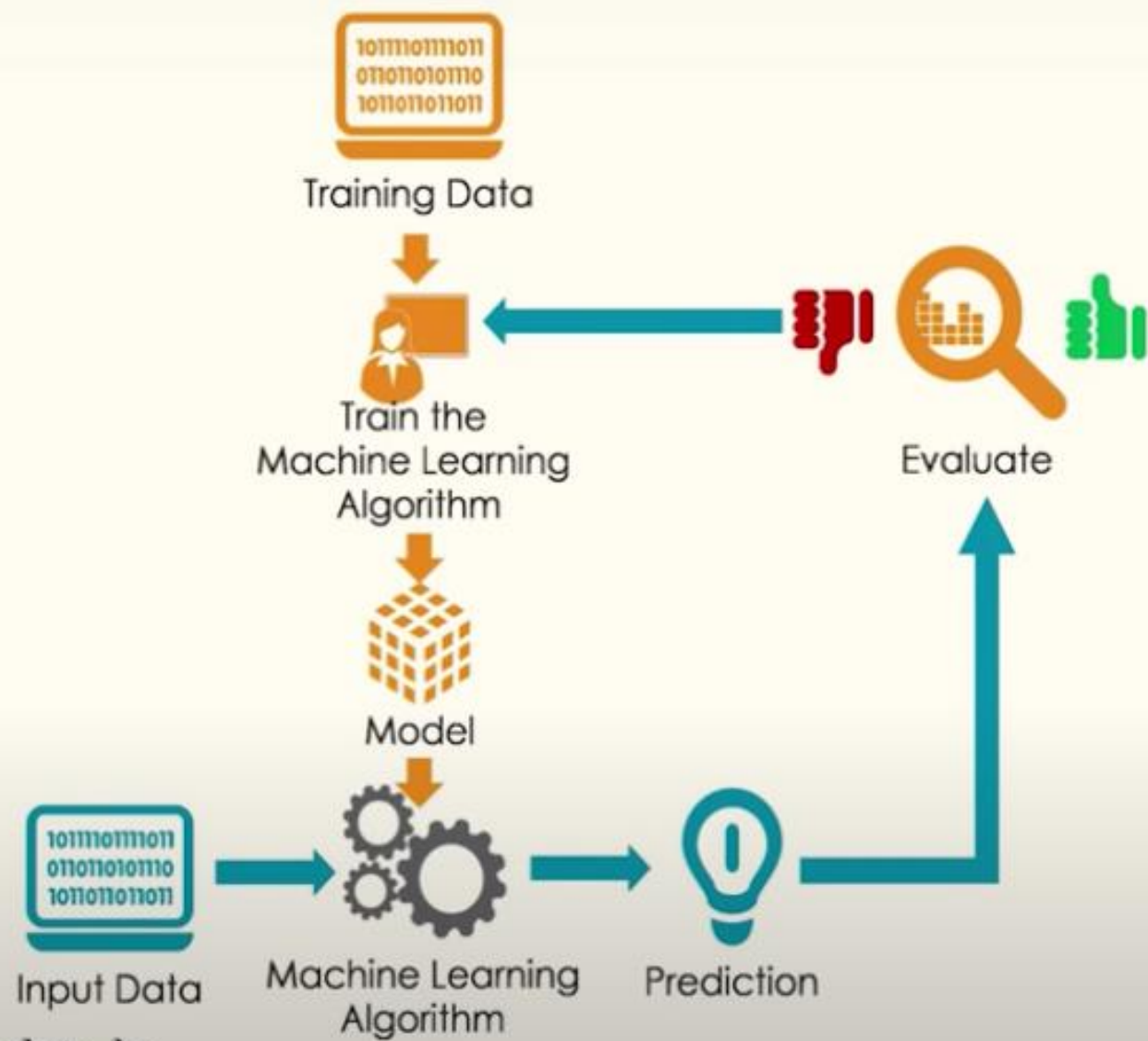


# Линейные модели и градиентный спуск



<http://www.7wdata.be>

$X$  – множество **объектов**

$Y$  – множество **допустимых ответов**

$y^*$  – целевая функция,  $y^*: X \rightarrow Y$ ,  $y_i = y^*(x_i)$  известны только на **конечном** подмножестве объектов  $x_1, \dots, x_m$  из  $X$

Пары  $(x_i, y_i)$  – прецеденты

Совокупность пар таких пар при  $i$  из  $1, \dots, m$  – **обучающая выборка** ( $X_{train}$ )

$a$  – **решающая функция** (алгоритм), которая любому объекту из  $X$  ставит в соответствие допустимый ответ из  $Y$  и приближает целевую функцию  $y^*$

$X_{test}$  – **выборка прецедентов** для тестирования построенного алгоритма  $a$

Для решения задачи обучения по прецедентам в первую очередь фиксируется восстанавливаемой зависимости.

X

 $y^*$ 

Features

| PassengerId | Survived | Pclass | Name  | Sex    | Age | SibSp | Parch | Ticket          |
|-------------|----------|--------|---|--------|-----|-------|-------|-----------------|
| 1           | 0        | 3      | Braund, Mr. Owen Harris                             | male   | 22  | 1     | 0     | A/5 21171       |
| 2           | 1        | 1      | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 38  | 1     | 0     | PC 17599        |
| 3           | 1        | 3      | Heikkinen, Miss. Laina                              | female | 26  | 0     | 0     | STON/O2. 310128 |
| 4           | 1        | 1      | Futrelle, Mrs. Jacques Heath (Lily May Peel)        | female | 35  | 1     | 0     | 113803          |
| 5           | 0        | 3      | Allen, Mr. William Henry                            | male   | 35  | 0     | 0     | 373450          |
| 6           | 0        | 3      | Moran, Mr. James                                    | male   |     | 0     | 0     | 330877          |
| 7           | 0        | 1      | McCarthy, Mr. Timothy J                             | male   | 54  | 0     | 0     | 17463           |
| 8           | 0        | 3      | Palsson, Master. Gosta Leonard                      | male   | 2   | 3     | 1     | 349909          |

 $Y = \{0,1\}$

**Признак** (feature)  $f$  объекта  $x$  — это результат измерения некоторой характеристики объекта. Формально признаком называется отображение  $f : X \rightarrow D_f$ , где  $D_f$  — множество допустимых значений признака. В частности, любой алгоритм  $a : X \rightarrow Y$  также можно рассматривать как признак

Пусть дан набор признаков  $f_1(x), \dots, f_n(x)$ .

**Признаковое описание объекта**  $x$  — вектор (одномерный массив)  $(f_1, \dots, f_n)$ . Совокупность признаковых описаний всех объектов выборки длины  $m$ , записанную в виде таблицы размера  $mn$ , называют матрицей объектов–признаков.

# Как строится функция $a$ ?

**Обучающая выборка** — выборка, по которой производится настройка (оптимизация параметров) модели зависимости.

**Тестовая выборка** — выборка, по которой оценивается качество построенной модели.

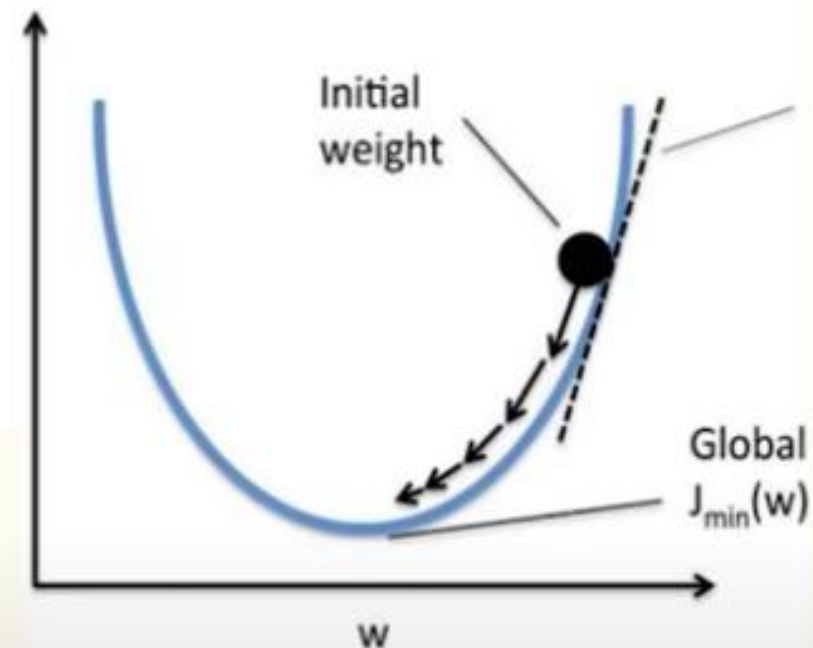
**Функционал качества (обучение с учителем)** — определяется как средняя ошибка ответов, выданных алгоритмом, по всем объектам выборки.

$$L(\hat{y}, y) = I(\hat{y} \neq y),$$

0      1

$$\text{logloss} = -\frac{1}{l} \cdot \sum_{i=1}^l (y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i))$$

Минимизация функции ошибки

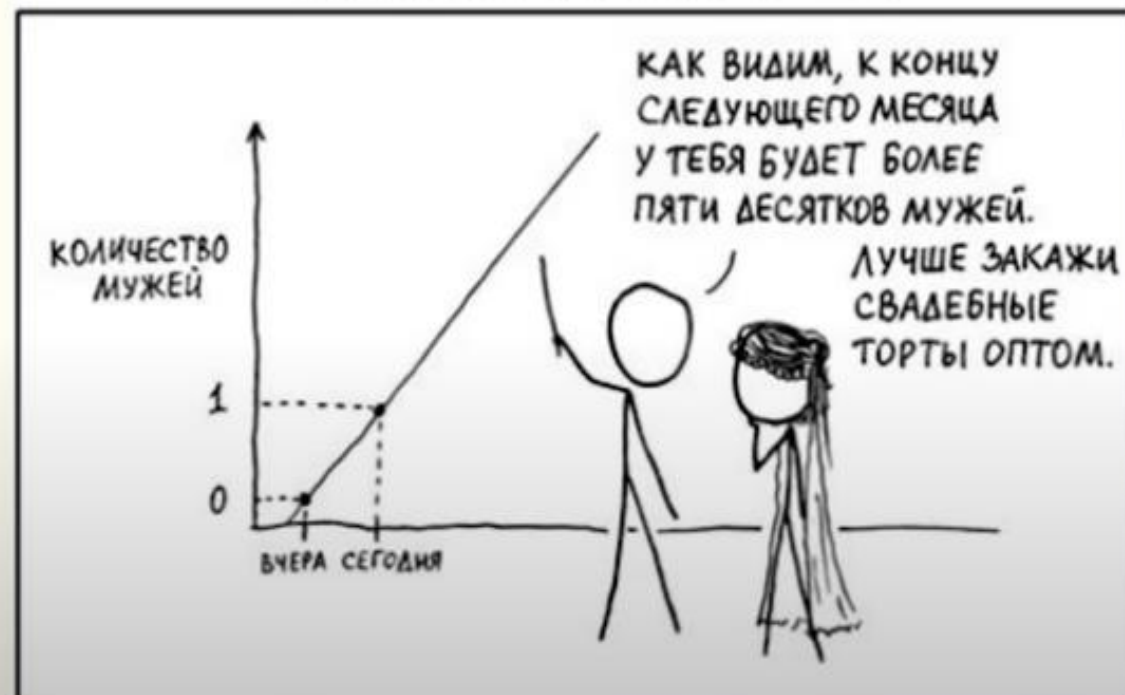


# Линейные модели

# Линейная регрессия

метод восстановления зависимости между двумя или более переменными

МОЁ ХОББИ: ЭКСТРАПОЛИРОВАТЬ

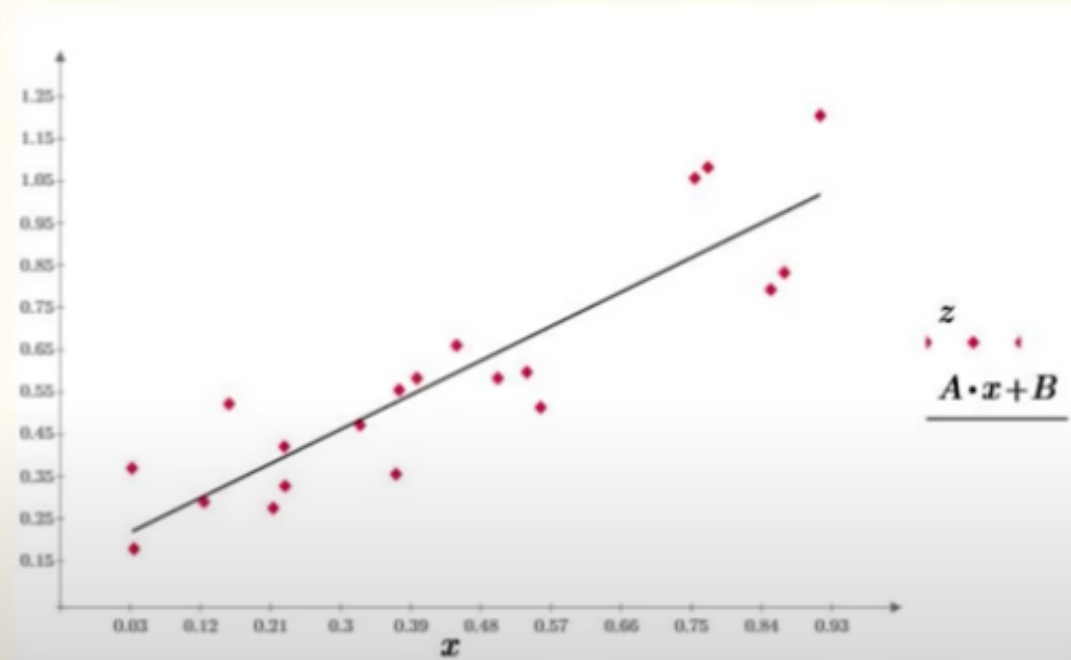




# Одномерная линейная регрессия

$(x, y)$  -- пары точек

**Задача:** построить  
предсказания по  $x$   
для неизвестных  $y$  в  
предположении, что  
 $y(x)$  -- линейная  
функция



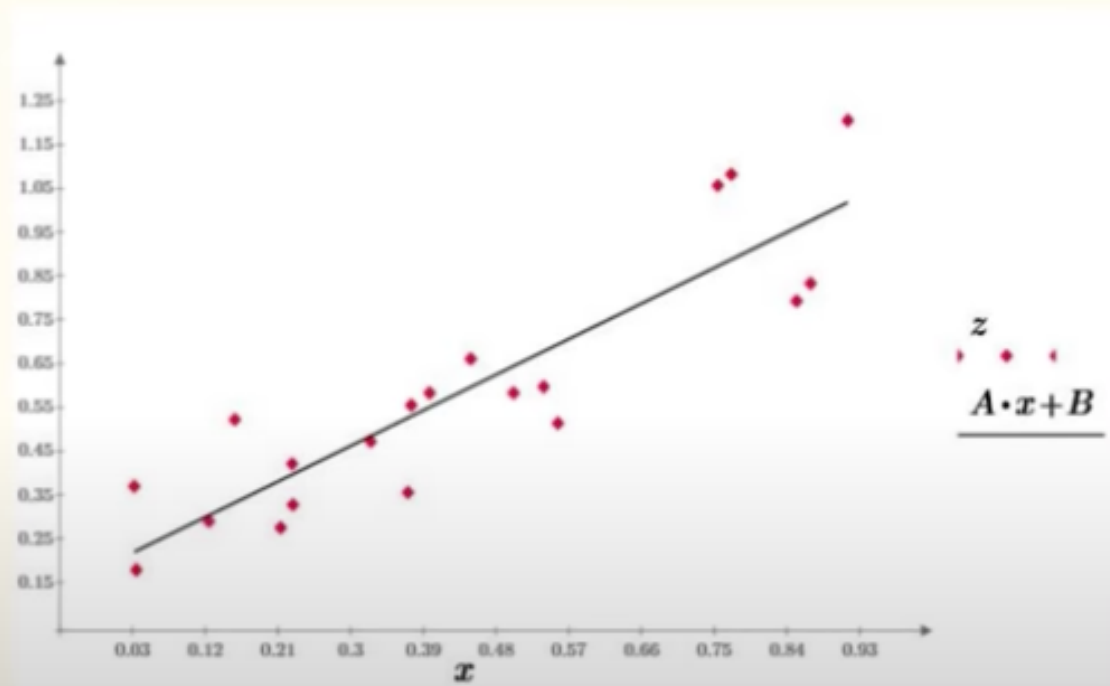
Чем характеризуется прямая?

# Одномерная линейная регрессия

$(x, y)$  -- пары точек

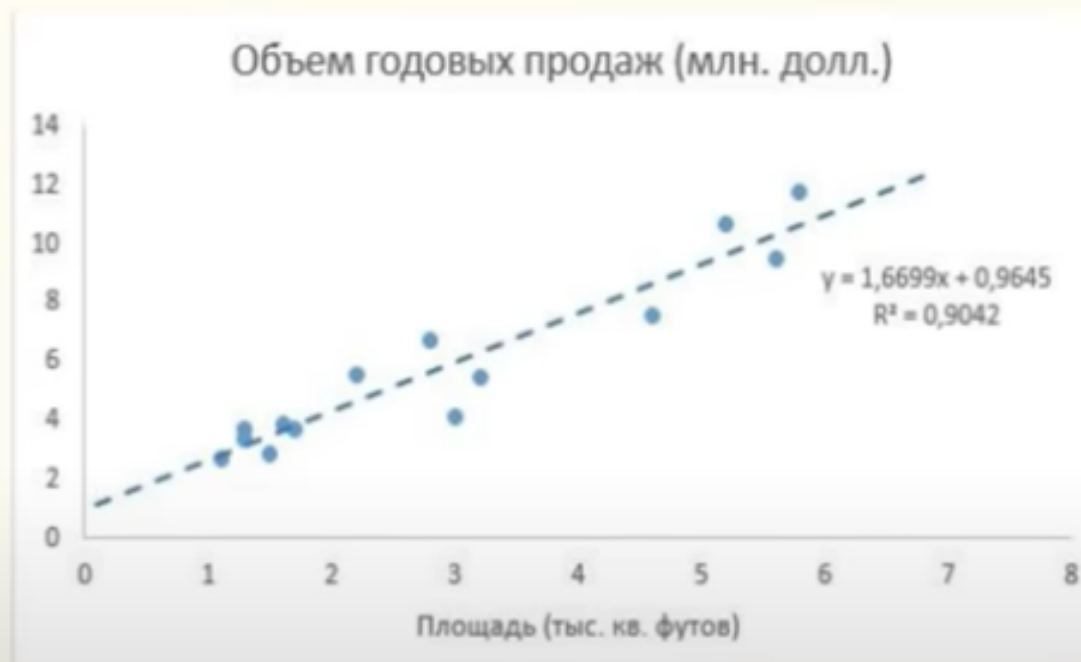
Задача: построить  
предсказания по  $x$  для  
неизвестных  $y$  в  
предположении, что  $y(x)$   
-- линейная функция

$$y = Ax + B$$



# Пример

Наша цель — предсказать объем годовых продаж для всех новых магазинов, зная их размеры.



$$y = Ax + B$$

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j.$$

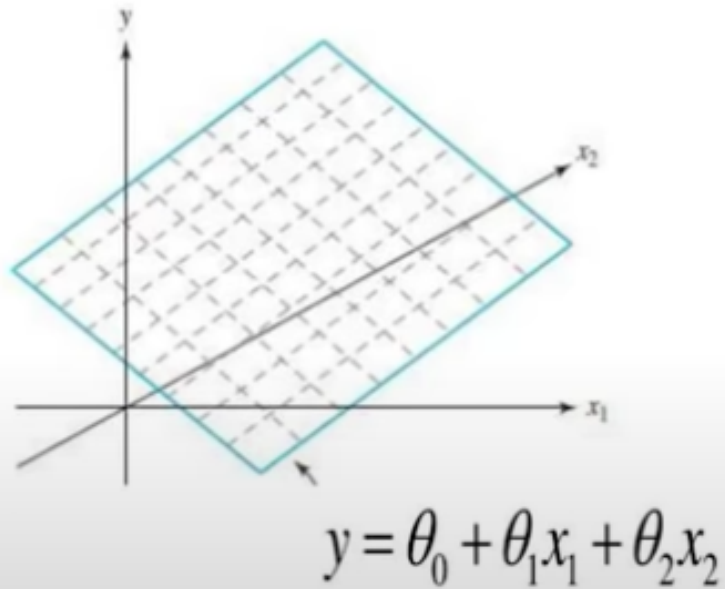
$w_j$  — веса признаков

$w_0$  — смещение (bias)

$$y = \mathbf{Ax} + \mathbf{B}$$

Например: местоположение, экономическая ситуация и проч.

## Двумерная регрессия



$$y_1 = ax_1 + b + \varepsilon_1$$

$$y_2 = ax_2 + b + \varepsilon_2$$

.....

$$y_n = ax_n + b + \varepsilon_n$$

```
houses = pd.read_csv("kc_house_data.csv")
```

```
houses.head()
```

|   | id         | date            | price     | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | ... | grade | sqft_above | sqft_base |
|---|------------|-----------------|-----------|----------|-----------|-------------|----------|--------|------------|------|-----|-------|------------|-----------|
| 0 | 7129300520 | 20141013T000000 | 221900.00 | 3        | 1.00      | 1180        | 5650     | 1.00   | 0          | 0    | ... | 7     | 1180       | 0         |
| 1 | 6414100192 | 20141209T000000 | 538000.00 | 3        | 2.25      | 2570        | 7242     | 2.00   | 0          | 0    | ... | 7     | 2170       | 400       |
| 2 | 5631500400 | 20150225T000000 | 180000.00 | 2        | 1.00      | 770         | 10000    | 1.00   | 0          | 0    | ... | 6     | 770        | 0         |
| 3 | 2487200875 | 20141209T000000 | 604000.00 | 3        | 3.00      | 1960        | 5000     | 1.00   | 0          | 0    | ... | 7     | 1050       | 910       |
| 4 | 1954400510 | 20150218T000000 | 510000.00 | 3        | 2.00      | 1680        | 8080     | 1.00   | 0          | 0    | ... | 8     | 1680       | 0         |

X

|   | id         | date            | price     | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | ... | grade | sqft_above | sqft_basement |
|---|------------|-----------------|-----------|----------|-----------|-------------|----------|--------|------------|------|-----|-------|------------|---------------|
| 0 | 7129300520 | 20141013T000000 | 221900.00 | 3        | 1.00      | 1180        | 5650     | 1.00   | 0          | 0    | ... | 7     | 1180       | 0             |
| 1 | 6414100192 | 20141209T000000 | 538000.00 | 3        | 2.25      | 2570        | 7242     | 2.00   | 0          | 0    | ... | 7     | 2170       | 400           |
| 2 | 5631500400 | 20150225T000000 | 180000.00 | 2        | 1.00      | 770         | 10000    | 1.00   | 0          | 0    | ... | 6     | 770        | 0             |
| 3 | 2487200875 | 20141209T000000 | 604000.00 | 4        | 3.00      | 1960        | 5000     | 1.00   | 0          | 0    | ... | 7     | 1050       | 910           |
| 4 | 1954400510 | 20150218T000000 | 510000.00 | 3        | 2.00      | 1680        | 8080     | 1.00   | 0          | 0    | ... | 8     | 1680       | 0             |

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j.$$

|   | id         | date            | price     | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | ... | grade | sqft_above | sqft_basement |
|---|------------|-----------------|-----------|----------|-----------|-------------|----------|--------|------------|------|-----|-------|------------|---------------|
| 0 | 7129300520 | 20141013T000000 | 221900.00 | 3        | 1.00      | 1180        | 5650     | 1.00   | 0          | 0    | ... | 7     | 1180       | 0             |
| 1 | 6414100192 | 20141209T000000 | 538000.00 | 3        | 2.25      | 2570        | 7242     | 2.00   | 0          | 0    | ... | 7     | 2170       | 400           |
| 2 | 5631500400 | 20150225T000000 | 180000.00 | 2        | 1.00      | 770         | 10000    | 1.00   | 0          | 0    | ... | 6     | 770        | 0             |
| 3 | 2487200875 | 20141209T000000 | 604000.00 | 3        | 3.00      | 1960        | 5000     | 1.00   | 0          | 0    | ... | 7     | 1050       | 910           |
| 4 | 1954400510 | 20150218T000000 | 510000.00 | 3        | 2.00      | 1680        | 8080     | 1.00   | 0          | 0    | ... | 8     | 1680       | 0             |

$$y = (y_1, \dots, y_\ell)$$

$$x = (x_1, \dots, x_\ell)$$

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

$$y^* = a(x) = w_0 + \sum_{j=1}^d w_j x_j$$

$$X = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$$



|   | id         | date            | price     | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | ... | grade | sqft_above | sqft_basement |
|---|------------|-----------------|-----------|----------|-----------|-------------|----------|--------|------------|------|-----|-------|------------|---------------|
| 0 | 7129300520 | 20141013T000000 | 221900.00 | 3        | 1.00      | 1180        | 5650     | 1.00   | 0          | 0    | ... | 7     | 1180       | 0             |
| 1 | 6414100192 | 20141209T000000 | 538000.00 | 3        | 2.25      | 2570        | 7242     | 2.00   | 0          | 0    | ... | 7     | 2170       | 400           |
| 2 | 5631500400 | 20150225T000000 | 180000.00 | 2        | 1.00      | 770         | 10000    | 1.00   | 0          | 0    | ... | 6     | 770        | 0             |
| 3 | 2487200875 | 20141209T000000 | 604000.00 | 4        | 3.00      | 1960        | 5000     | 1.00   | 0          | 0    | ... | 7     | 1050       | 910           |
| 4 | 1954400510 | 20150218T000000 | 510000.00 | 3        | 2.00      | 1680        | 8080     | 1.00   | 0          | 0    | ... | 8     | 1680       | 0             |

$$\overline{y} = (y_1, \dots, y_\ell)$$

$$\overline{x} = (x_1, \dots, x_\ell)$$

$$\overline{w} = (w_1, \dots, w_d)$$

$$\overline{y}^* = a(x) = w_0 + \sum_{j=1}^d w_j x_j.$$

$$\overline{y}^* = a(x) = w_0 + \langle \overline{w}, \overline{x} \rangle.$$

Скалярное произведение

$$X = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$$

Матричная запись

$$\begin{array}{l} y_1 = ax_1 + b + \varepsilon_1 \\ y_2 = ax_2 + b + \varepsilon_2 \\ \dots \\ y_n = ax_n + b + \varepsilon_n \end{array}$$

$$\vec{y} = X\vec{w} + \epsilon,$$

Непрогнозируемая ошибка

Фиктивная переменная (случай 2х параметров, n наблюдений)

$$X\omega = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ .. & .. \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} ax_1 + b \\ ax_2 + b \\ ..... \\ ax_n + b \end{pmatrix}$$

$$\vec{y} = X\vec{w} + \epsilon,$$

$$\omega = \begin{pmatrix} b \\ a \end{pmatrix}$$

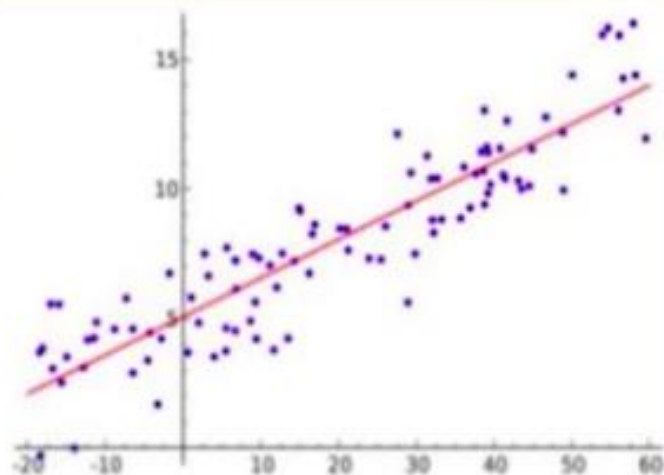
$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ .. & .. \\ 1 & x_n \end{pmatrix}$$

# Линейная регрессия

$$\vec{y} = X\vec{w} + \epsilon,$$

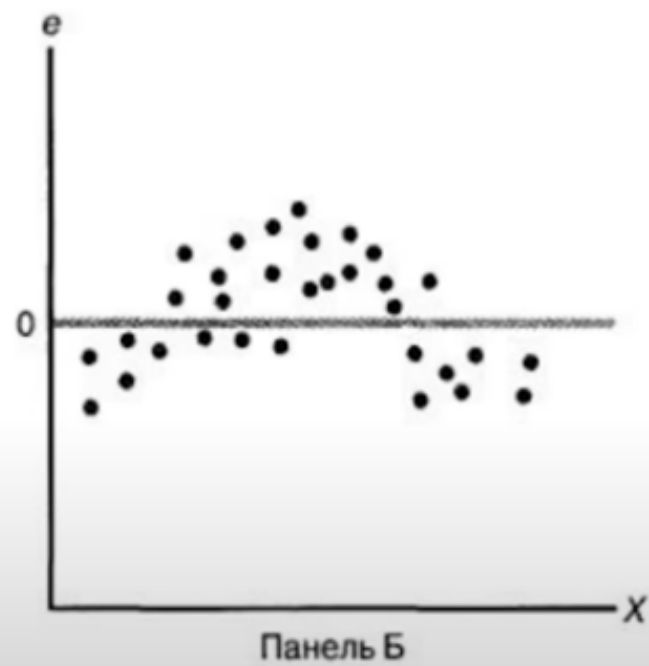
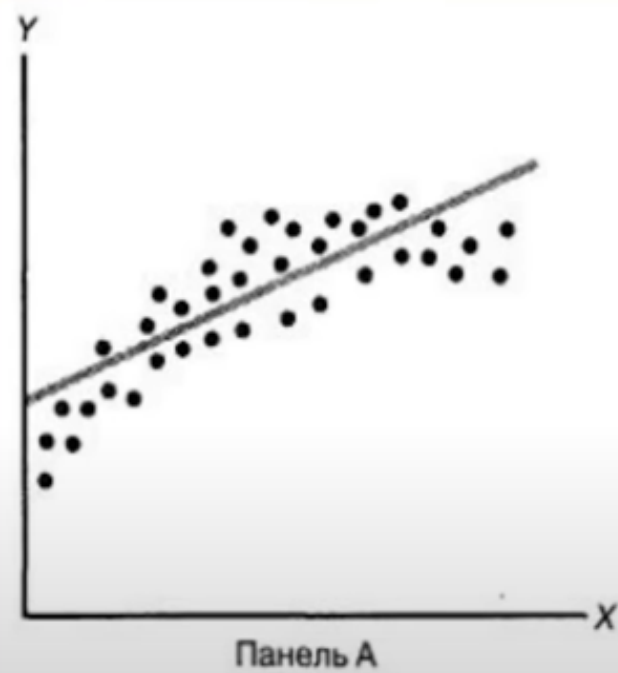
где

- $\vec{y} \in \mathbb{R}^n$  – объясняемая (или целевая) переменная;
- $w$  – вектор параметров модели (в машинном обучении эти параметры часто называют весами);
- $X$  – матрица наблюдений и признаков размерности  $n$  строк на  $m + 1$  столбцов (включая фиктивную единичную колонку слева) с полным рангом по столбцам:  $\text{rank}(X) = m + 1$ ;
- $\epsilon$  – случайная переменная, соответствующая случайной, непрогнозируемой ошибке модели.

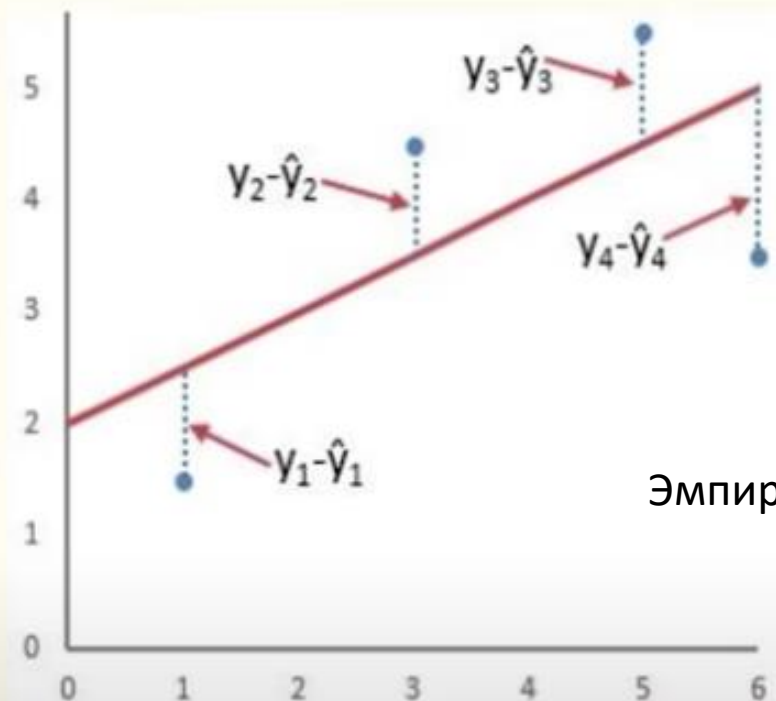


# Обучение линейной регрессии

# Построение прогноза



# Метод наименьших квадратов (одномерный случай)



Функция потерь — квадратичная:

$$\mathcal{L}(a, y) = (a - y)^2$$

Метод обучения — метод наименьших квадратов:

Эмпирический риск

$$Q(w) = \sum_{i=1}^{\ell} (a(x_i, w) - y_i)^2 \rightarrow \min_w$$

Как можно минимизировать  
эмпирический риск?

## МНК (многомерный случай)

То, что нам выдал  
алгоритм

Реальные  
значения

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} (f(x_i, \alpha) - y_i)^2 = \|F\alpha - y\|^2 \rightarrow \min_{\alpha}.$$

$$f(x, \alpha) = \sum_{j=1}^n \alpha_j f_j(x), \quad \alpha \in \mathbb{R}^n.$$

Квадратичная  
норма

Матричные обозначения:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad y_{\ell \times 1} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \alpha_{n \times 1} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}.$$



## Метод градиентного спуска

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} (f(x_i, \alpha) - y_i)^2 = \|F\alpha - y\|^2 \rightarrow \min_{\alpha}$$

Нужно подобрать вектор параметров

Минимизируем ошибку путем минимизации  
функции!

# Поиск минимума функции

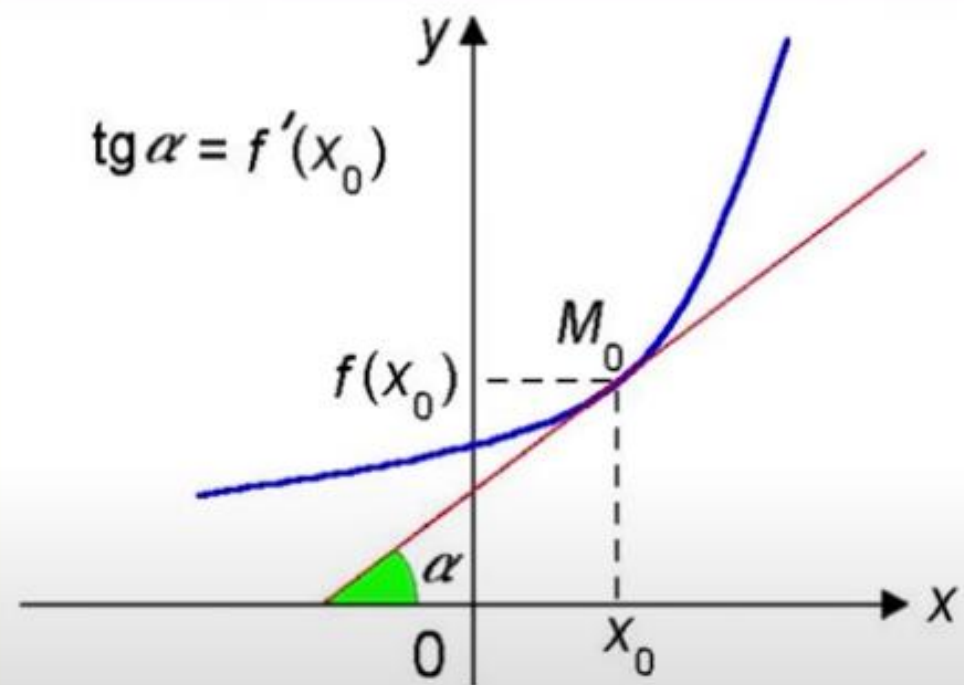
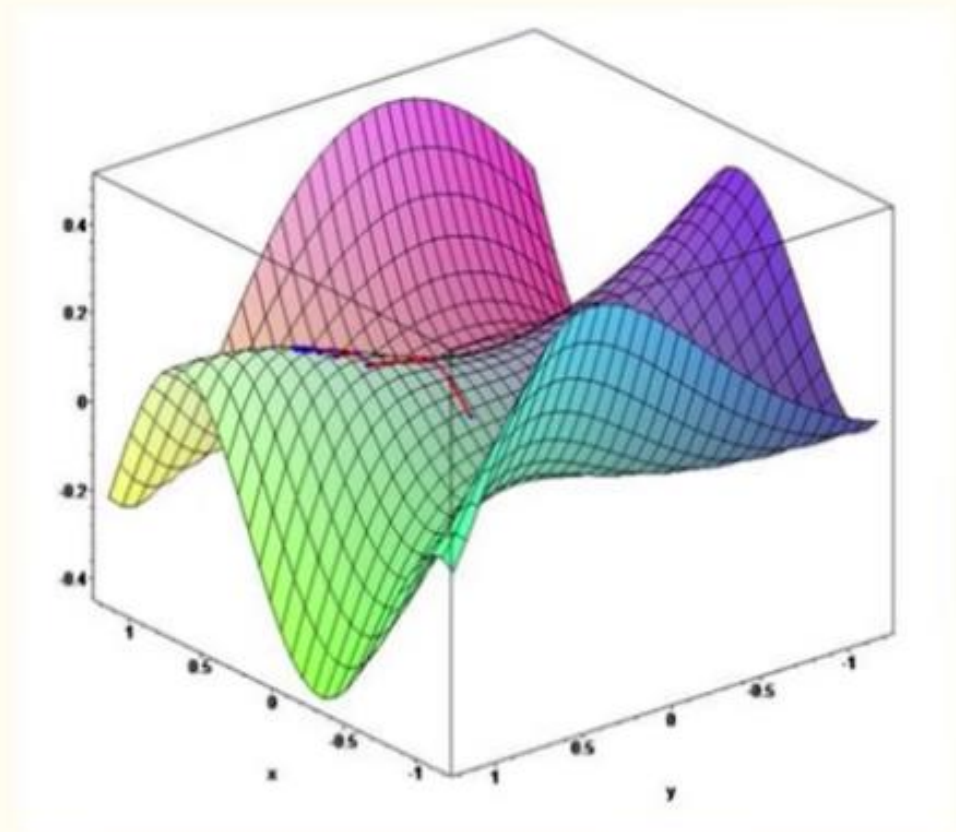


Рис. 1

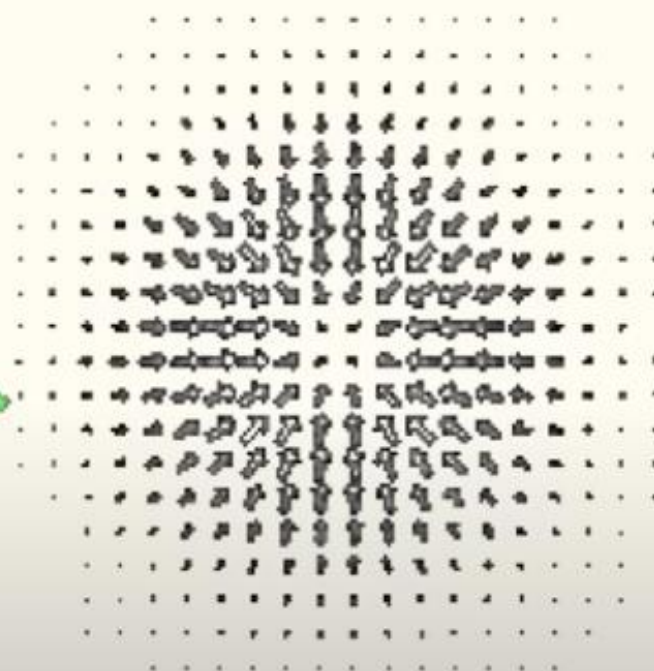
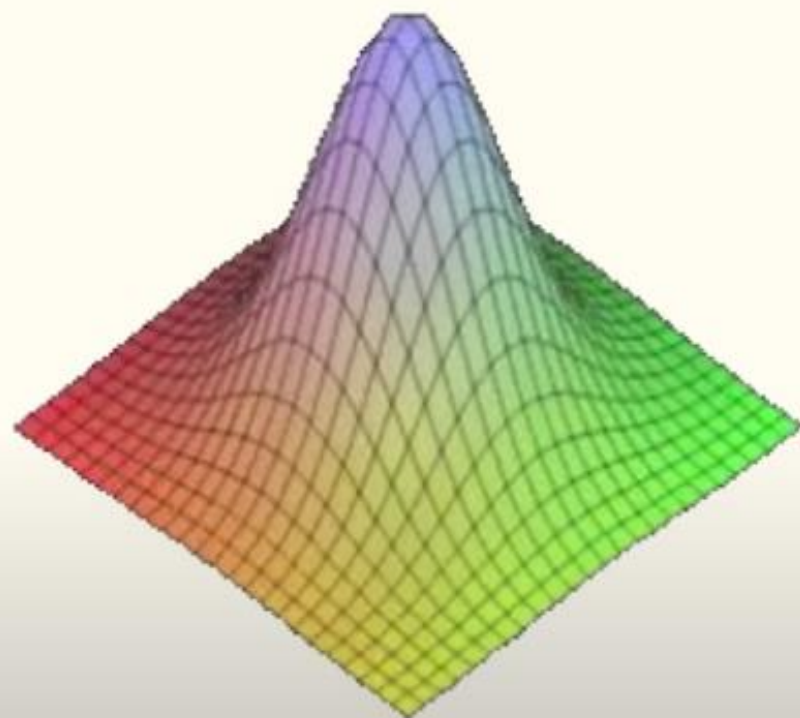
# Многомерный случай



Градиент

$\nabla \varphi$

$\text{grad } \varphi$



Локальный минимум

# Градиентный спуск (GD)

$$Q(w) = \sum_{i=1}^{\ell} \mathcal{L}_i(w) \rightarrow \min_w$$

Сумма функции ошибки

Численная минимизация методом *градиентного спуска*:

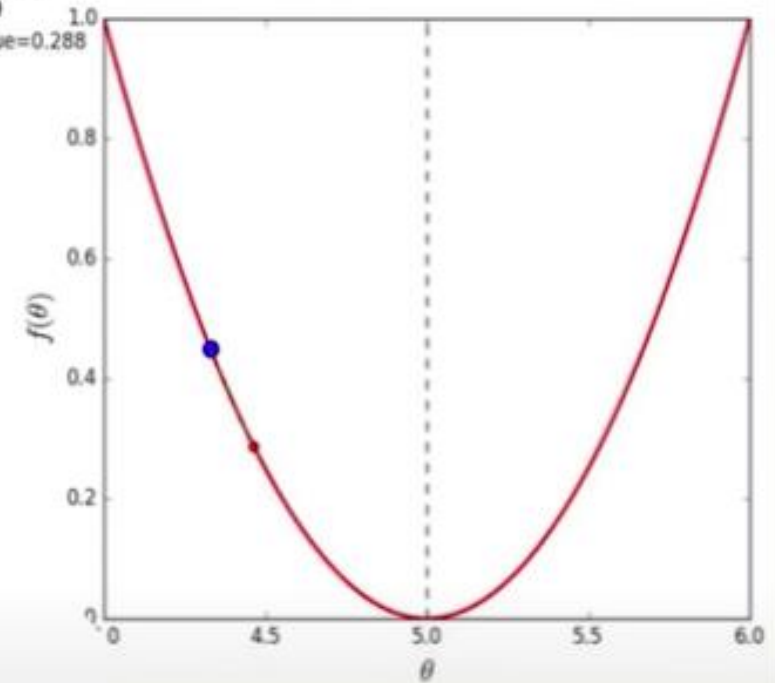
$w^{(0)}$  := начальное приближение:

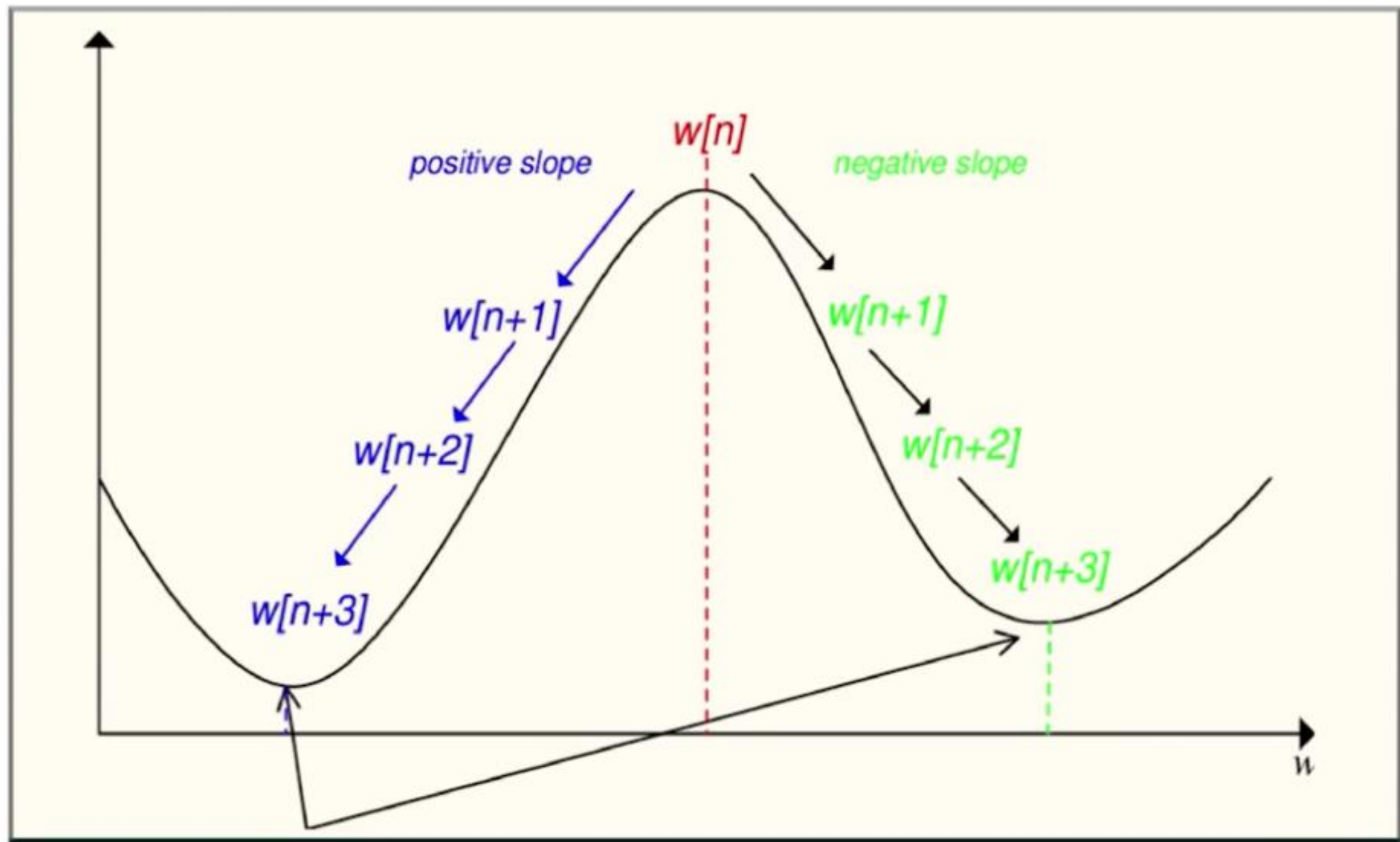
$$w^{(t+1)} := w^{(t)} - h \cdot \nabla Q(w^{(t)}),$$

где  $h$  — *градиентный шаг*, называемый также *темпом обучения* Learning raid – шаг спуска

$$w^{(t+1)} := w^{(t)} - h \sum_{i=1}^{\ell} \nabla \mathcal{L}_i(w^{(t)}).$$

Rate: 0.1  
Step: 9  
Func value=0.288  
 $\theta=4.463$





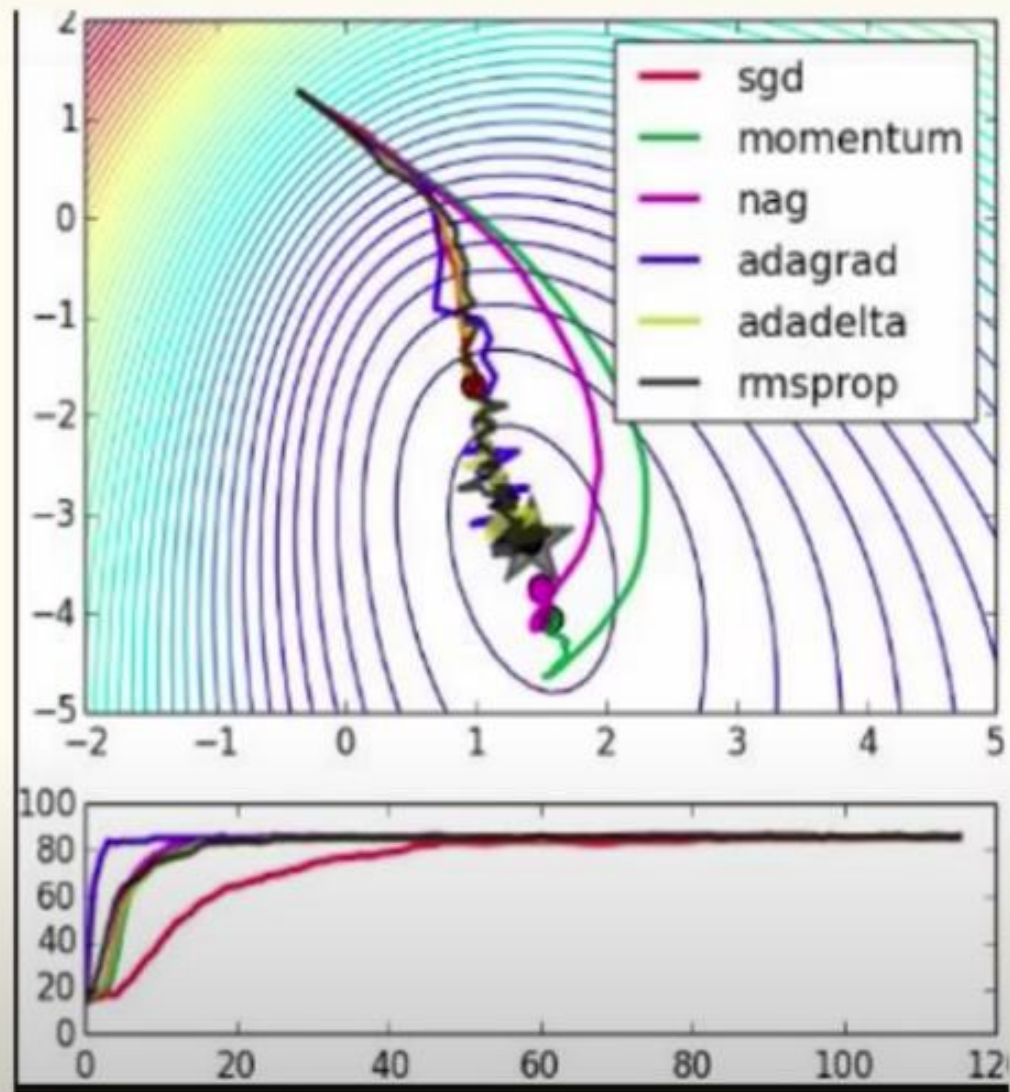
## Ускорения сходимости к минимуму?

1. Брать по одной новой паре  $(x, y)$  и сразу обновлять вектор весов
2. Просматривать не в одном порядке, а в случайном

$$\text{GD} + 1) + 2) = \text{SGD}$$

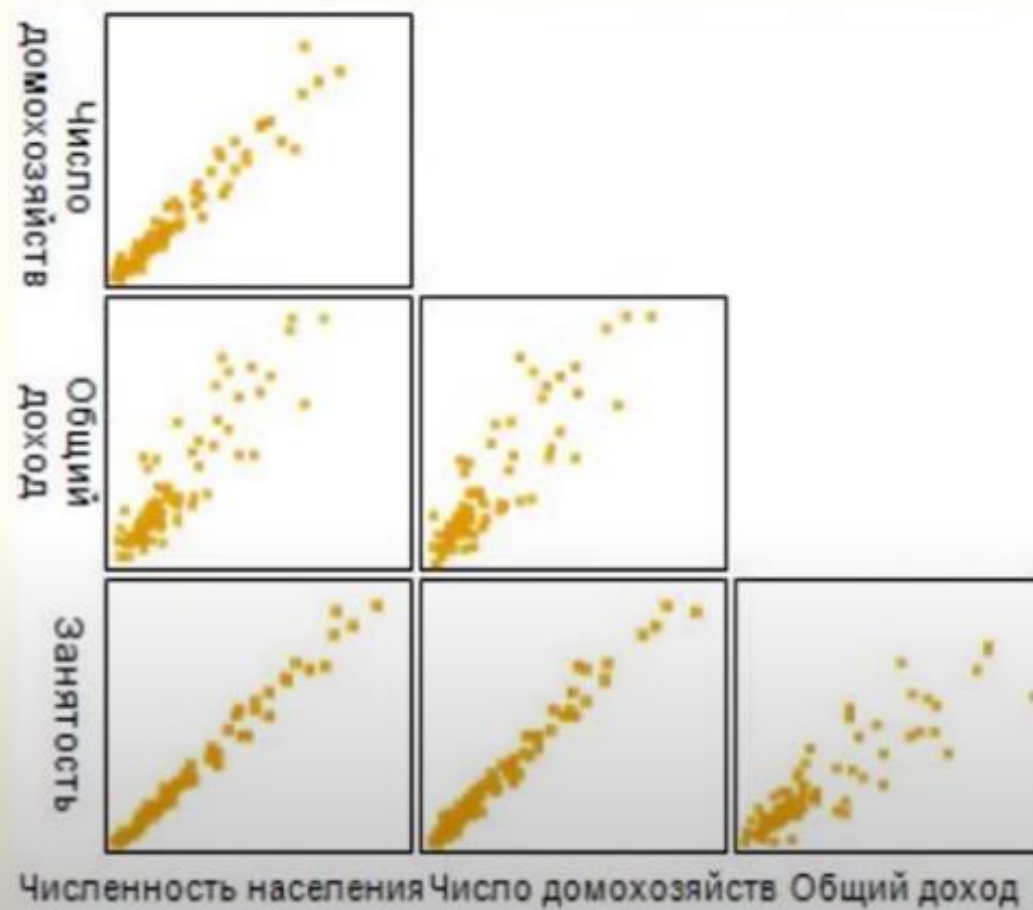
(стохастический градиентный спуск)





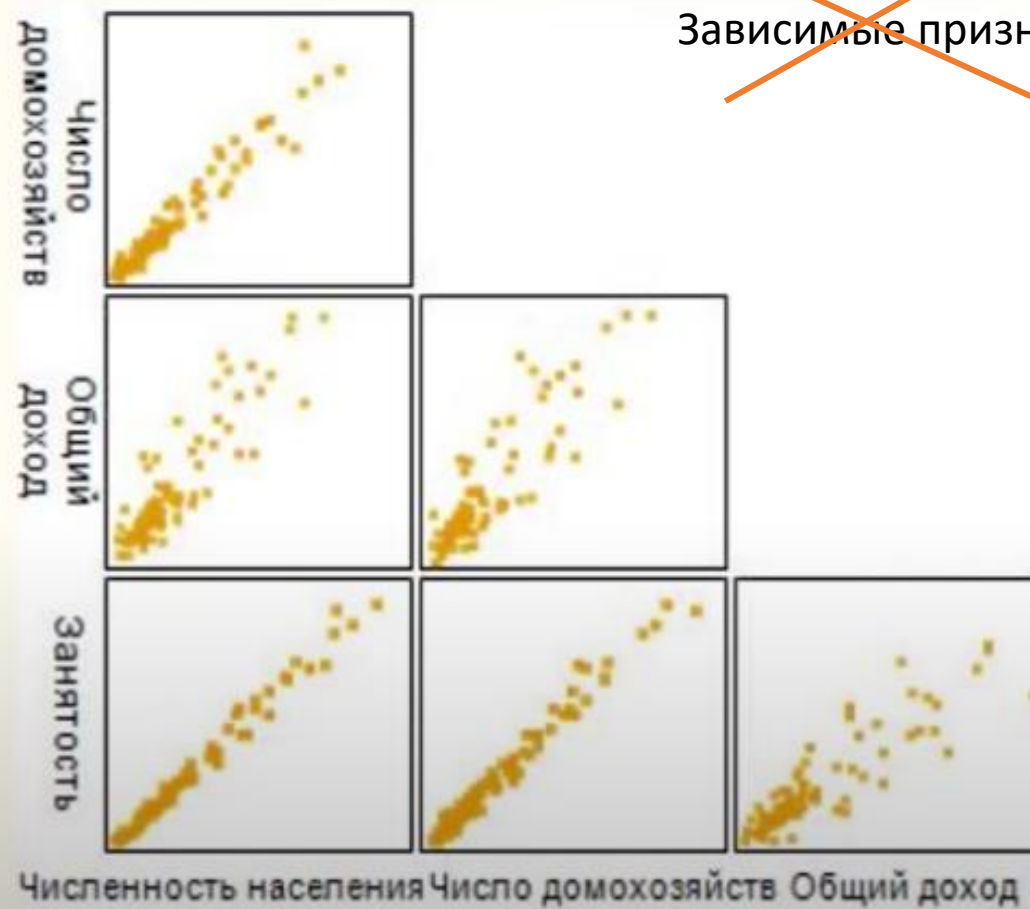


# Problems?



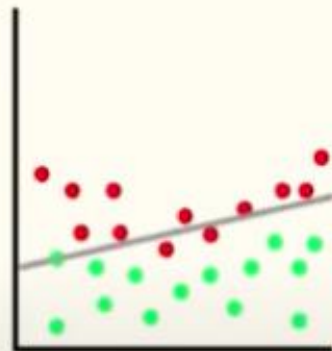
# Problems?

~~Зависимые признаки~~

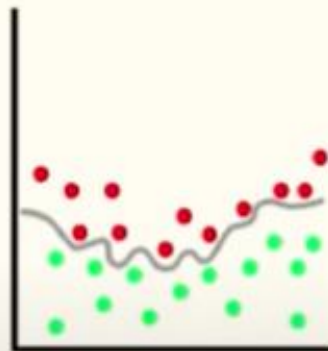


# Problems?

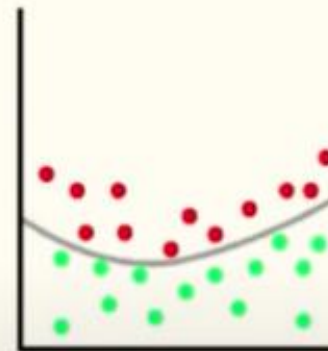
learning & regularization



Underfitting



Overfitting




Balanced

Что делать?

1. Регуляризация: штраф за увеличение нормы вектора весов
2. Отбор признаков

Коэффициент  
регуляризации весов



$$Q_{\tau}(\alpha) = \|F\alpha - y\|^2 + \frac{1}{\sigma}\|\alpha\|^2,$$

где  $\tau = \frac{1}{\sigma}$  — неотрицательный *параметр регуляризации*.

# Связь линейного классификатора и нейрона

# Линейный классификатор

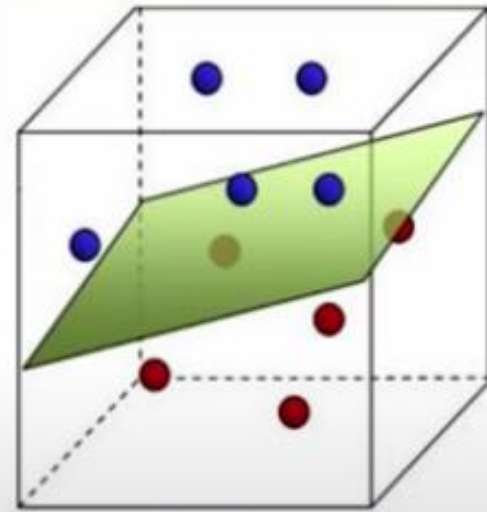
Как из линейной регрессии сделать линейную классификацию?

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j.$$

# Линейный классификатор

Как из линейной регрессии сделать линейную классификацию?

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j.$$



$$a(x, w) = \text{sign} \langle x, w \rangle = \text{sign} \sum_{j=1}^n w_j f_j(x)$$

## Линейный классификатор как модель нейрона

Линейная модель нейрона МакКаллока-Питтса [1943]:

$$a(x, w) = \sigma(\langle w, x \rangle) = \sigma\left(\sum_{j=1}^n w_j f_j(x) - w_0\right)$$

$\sigma(z)$  — функция активации (например, sign),

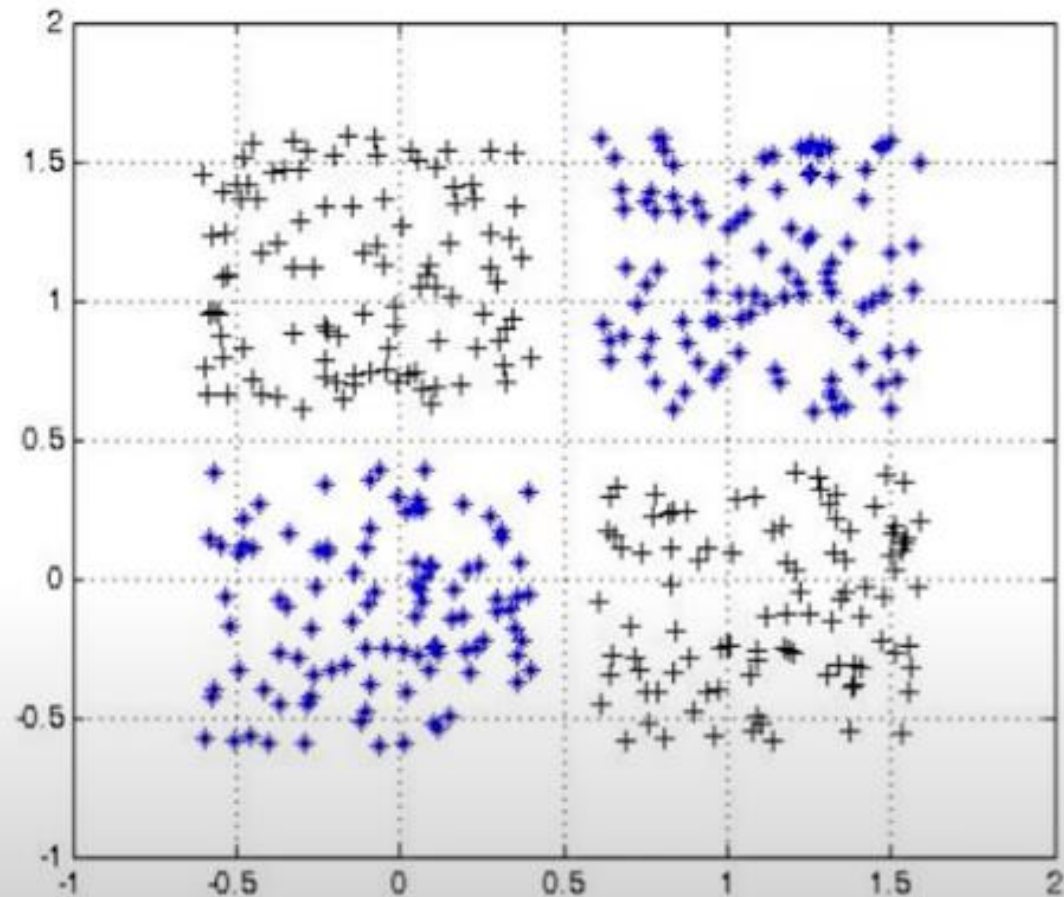
$w_j$  — весовые коэффициенты синаптических связей,

$w_0$  — порог активации,



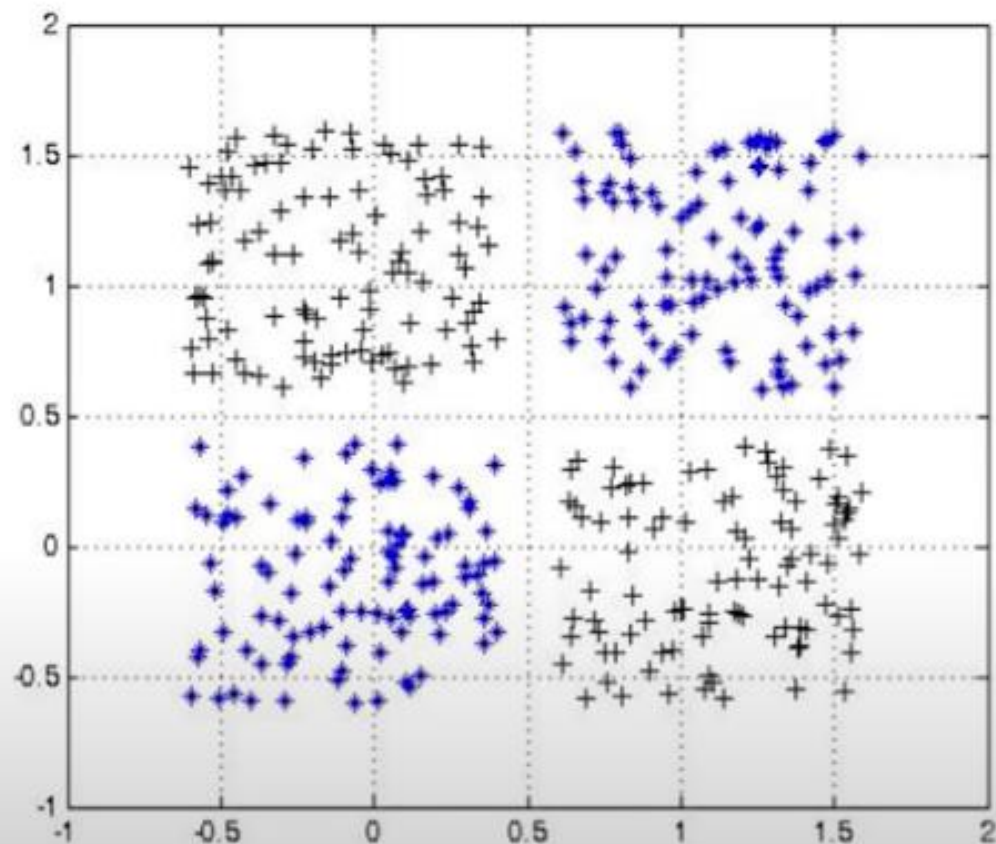
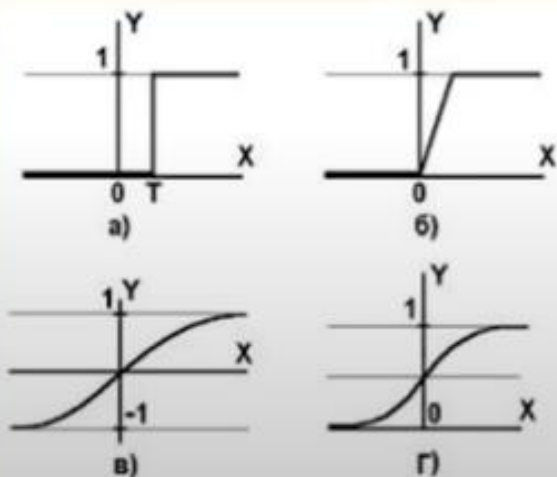
# ХОР-проблема

Как провести  
разделяющую  
гиперплоскость?



# ХОР-проблема

Как провести  
разделяющую  
гиперплоскость?



## Линейный классификатор как модель нейрона

Линейная модель нейрона МакКаллока-Питтса [1943]:

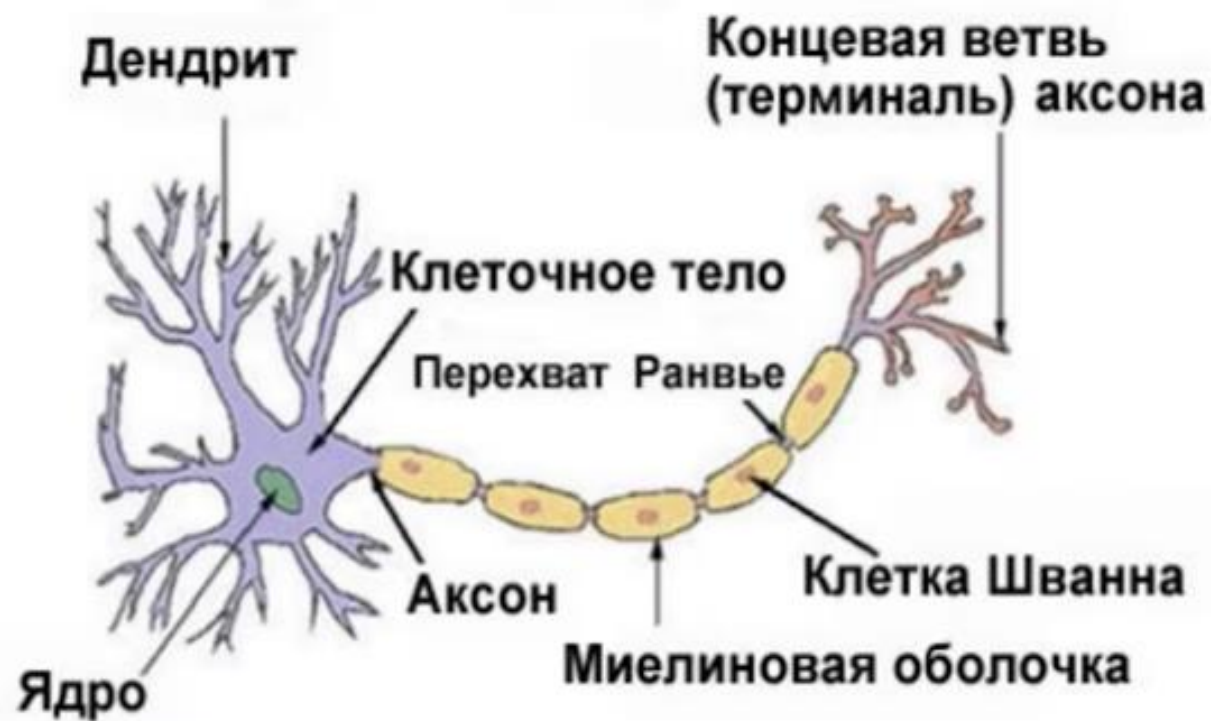
$$a(x, w) = \sigma(\langle w, x \rangle) = \sigma\left(\sum_{j=1}^n w_j f_j(x) - w_0\right)$$

$\sigma(z)$  — функция активации (например, sign),

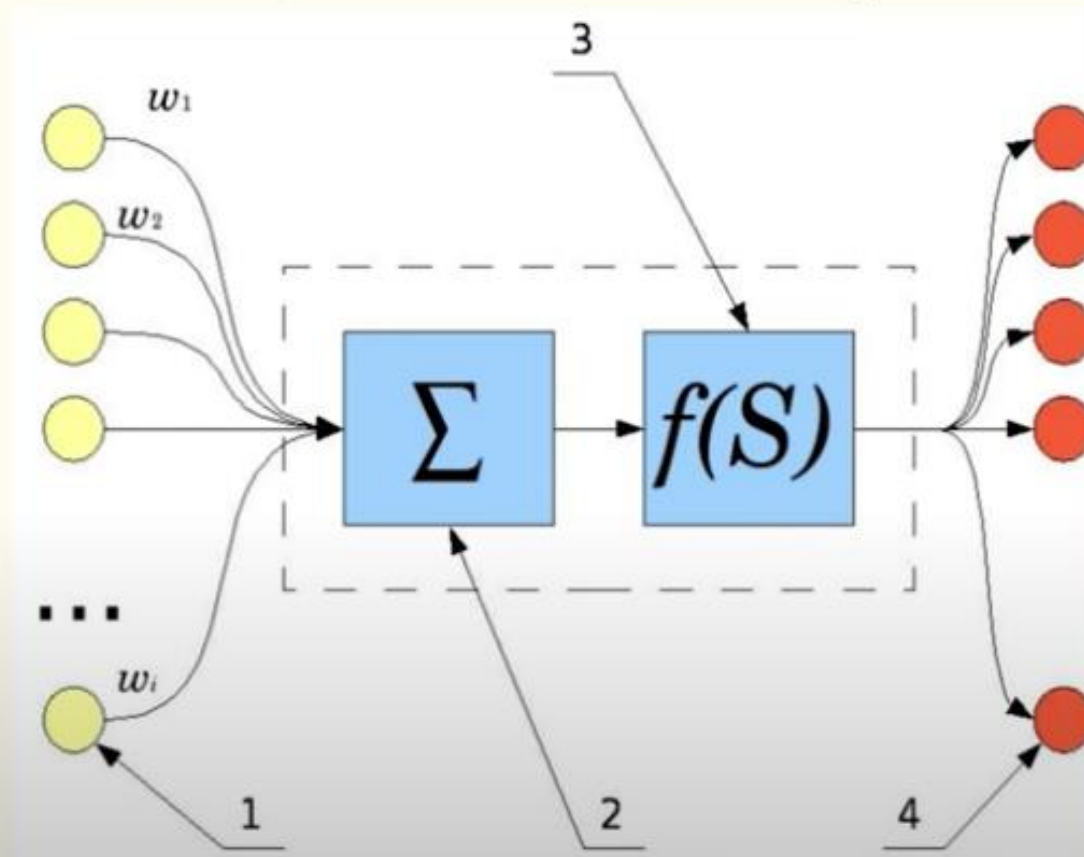
$w_j$  — весовые коэффициенты синаптических связей,

$w_0$  — порог активации,

# Биологический нейрон человека



# Искусственный нейрон



# ХОР-проблема

Как провести  
разделяющую  
гиперплоскость?

