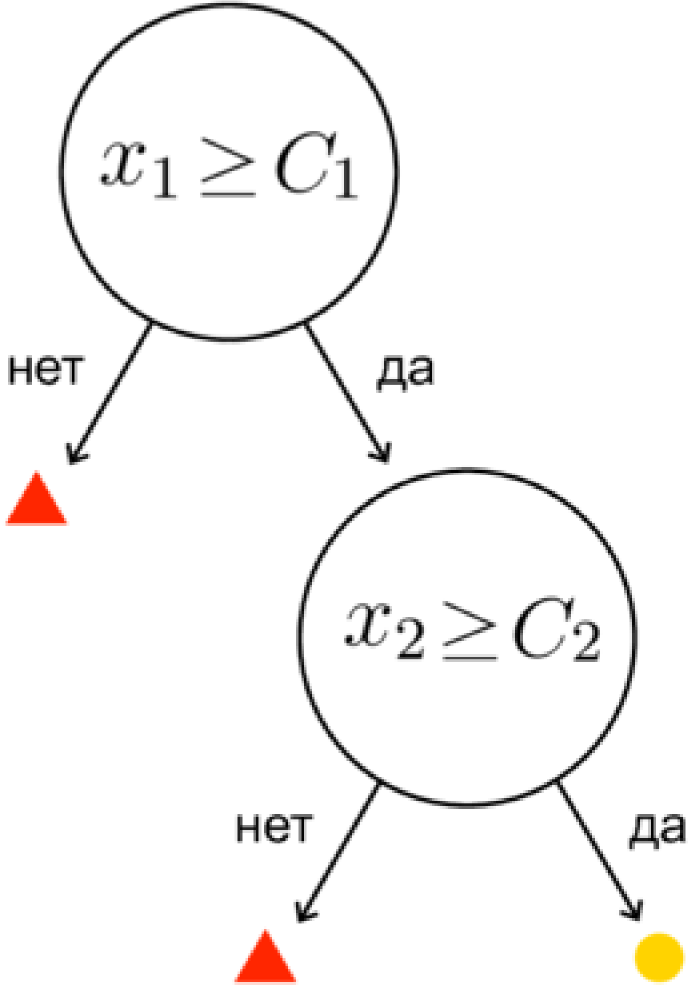
Решающие деревья

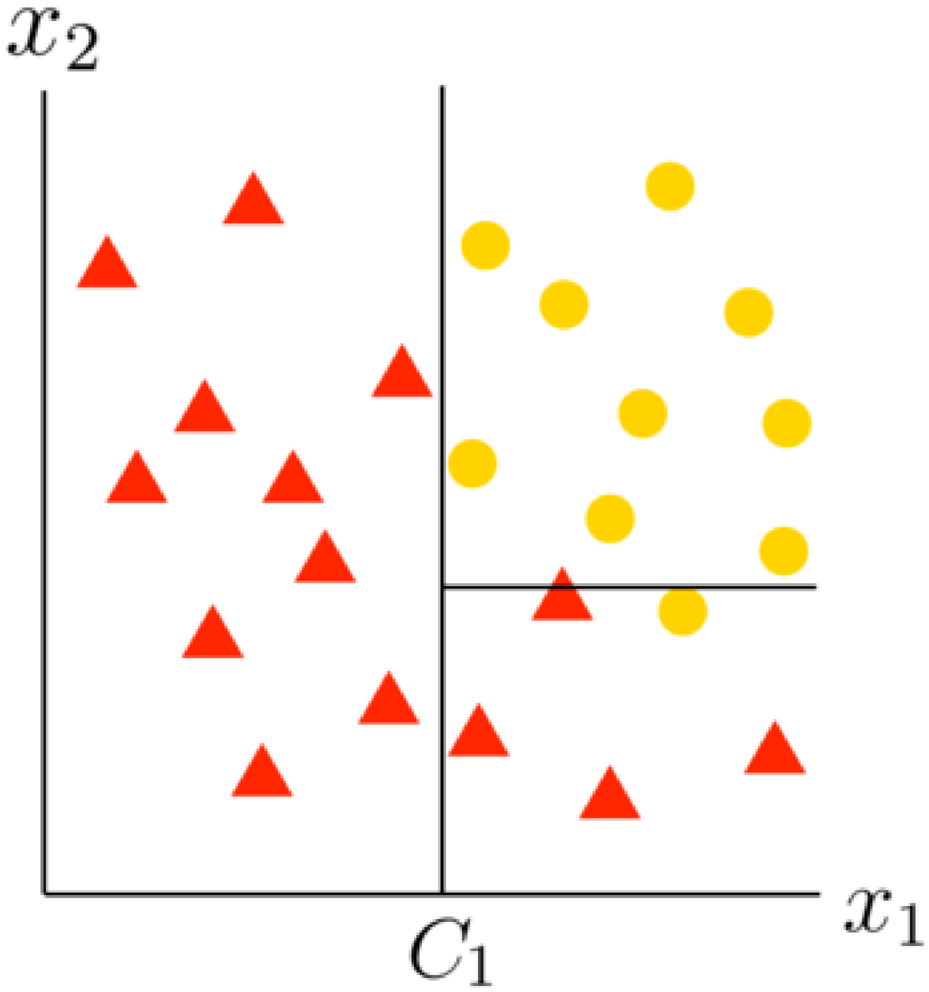


Какие условия будут в дереве?

Попробуем использовать пороговые условия перехода в виде пороговых правил: $x > C$

Решающие деревья



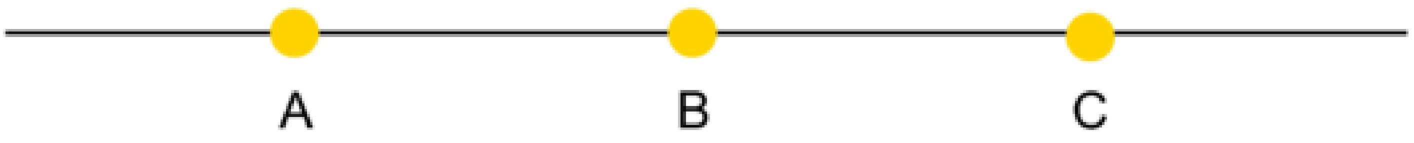


Каждый раз берём наиболее «информативное» разделение текущей области

Преимущества логических алгоритмов перед метрическими

1. Придумать правильную меру сходства – значит почти решить задачу, это сложно. А решающие деревья не используют метрики

2. Единственное, что используют деревья – точка B ближе к A, чем C по данному признаку



3. Устойчивы к монотонным преобразованиям признаков

Метод опорных векторов (SVM)

Немного истории

Вапник Владимир
Наумович

Червоненкис Алексей
Яковлевич

Линейные методы классификации

Дано:

Обучающая выборка $X^l = \{(x_i, y_i) \mid i = 1, \ldots, l\}$

x_i – объекты, векторы из множества $X = \mathbb{R}^n$

y_i – ответы из множества $Y = \{-1; +1\}$

Найти:

Параметры $w \in \mathbb{R}^n$ линейной модели классификации

$a(x; w) = \operatorname{sign}(\langle x, w \rangle)$

Понятие отступа

Отступ положителен

объект правильно расклассифицирован

Отступ отрицателен

объект неправильно расклассифицирован

$$M_i(w) = (\langle x_i, w \rangle) y_i$$

– отступ объекта x_i

Задача SVM

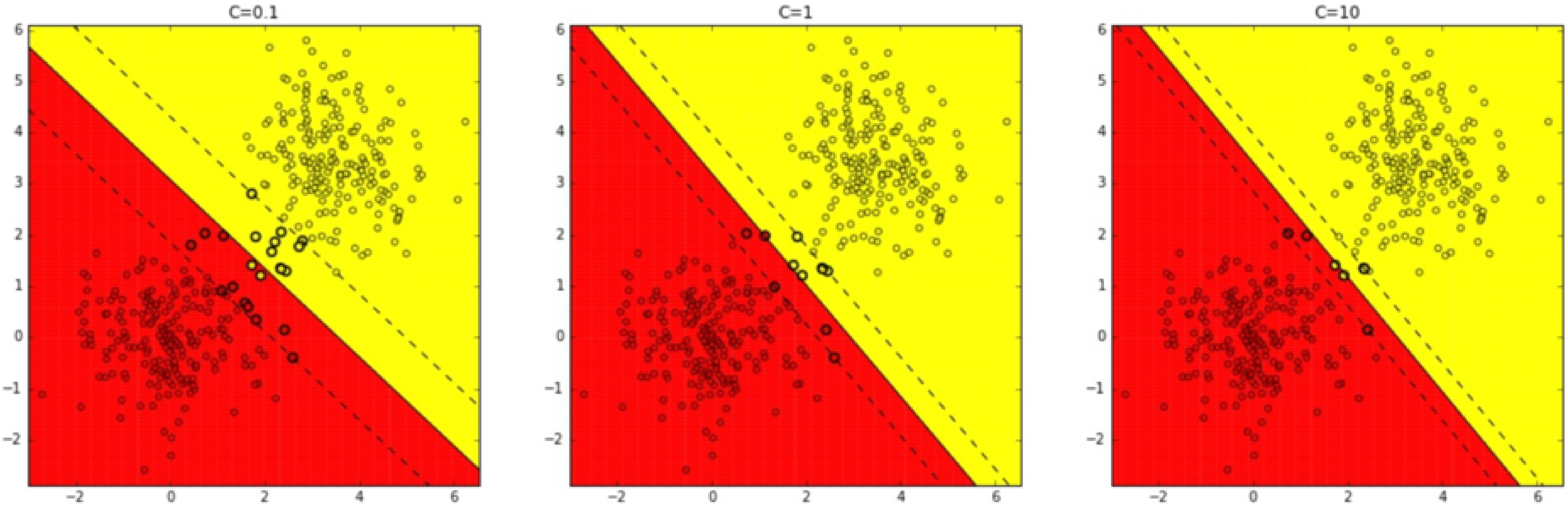
Хотим минимизировать общую потерю:

$$\sum_{i=1}^{l} [a(x_i, w) \neq y_i] = \sum_{i=1}^{l} [M_i(w) < 0] \to \min_w$$

Заменим общую потерю на чуть большую функцию и будем минимизировать её:

$$\sum_{i=1}^{l} [M_i(w) < 0] \leqslant \sum_{i=1}^{l} [1 - M_i(w)]_+ + \frac{1}{2C} \|w\|^2 \to \min_w$$

Применение SVM с разными константами



$$\sum_{i=1}^{l} [1 - M_i(w)]_+ + \frac{1}{2C} \|w\|^2 \to \min_{w}$$

Двойственная задача

Задача SVM эквивалентна такой задаче:

$$\begin{cases} -\sum\limits_{i=1}^{l} \lambda_i + \frac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l} \lambda_i\lambda_j y_i y_j \boxed{\langle x_i, x_j\rangle} \to \min\limits_{\lambda}, \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, l, \\ \sum\limits_{i=1}^{l} \lambda_i y_i = 0. \end{cases}$$

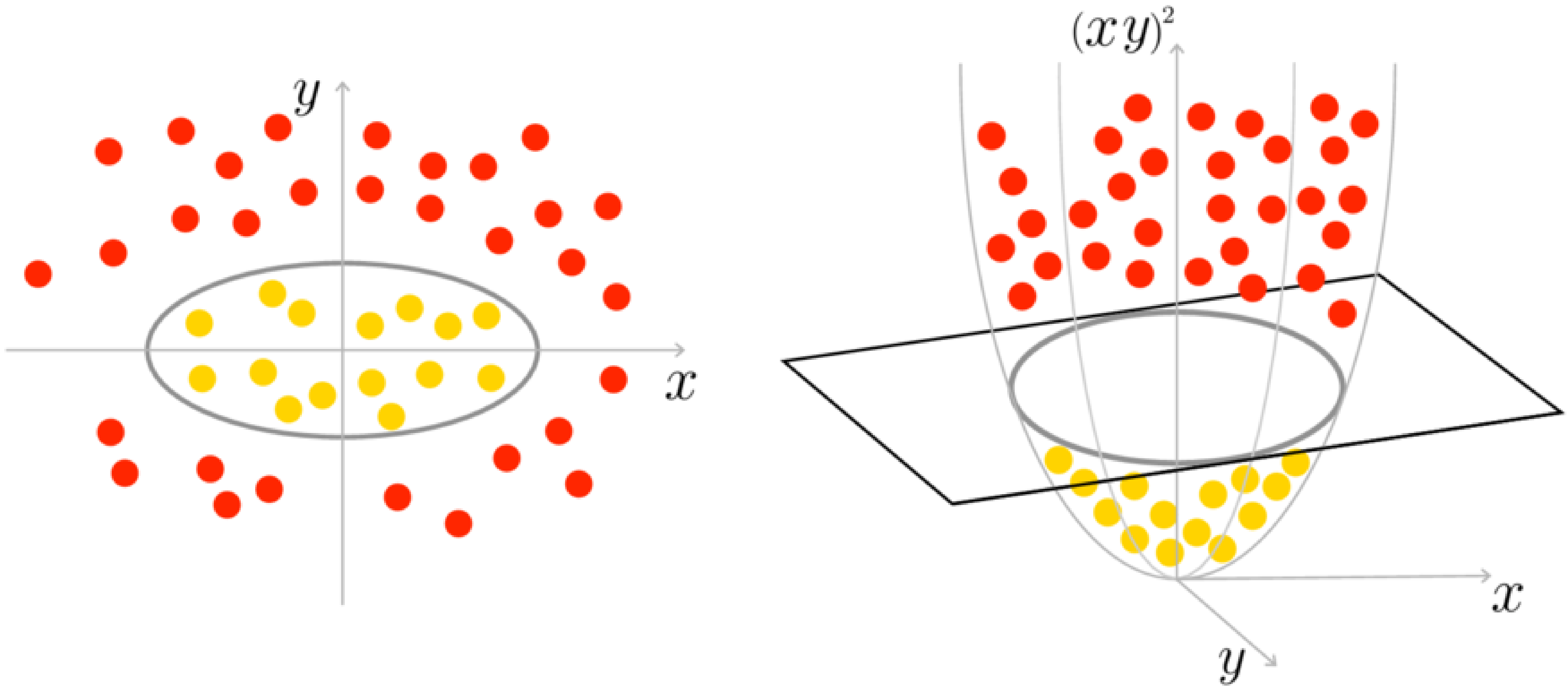
Нелинейный SVM

В двойственной задаче элементы x_i участвуют только в скалярных произведениях.

И постановка задачи, и сам классификатор зависят только от способности вычислять скалярные произведения объектов.

Новая идея: заменить скалярное произведение функцией от двух переменных, если её можно считать скалярным произведением в некотором (пусть другом) пространстве.

Нелинейный SVM: пример



Понятие ядра

Функция $K : (X, X) \to \mathbb{R}$ называется *ядром*,
если $\forall x, x' \in X \quad K(x, x') = \langle \psi(x), \psi(x') \rangle$

для некоторой функции $\psi : X \to \mathcal{H}$,
где $\mathcal{H}$ – пространство со скалярным произведением

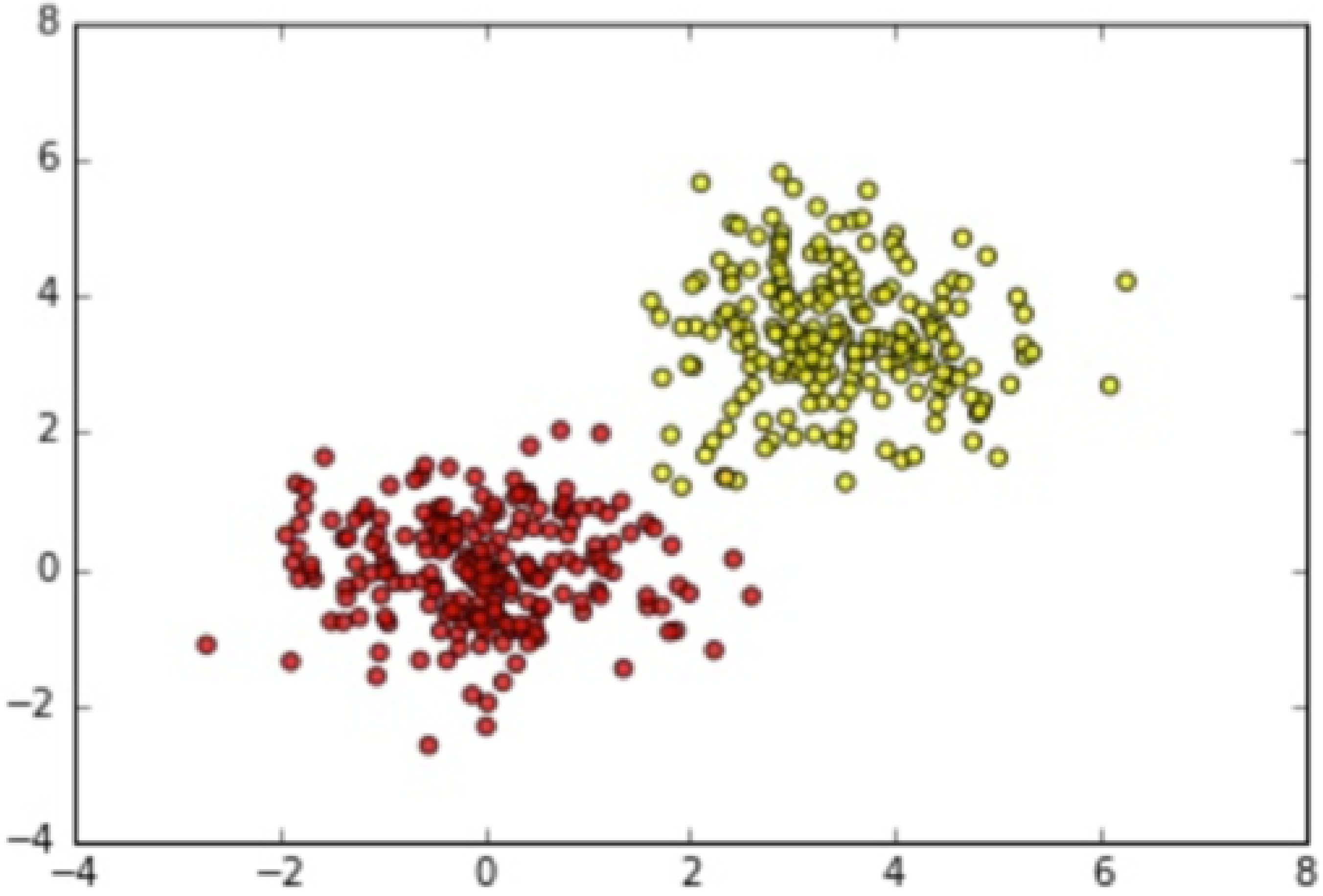
$$K(x, x') = \langle x, x' \rangle^2$$

– квадратичное ядро

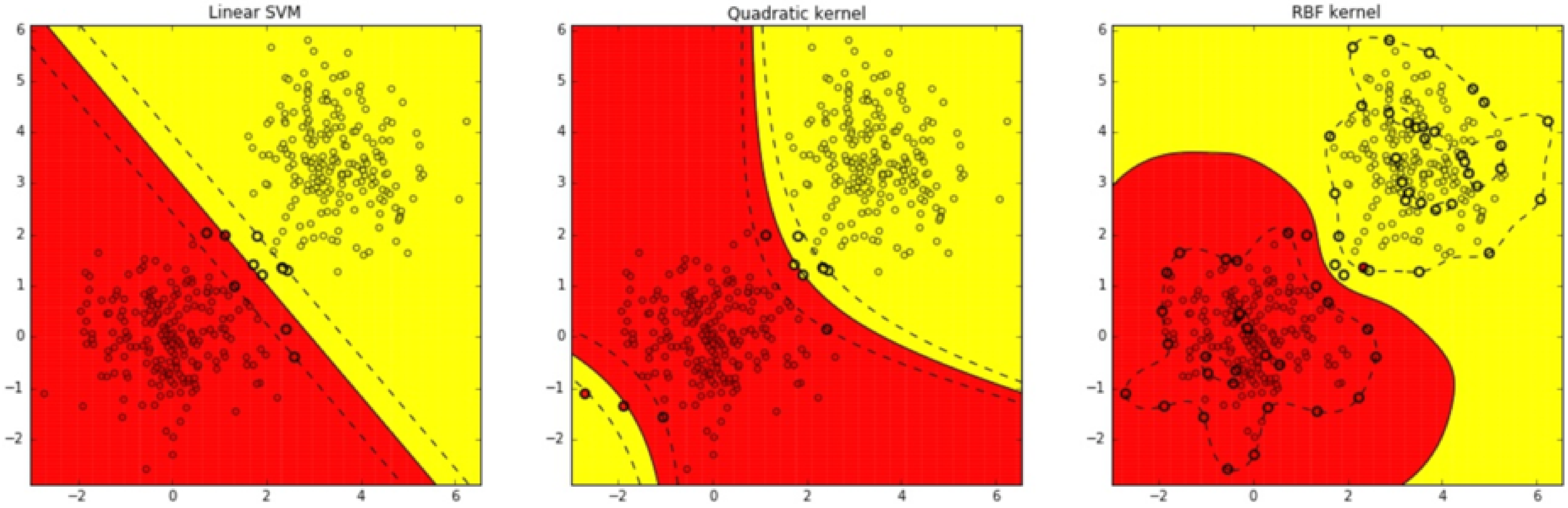
$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

– RBF-ядро

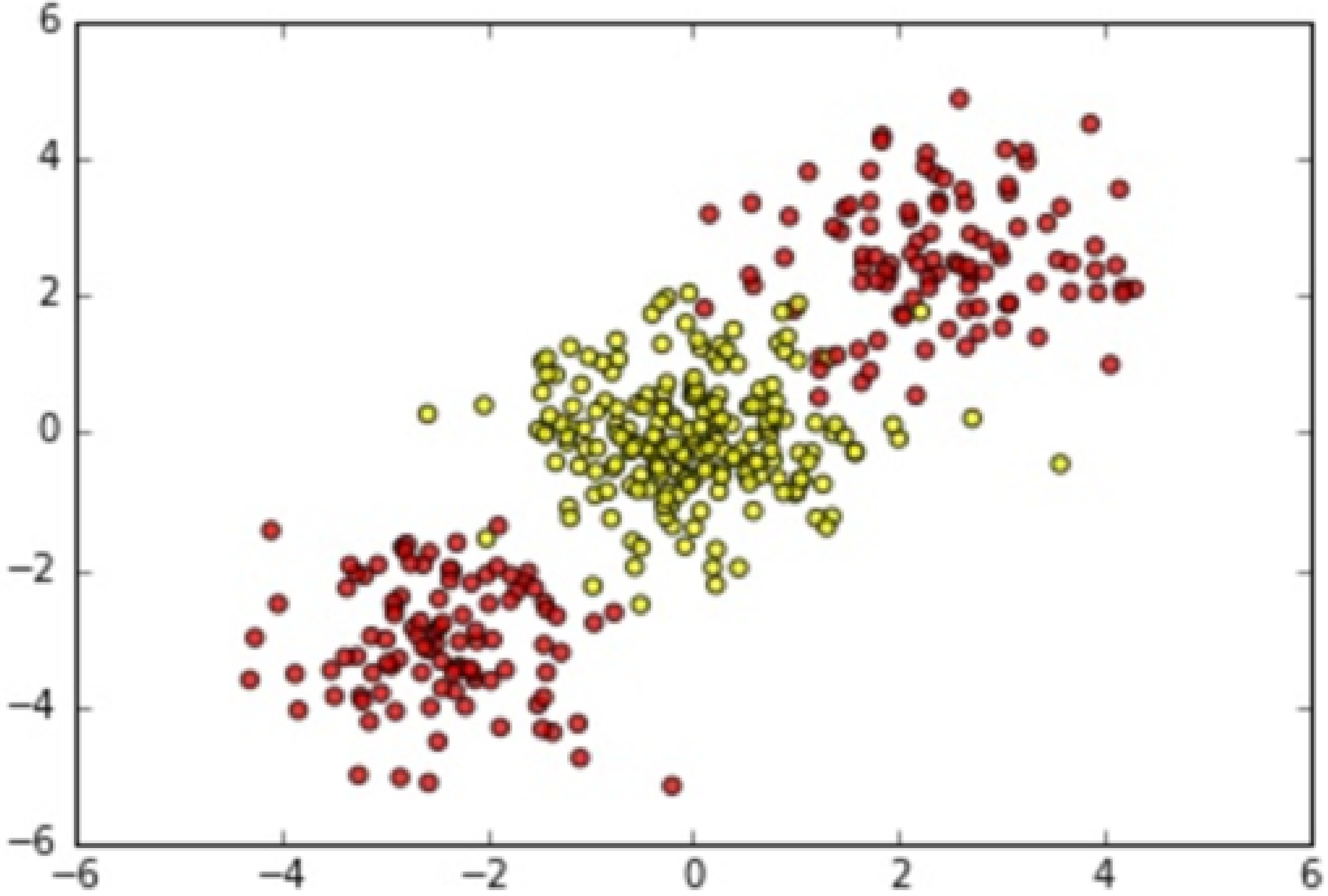
Линейно разделимая выборка



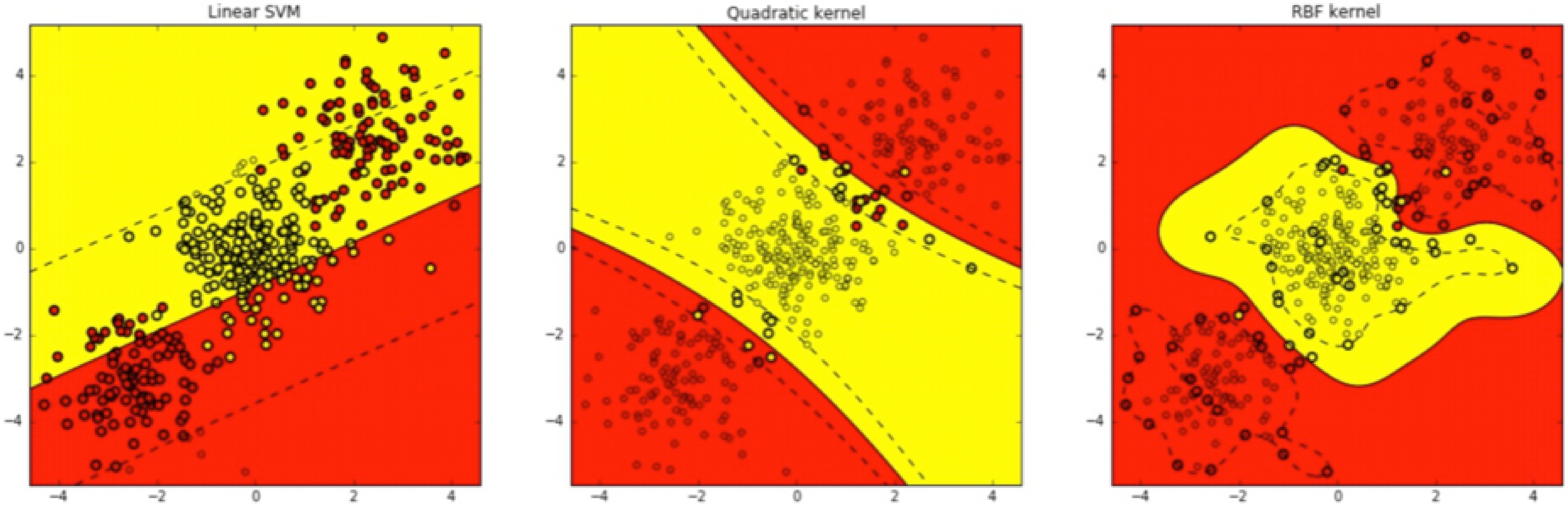
Применение SVM с разными ядрами



Линейно не разделимая выборка

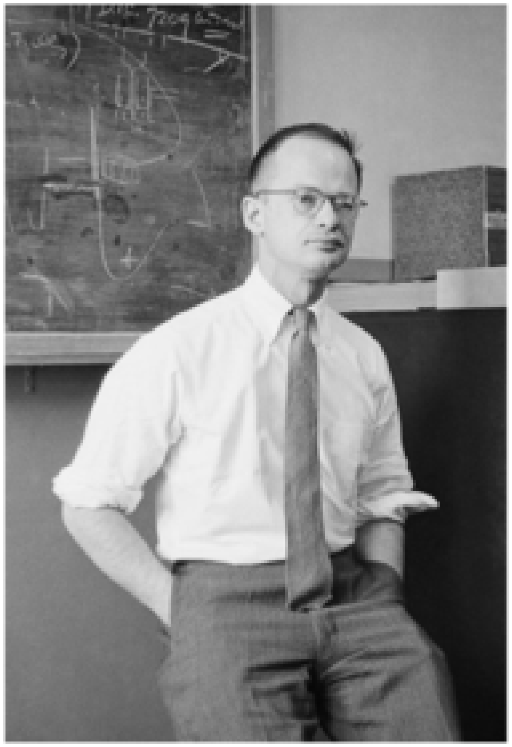


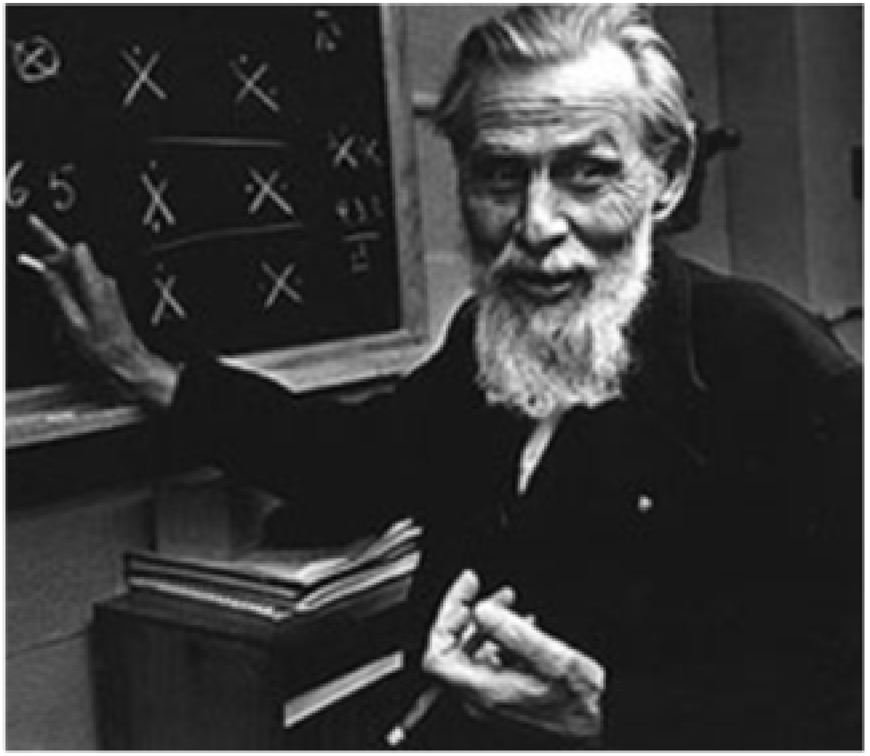
Применение SVM с разными ядрами



Нейронные сети

И снова немного истории

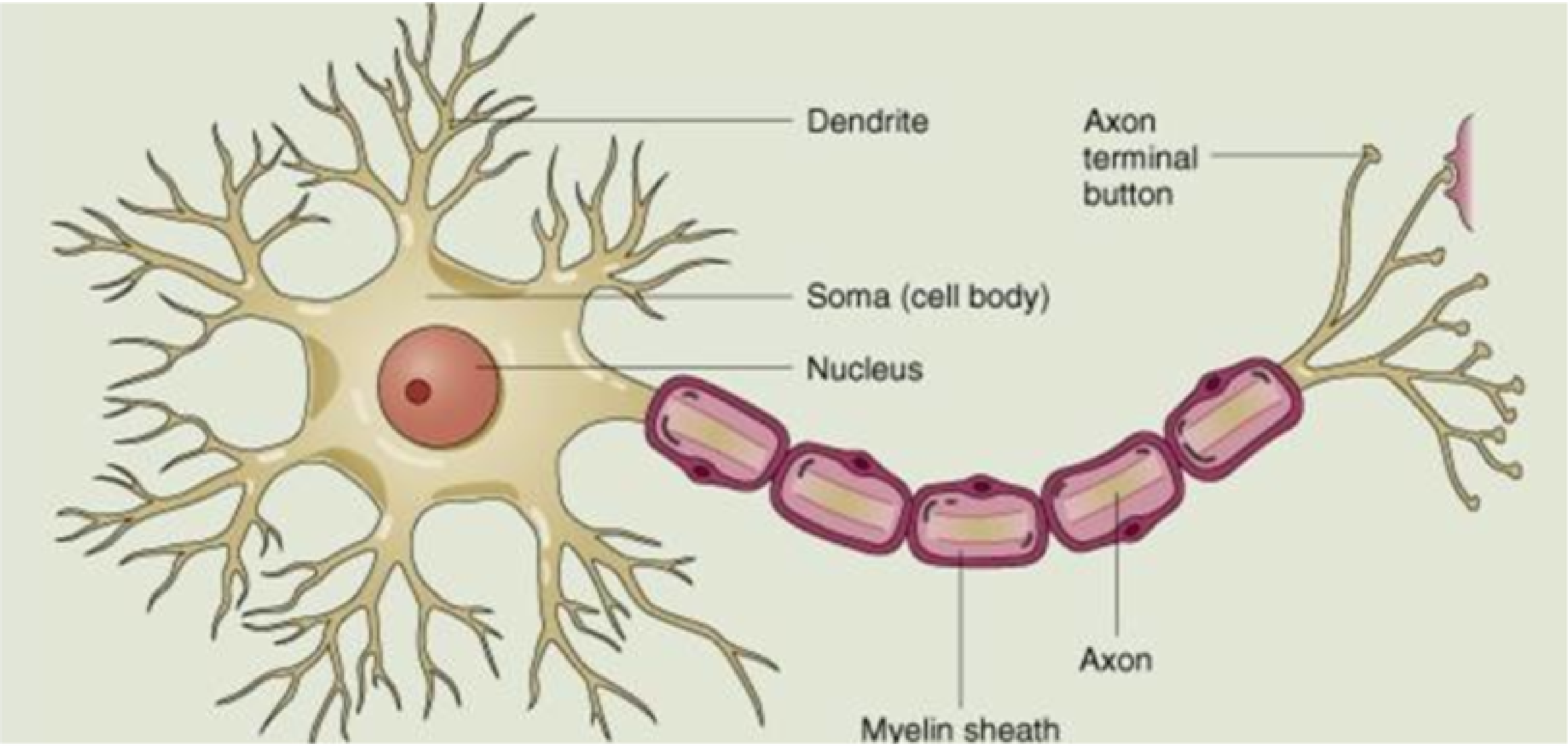




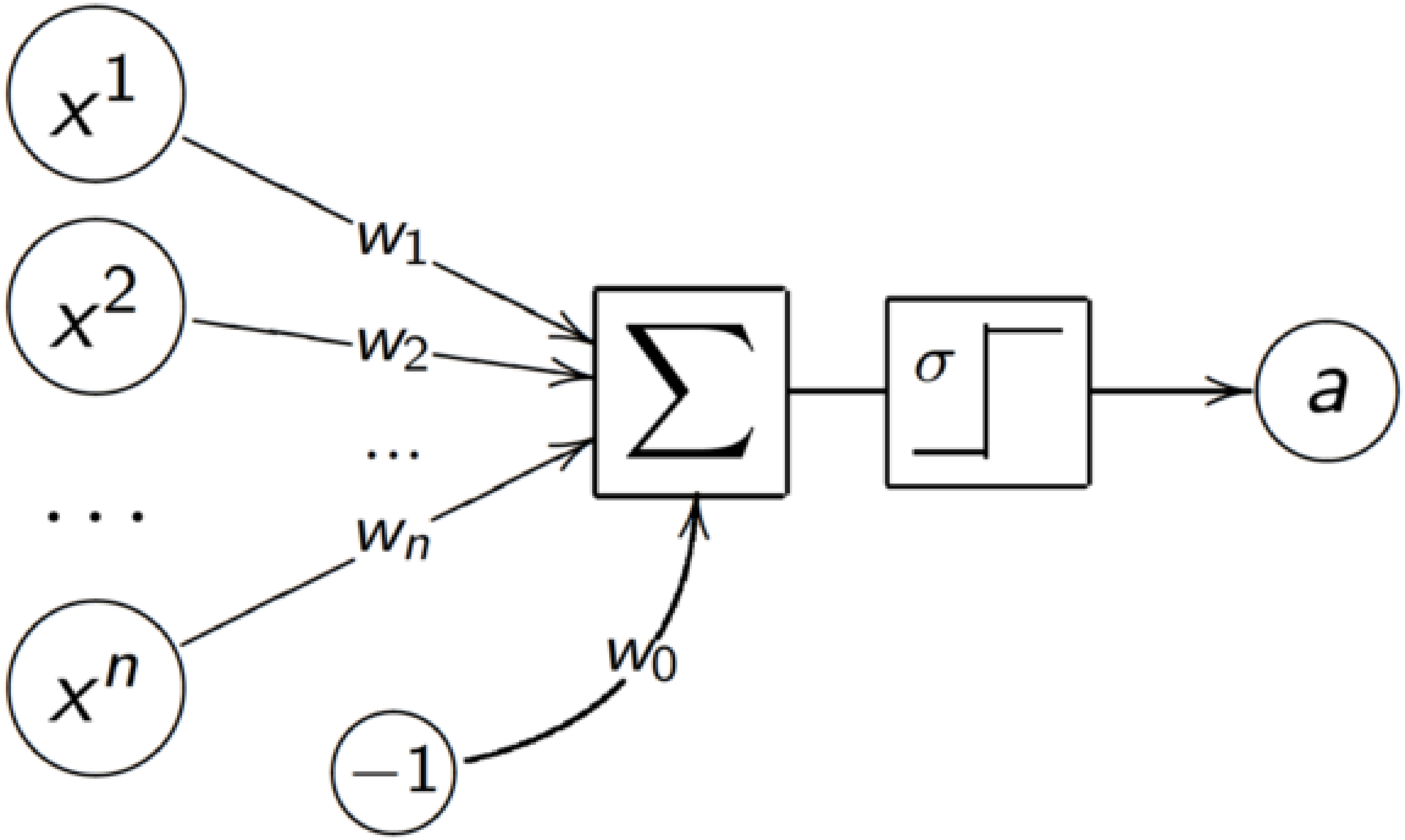
Уоррен Маккалок, Уолтер Питтс
Первый формальный нейрон – 1943

Фрэнк Розенблатт
Первая искусственная
нейронная сеть – 1957

Нейрон


Dendrite
Axon terminal button
Soma (cell body)
Nucleus
Axon
Myelin sheath

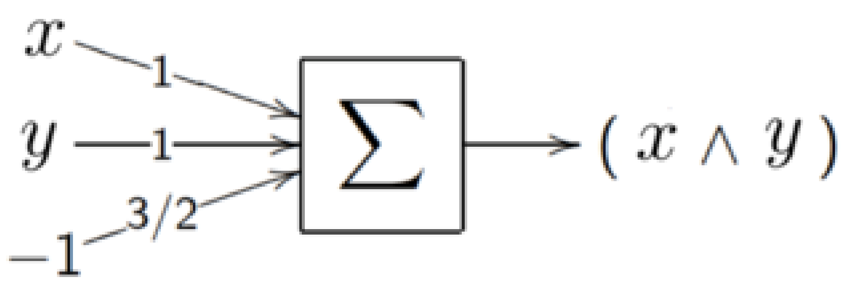
Линейная модель нейрона Маккалока-Питтса

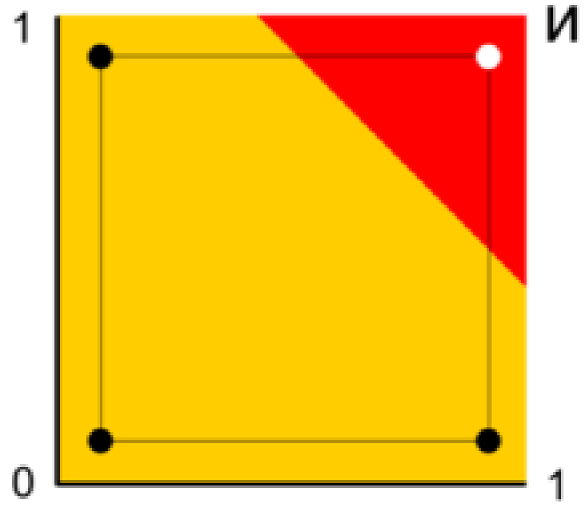


Нейронная реализация логических функций

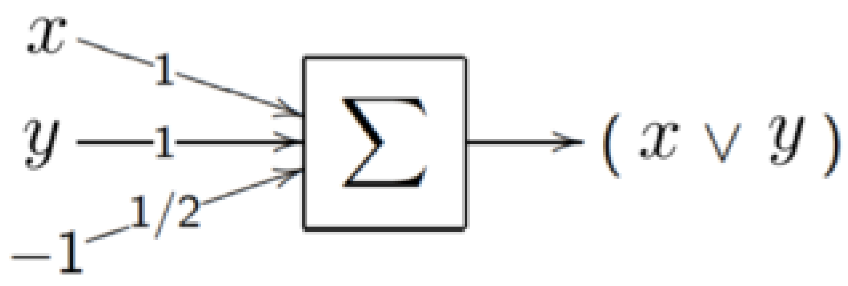
Функции И, ИЛИ от бинарных переменных x и y

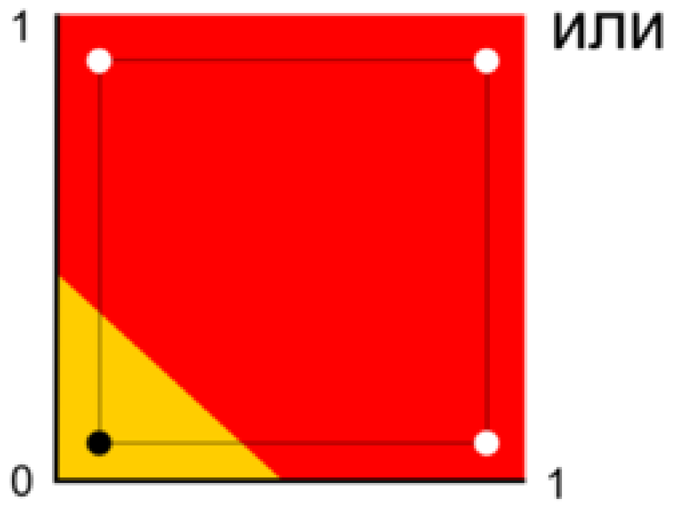
$$x \wedge y = \left[x + y - \frac{3}{2} > 0 \right]$$





$$x \vee y = \left[x + y - \frac{1}{2} > 0 \right]$$

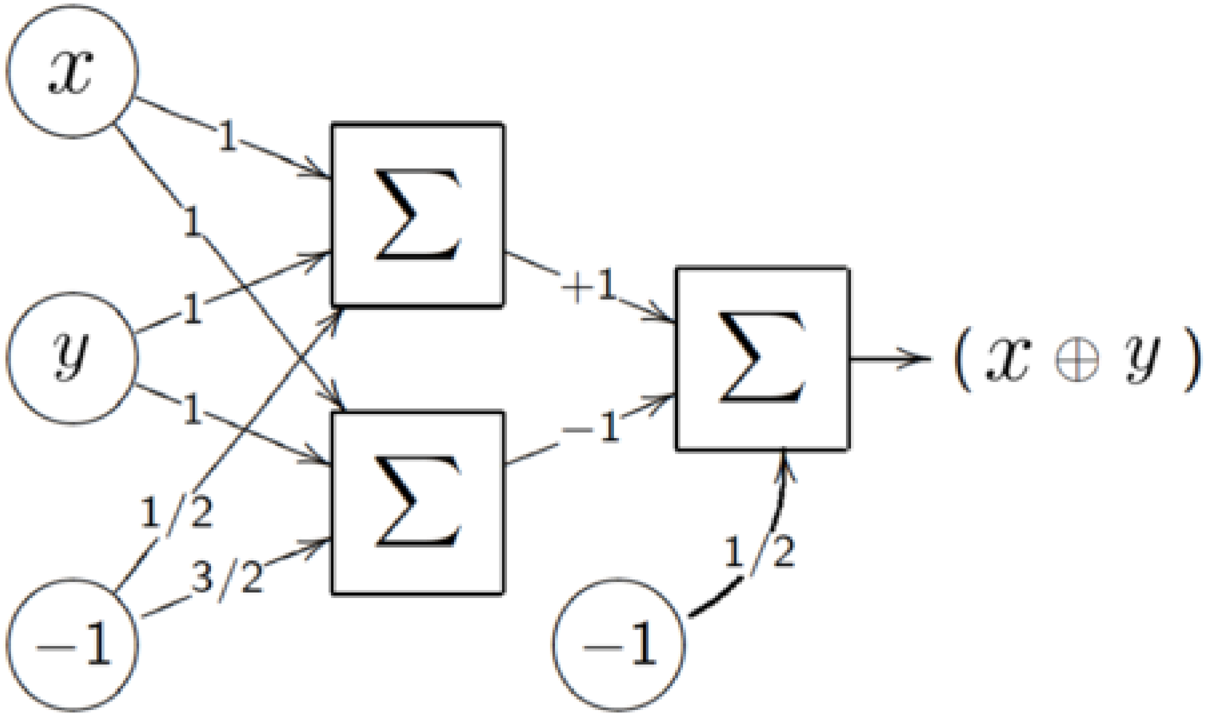


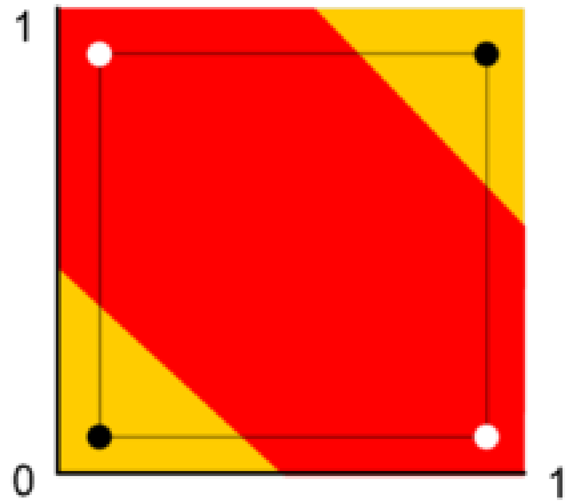


Реализация исключающего ИЛИ

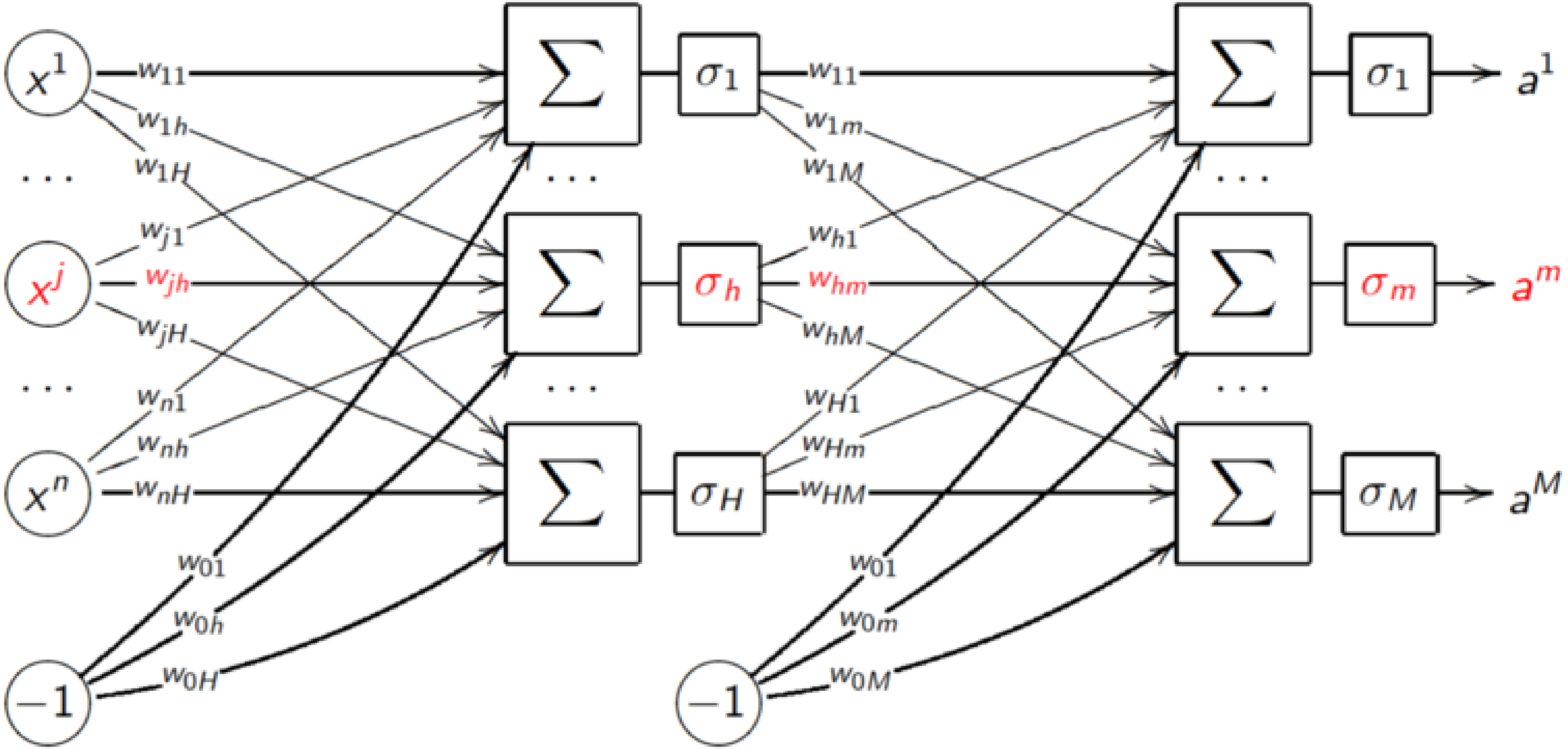
Функция $x \oplus y = [x \neq y]$ не реализуема ни одним нейроном. Сеть (двухслойная суперпозиция) функций И, ИЛИ, НЕ

$$x \oplus y = \left[(x \vee y) - (x \wedge y) - \frac{1}{2} > 0\right]$$





Многослойная нейронная сеть



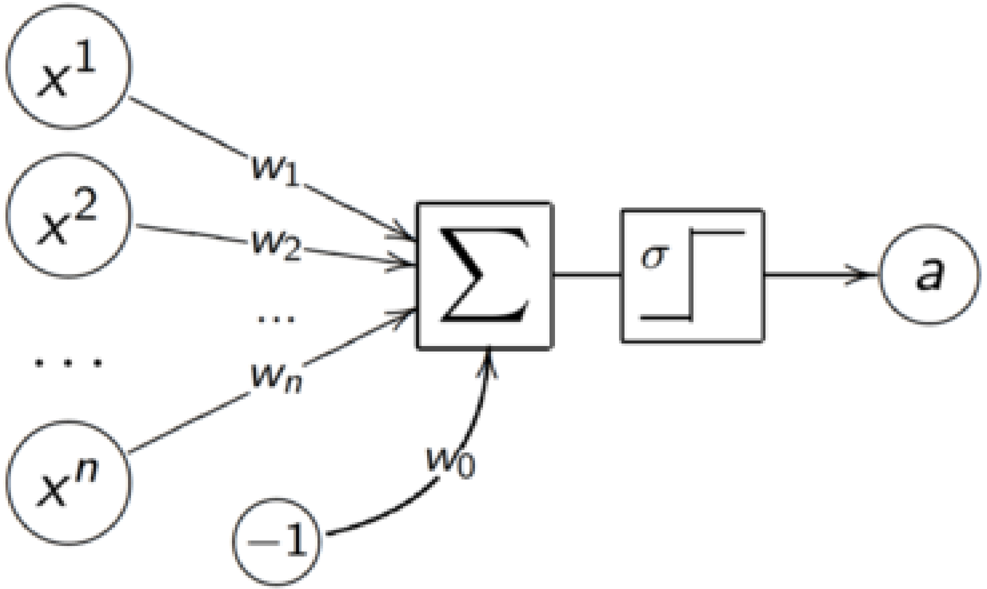
Представление функций нейросетью

- Двухслойная сеть в $\{0, 1\}^n$ позволяет реализовать произвольную булеву функцию
- Двухслойная сеть в $\mathbb{R}^n$ позволяет отделить произвольный многогранник
- Трёхслойная сеть в $\mathbb{R}^n$ позволяет отделить произвольную многогранную область, не обязательно выпуклую и связную
- С помощью линейных операций и одной нелинейной функции активации можно приблизить любую непрерывную функцию с любой заданной точностью

Персептрон Розенблатта

Задача классификации с двумя классами $y_i \in \{0, 1\}$
и бинарными признаками $f_j(x) \in \{0, 1\}$

$$a(x) = [\langle x, w \rangle > 0]$$



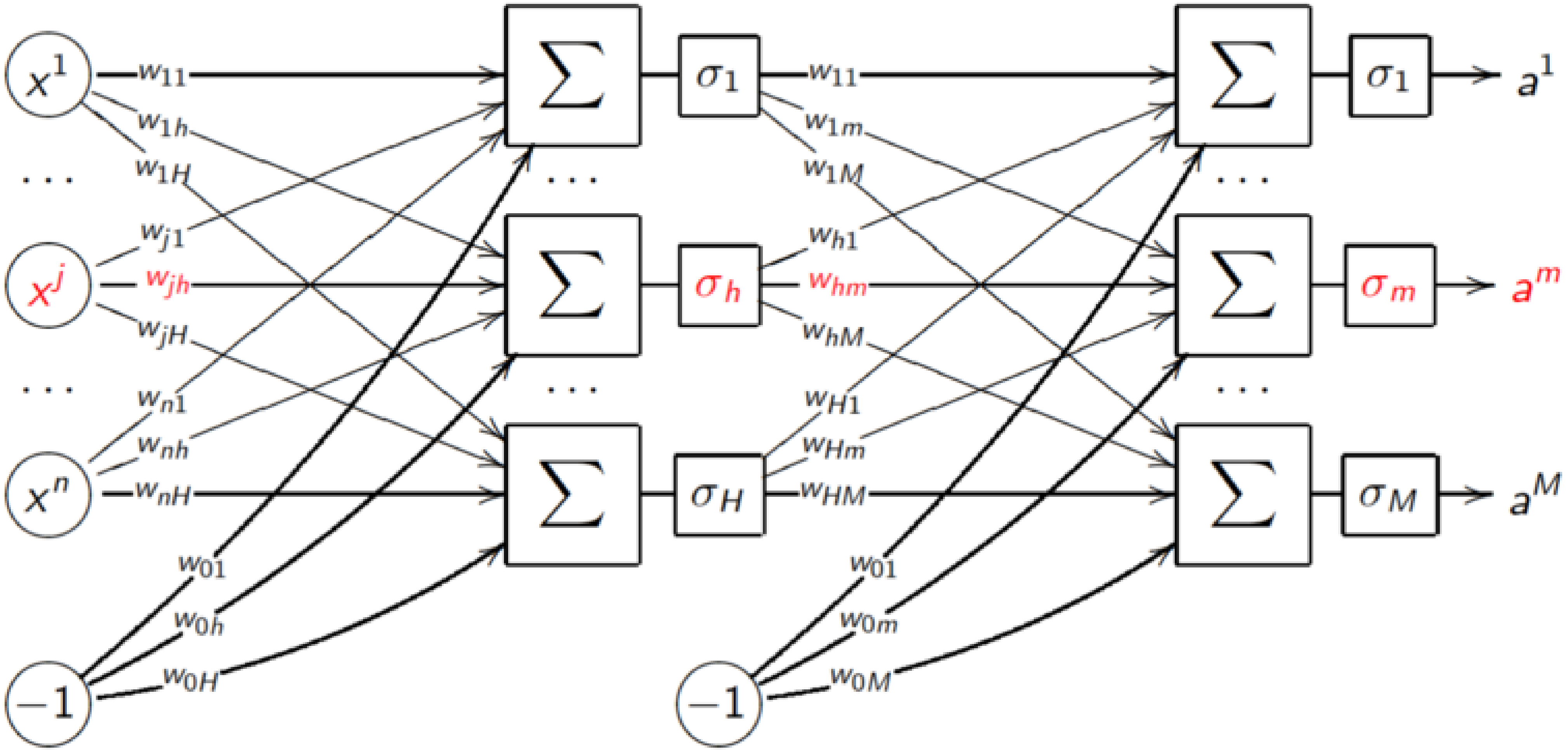
Исправление вектора весов в случае ошибки:

Если $a(x_i, w) = y_i$, то w менять не нужно

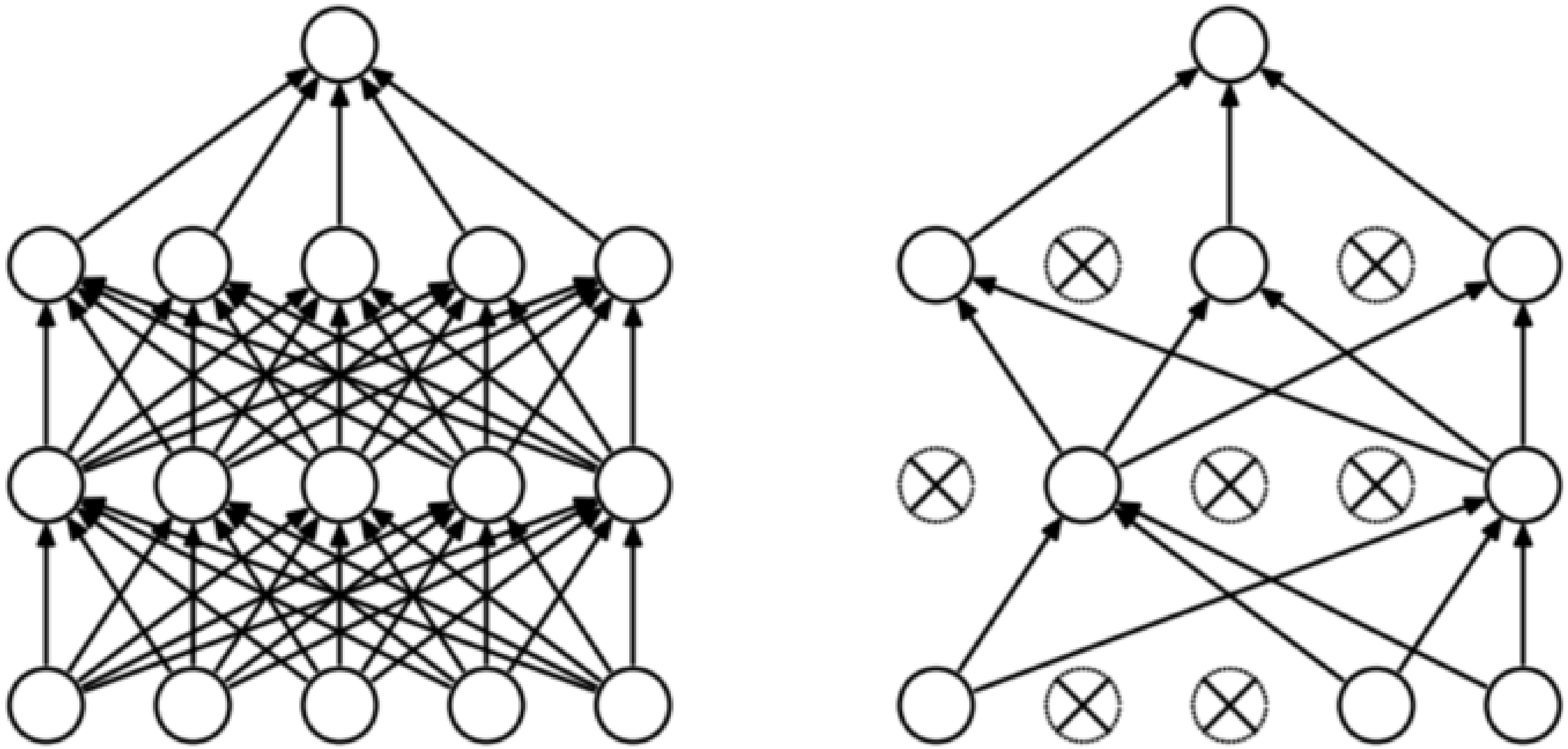
Если $a(x_i, w) = 0$, а $y_i = 1$, то $w_j := w_j + h \cdot f_j(x_i)$

Если $a(x_i, w) = 1$, а $y_i = 0$, то $w_j := w_j - h \cdot f_j(x_i)$

Многослойная нейронная сеть

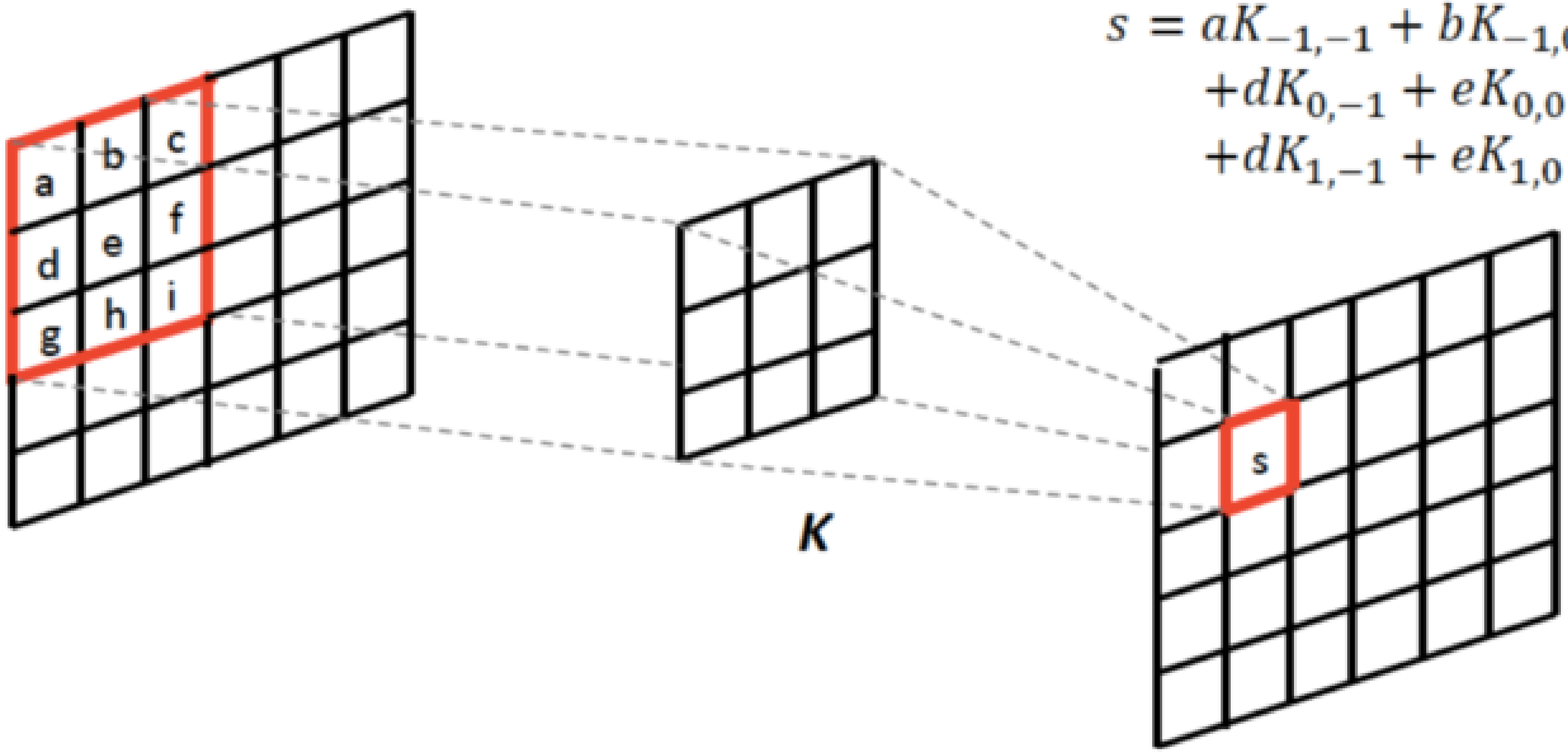


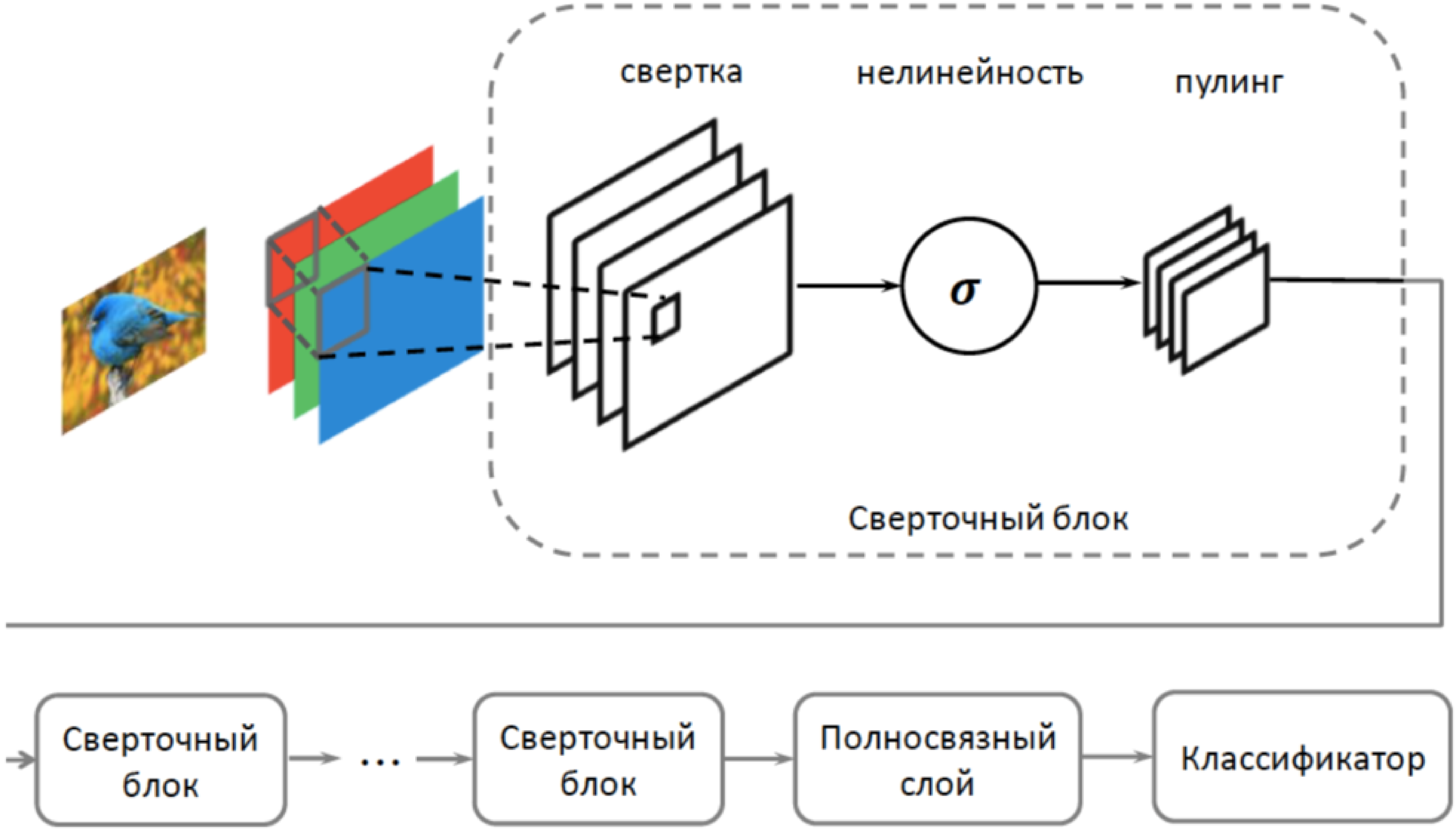
Слой Dropout

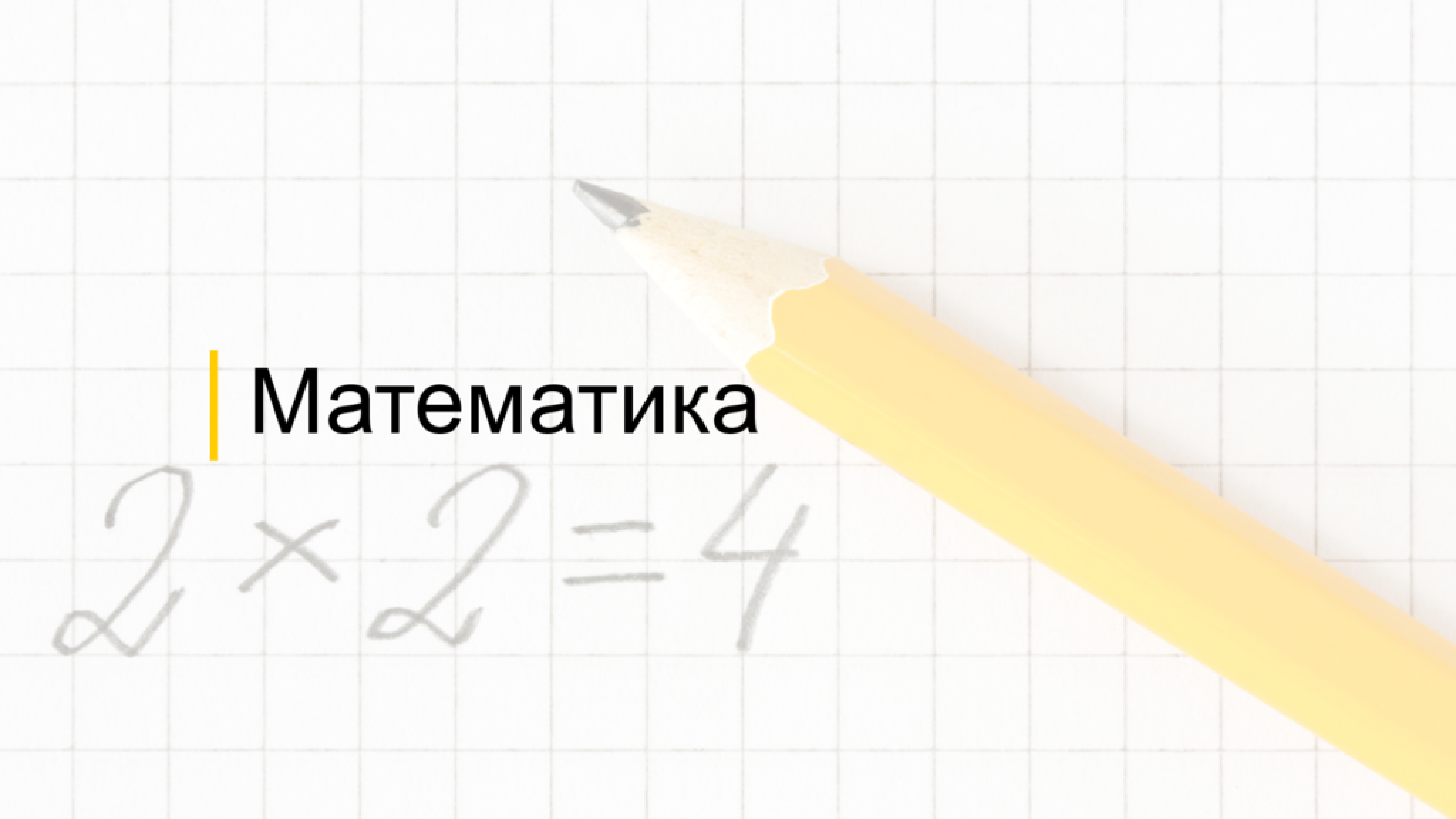


Принцип действия свертки

$$
\begin{aligned}
s = aK_{-1,-1} + bK_{-1,0} + cK_{-1,} \\
+dK_{0,-1} + eK_{0,0} + aK_{0,1} \\
+dK_{1,-1} + eK_{1,0} + aK_{1,1}
\end{aligned}
$$




свертка
нелинейность
пулинг
σ
Сверточный блок
Сверточный блок
...
Сверточный блок
Полносвязный слой
Классификатор


Математика
2 × 2 = 4

Математические дисциплины

Математический анализ

Теория алгоритмов

Теория информации

Линейная алгебра

Функциональный анализ

Теория вероятностей

Математическая статистика

Оптимизация