ML

НА ПАЛЬЦАХ

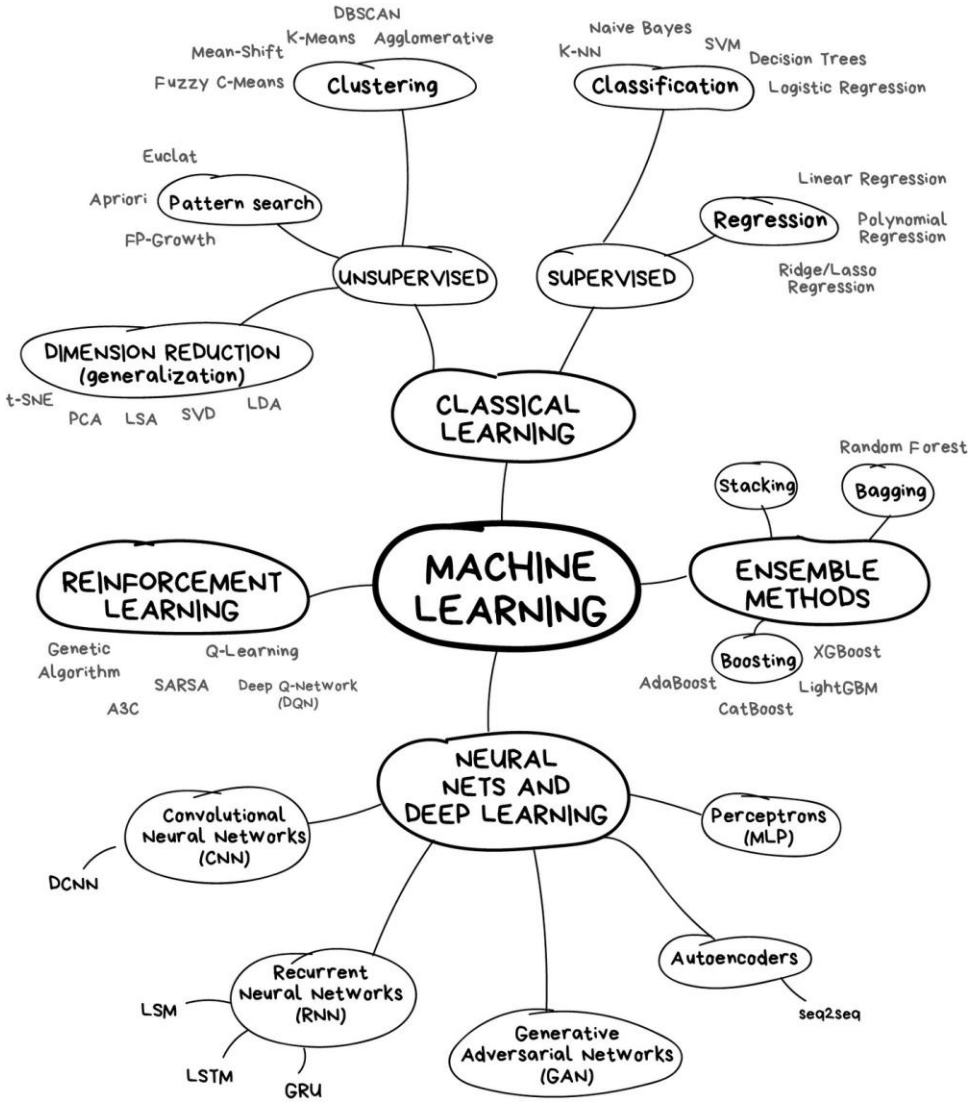
Петров Никита

ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ТЕХ, КТО ХОЧЕТ ВКАТИТЬСЯ В ТЕМАТИКУ И ПОНЯТЬ ЧТО К ЧЕМУ

ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ

Построение моделей на основе входных данных, способных «обобщать» эти входные данные

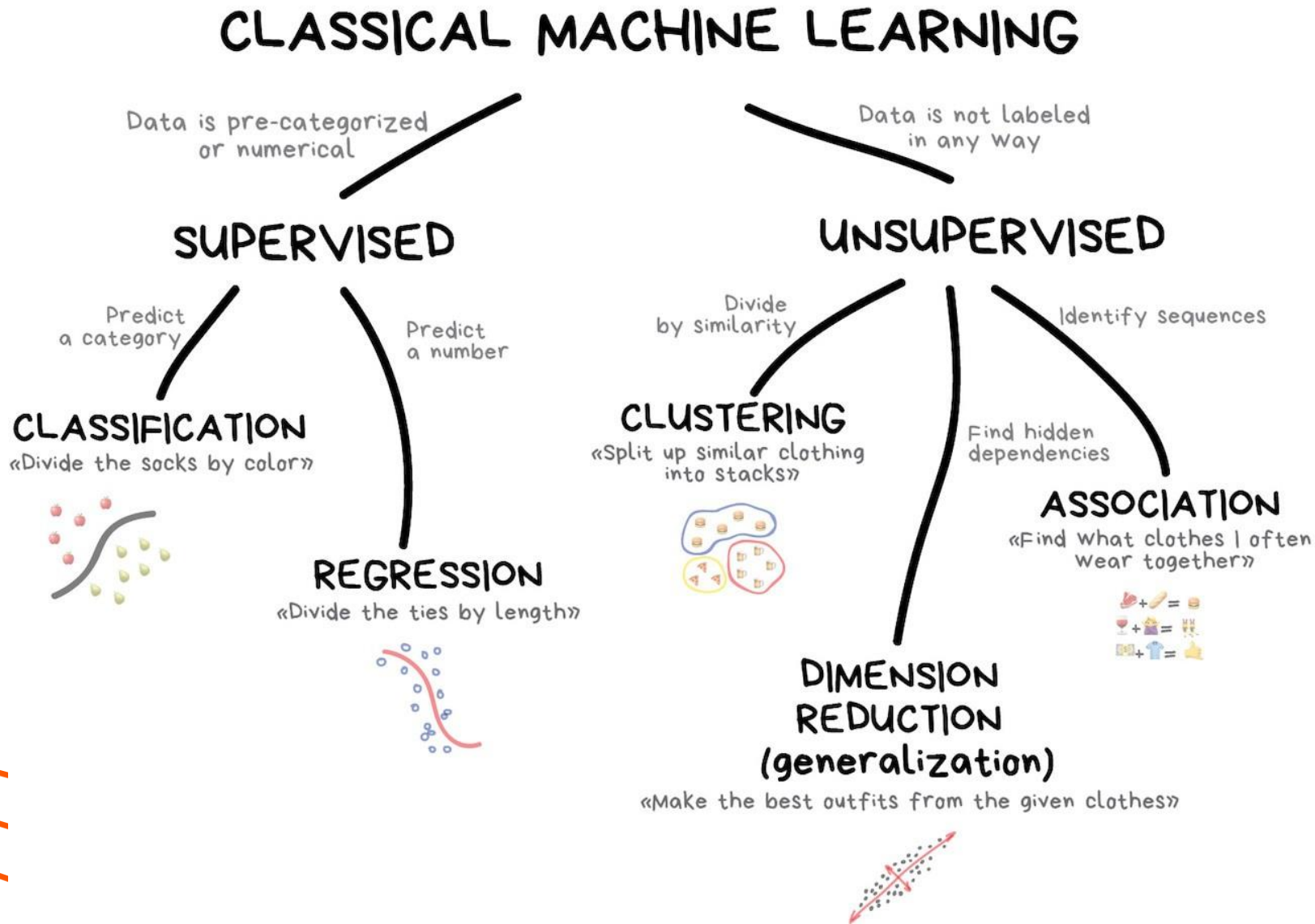
- Задача регрессии
- Задача классификации
- Задача кластеризации
- Другие



ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ

- Линейные модели
- Вероятностные модели
- Метрические методы
- Решающие деревья
- Нейронные сети

ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ



ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{in} \end{bmatrix} = [x_{i1}, x_{i2}, \dots, x_{in}]^T \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_l \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ . & . & . & . \\ x_{l1} & x_{l2} & \dots & x_{ln} \end{bmatrix}$$

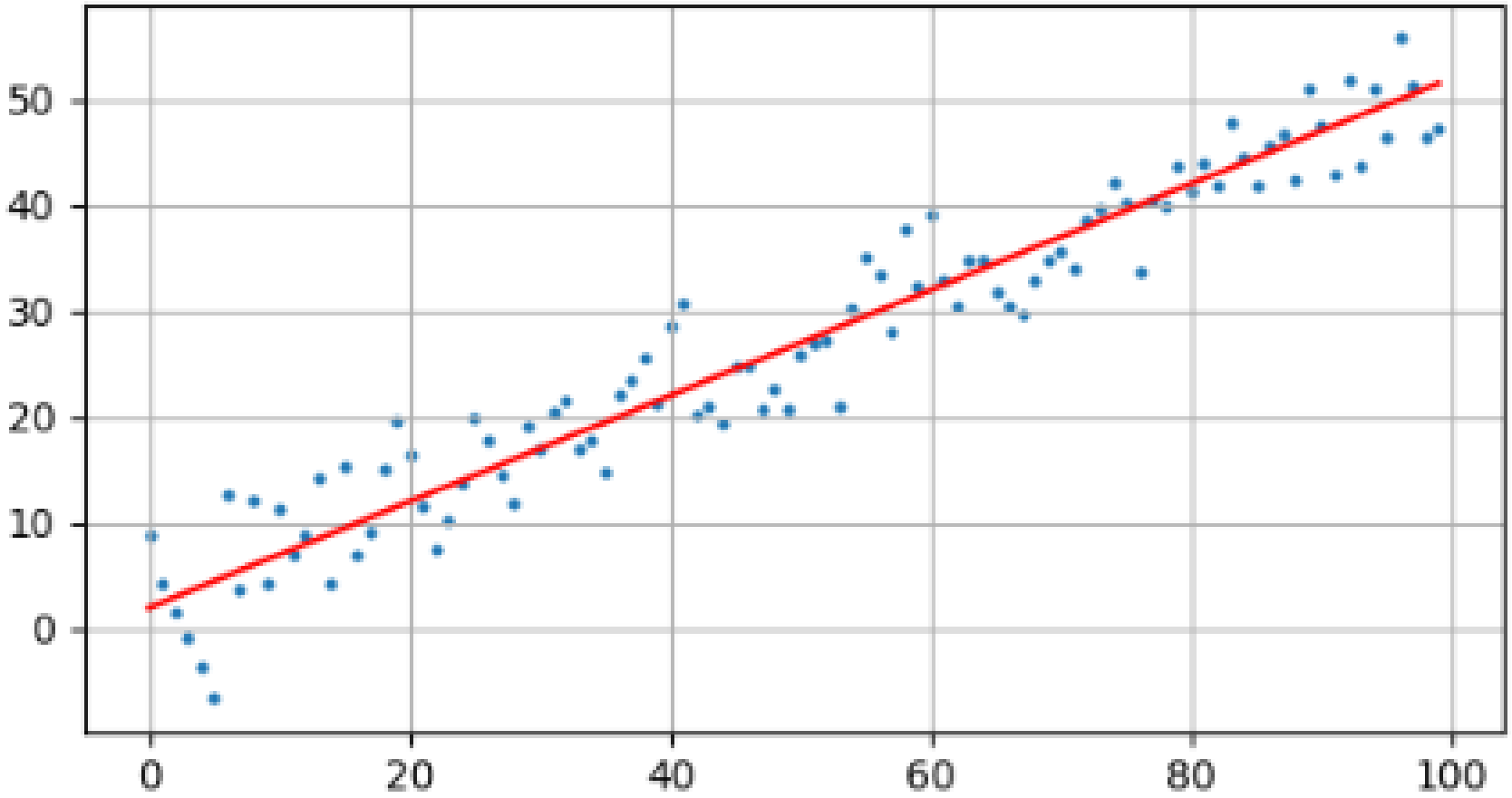
$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = [y_1, y_2, \dots, y_m]^T$$

ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ

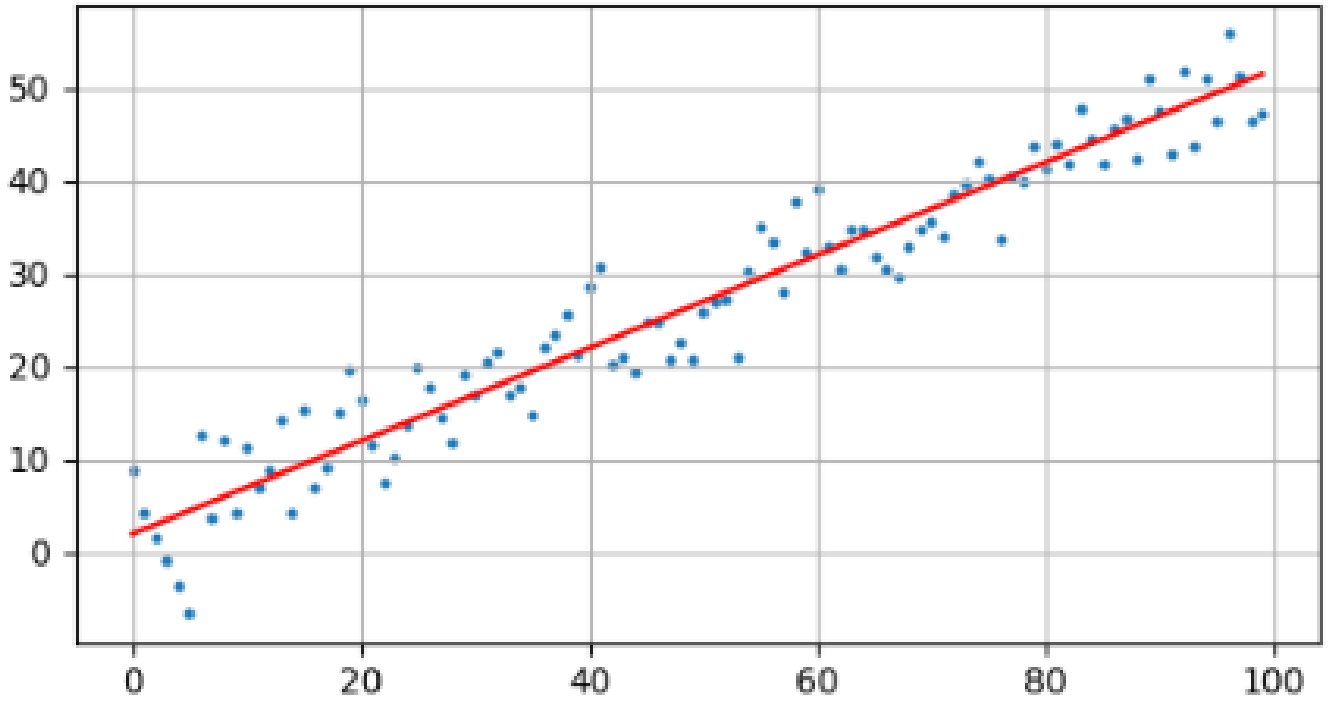
	v.id	on road old	on road now	years	km	rating	condition	economy	top speed	hp	torque	current price
0	1	535651	798186	3	78945	1	2	14	177	73	123	351318.0
1	2	591911	861056	6	117220	5	9	9	148	74	95	285001.5
2	3	686990	770762	2	132538	2	8	15	181	53	97	215386.0
3	4	573999	722381	4	101065	4	3	11	197	54	116	244295.5
4	5	691388	811335	6	61559	3	9	12	160	53	105	531114.5
...	...	...	...	...	...	...	...	...	...	...	...	...
995	996	633238	743850	5	125092	1	6	11	171	95	97	190744.0
996	997	599626	848195	4	83370	2	9	14	161	101	120	419748.0
997	998	646344	842733	7	86722	1	8	9	196	113	89	405871.0
998	999	535559	732439	2	140478	4	5	9	184	112	128	74398.0
999	1000	590105	779743	5	67295	4	2	8	199	99	96	414938.5

1000 rows × 12 columns

ПРОСТЕЙШАЯ ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ – ЛИНЕЙНАЯ РЕГРЕССИЯ



ФУНКЦИЯ ПОТЕРЬ

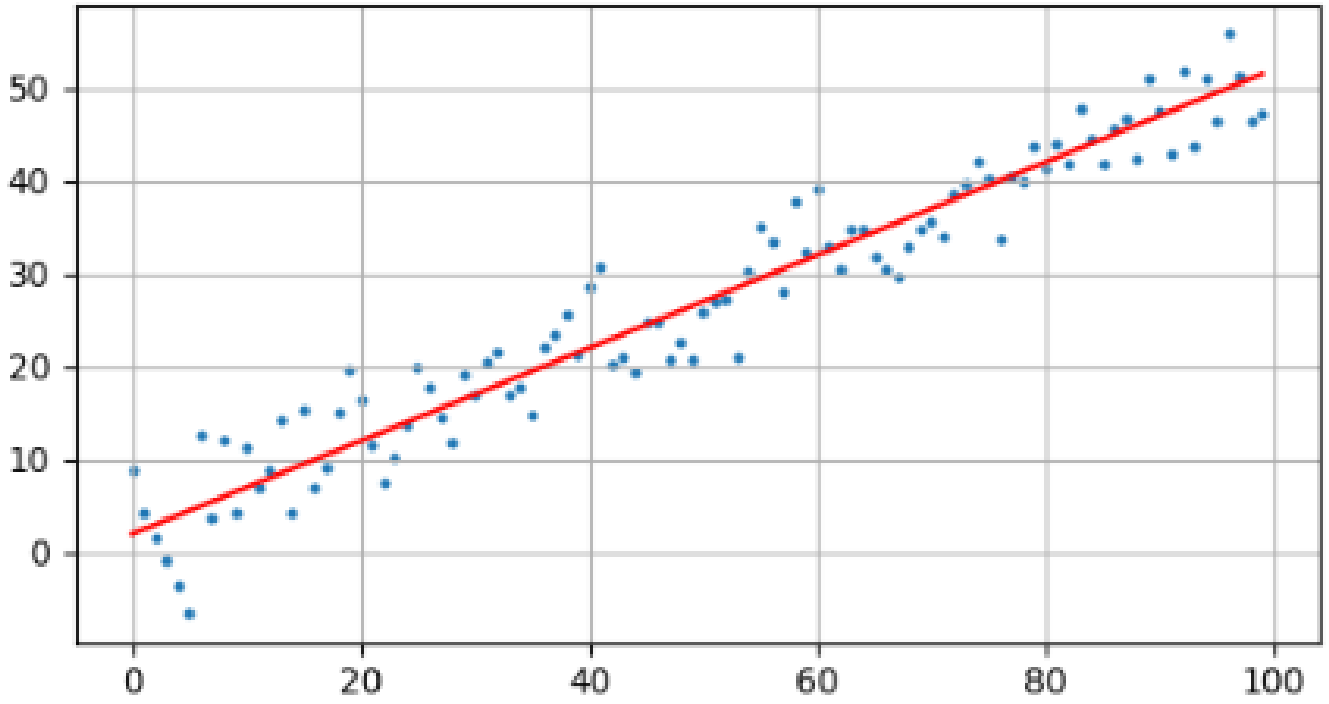


- $L(a,x) = |a(x) - y(x)|$ - абсолютная ошибка;
- $L(a,x) = (a(x) - y(x))^2$ - квадратичная ошибка.

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} L(a, x_i)$$

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} (a(x_i) - y(x_i))^2 = \frac{1}{l}\sum_{i=1}^{l} (kx_i + b - y(x_i))^2$$

ФУНКЦИЯ ПОТЕРЬ



- $L(a,x) = |a(x) - y(x)|$ - абсолютная ошибка;
- $L(a,x) = (a(x) - y(x))^2$ - квадратичная ошибка.

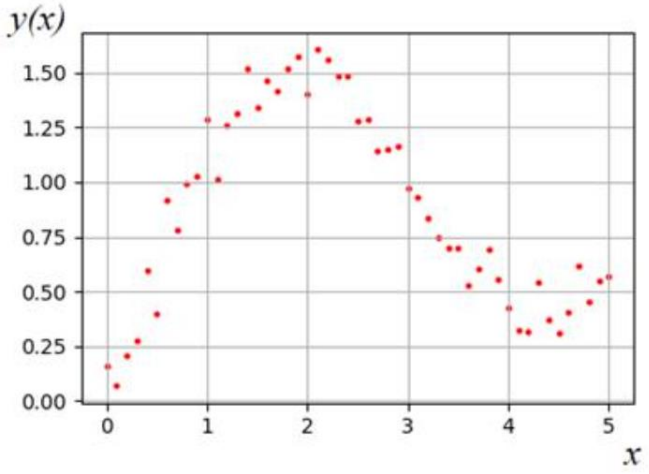
$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} L(a, x_i)$$

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} (a(x_i) - y(x_i))^2 = \frac{1}{l}\sum_{i=1}^{l} (kx_i + b - y(x_i))^2$$

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

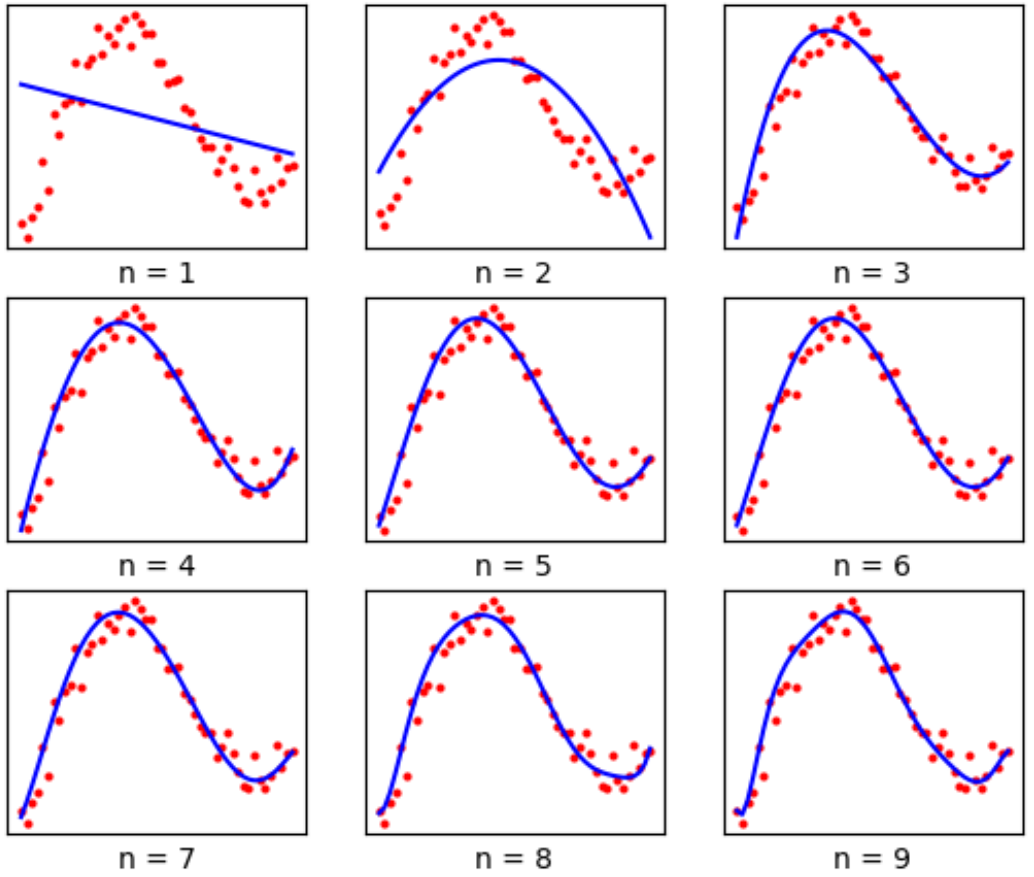
НЕМНОГО ПРО ПРИЗНАКИ В ЛИНЕЙНЫХ МОДЕЛЯХ

$$y(x) = \sin(x) + 0,3x + \varepsilon(x), \quad x \in [0;5]$$

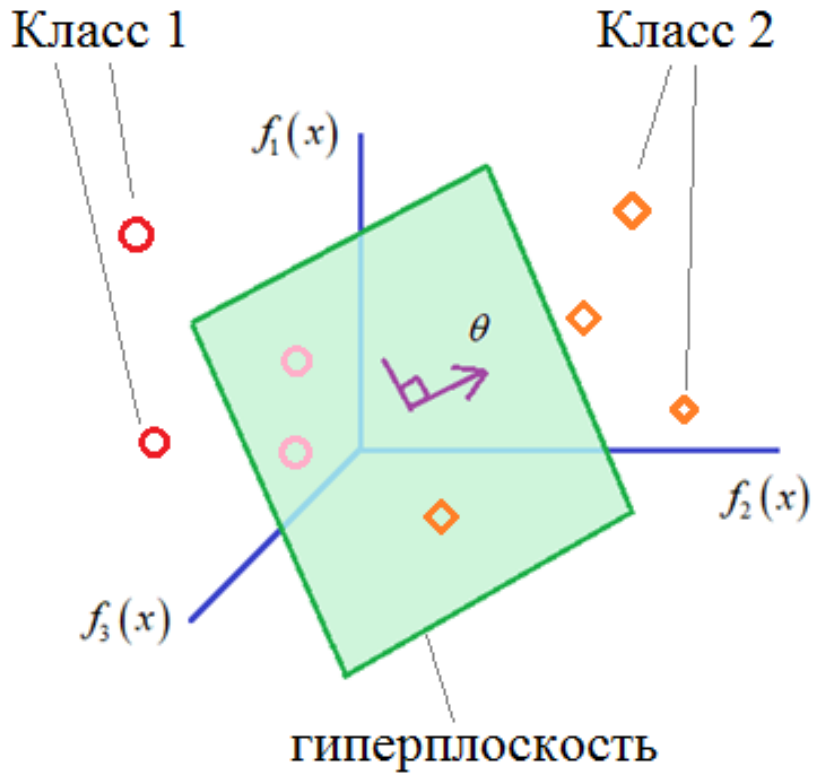


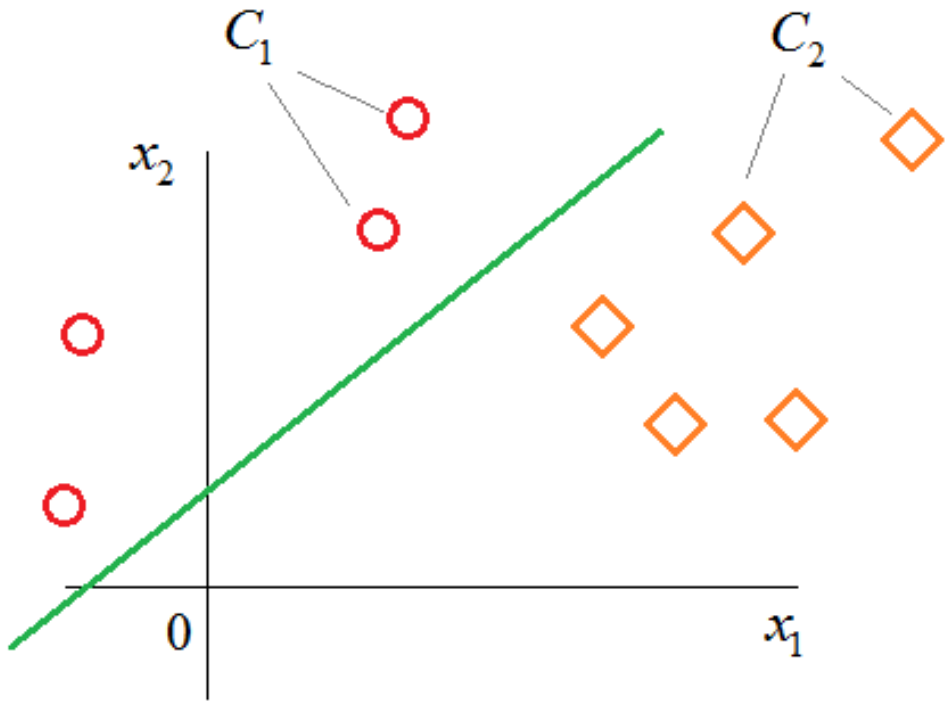
$$f_0(x) = 1$$
$$f_1(x) = x$$
$$f_2(x) = x^2$$
$$\ldots$$
$$f_n(x) = x^n$$

$$a(x) = 1 \cdot \theta_1 + x \cdot \theta_2 + x^2 \cdot \theta_3 + \ldots + x^n \cdot \theta_{n+1}$$



БИНАРНАЯ КЛАССИФИКАЦИЯ

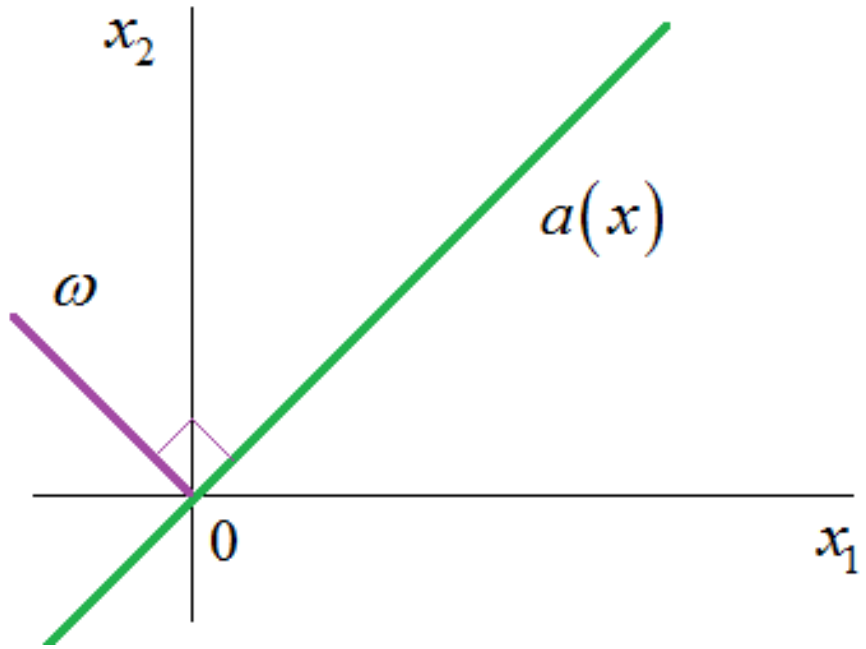


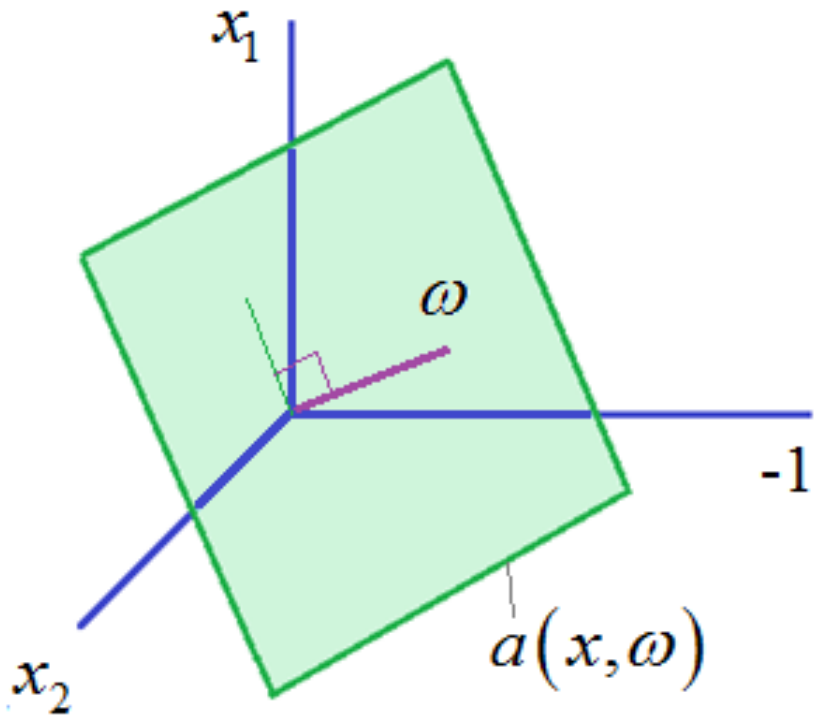


$$x_i = [x_{i1}, x_{i2}]^T, \quad i = 1, 2, \ldots$$

$$y_i = \begin{cases} +1, & \text{если } x_i \in C_1 \\ -1, & \text{если } x_i \in C_2 \end{cases}$$

БИНАРНАЯ КЛАССИФИКАЦИЯ





НЕМНОГО О ВЕРОЯТНОСТНОМ ПОДХОДЕ

Доход	Наличие жилья	Наличие работы	Число детей	Возврат кредита
10 000	1	1	2	1
10 000	1	1	2	-1
10 000	1	1	2	-1
10 000	1	1	2	1
10 000	1	1	2	-1
10 000	1	1	2	1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	1

$$a(x,\omega) = sign(\langle \omega, x \rangle)$$

$$P(y \mid x, \omega)$$

$$x = [10\ 000,\ 1, 1, 2]^T$$

$$P(y = +1 \mid x, \omega) = \frac{3}{6} = 0,5$$

$$P(y = -1 \mid x, \omega) = 1 - P(y = +1 \mid x, \omega) = 1 - 0,5 = 0,5$$

$$\omega_* = \arg\max_{\omega} P(y \mid x, \omega)$$

$$P(y_1, \ldots, y_l \mid x_1, \ldots, x_l, \omega_1, \ldots, \omega_n) = \prod_{i=1}^{l} P(y_i \mid x_i, \omega)$$

$$\log L(\omega) = \sum_{i=1}^{l} \log P(y_i \mid x_i, \omega) \to \max_{\omega}$$

$$Q(a, X^l) = \sum_{i=1}^{l} L_i(x_i, \omega) = -\sum_{i=1}^{l} \log P(y_i \mid x_i, \omega) \to \min_{\omega}$$

НЕМНОГО О ВЕРОЯТНОСТНОМ ПОДХОДЕ

Доход	Наличие жилья	Наличие работы	Число детей	Возврат кредита
10 000	1	1	2	1
10 000	1	1	2	-1
10 000	1	1	2	-1
10 000	1	1	2	1
10 000	1	1	2	-1
10 000	1	1	2	1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	-1
15 000	1	0	3	1

$$a(x,\omega) = sign(\langle \omega, x \rangle)$$

$$P(y \mid x, \omega)$$

$$x = [10\ 000,\ 1, 1, 2]^T$$

$$P(y = +1 \mid x, \omega) = \frac{3}{6} = 0,5$$

$$P(y = -1 \mid x, \omega) = 1 - P(y = +1 \mid x, \omega) = 1 - 0,5 = 0,5$$

$$\omega_* = \arg\max_{\omega} P(y \mid x, \omega)$$

$$P(y_1, \ldots, y_l \mid x_1, \ldots, x_l, \omega_1, \ldots, \omega_n) = \prod_{i=1}^{l} P(y_i \mid x_i, \omega)$$

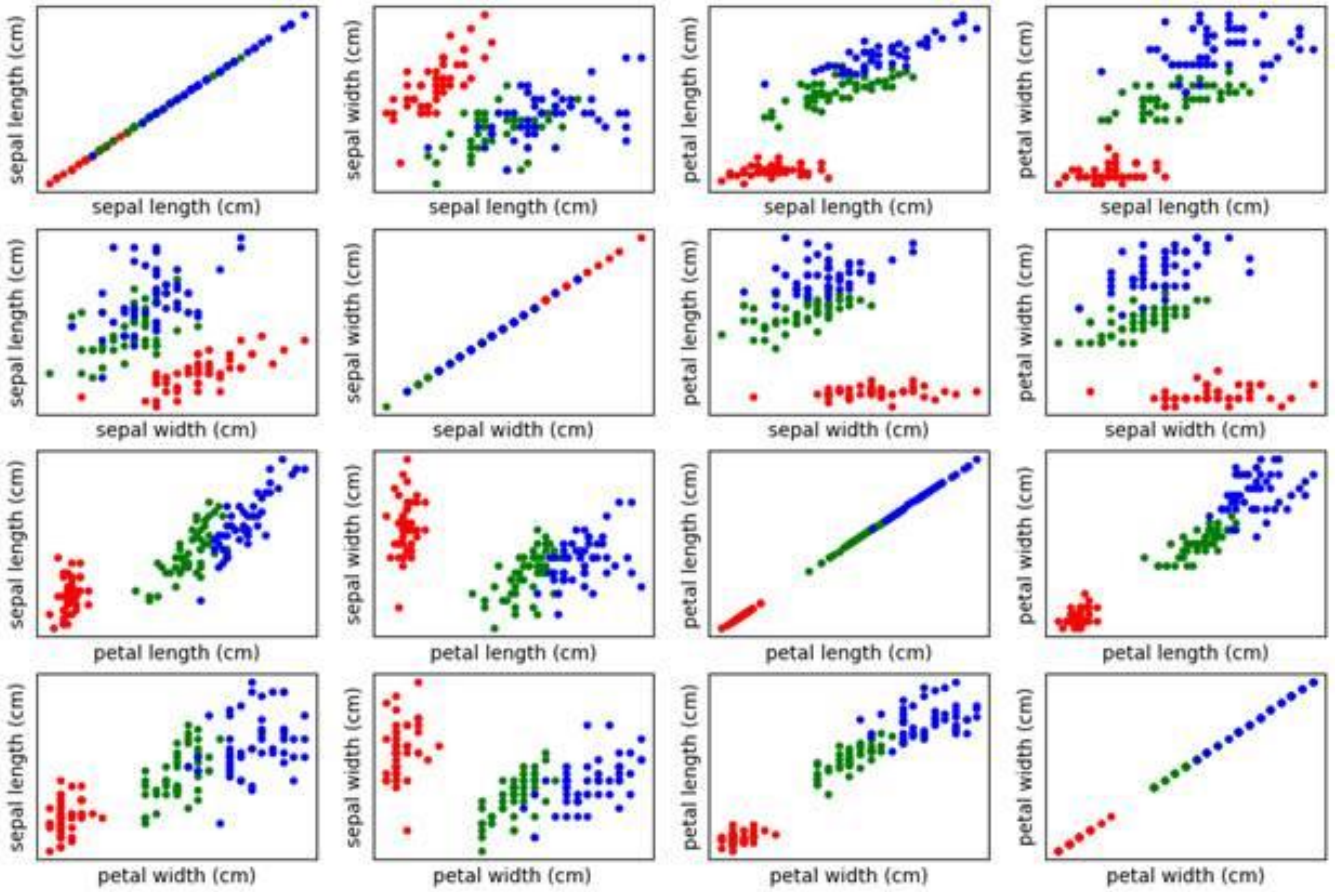
$$\log L(\omega) = \sum_{i=1}^{l} \log P(y_i \mid x_i, \omega) \to \max_{\omega}$$

$$Q(a, X^l) = \sum_{i=1}^{l} L_i(x_i, \omega) = -\sum_{i=1}^{l} \log P(y_i \mid x_i, \omega) \to \min_{\omega}$$

О МЕТРИЧЕСКИХ МЕТОДАХ

- ирис щетинистый (iris setosa);
- ирис виргинский (iris virginica);
- ирис разноцветный (iris versicolor).

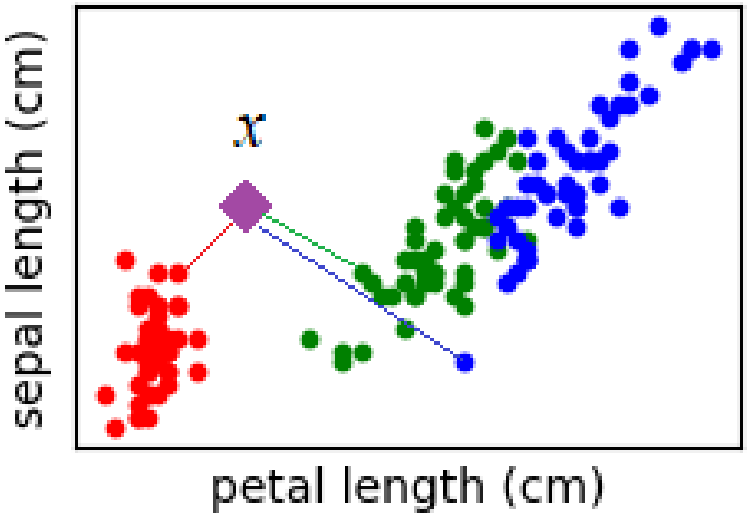
- длина наружной доли околоцветника (sepal length);
- ширина наружной доли околоцветника (sepal width);
- длина внутренней доли околоцветника (petal length);
- ширина внутренней доли околоцветника (petal width).

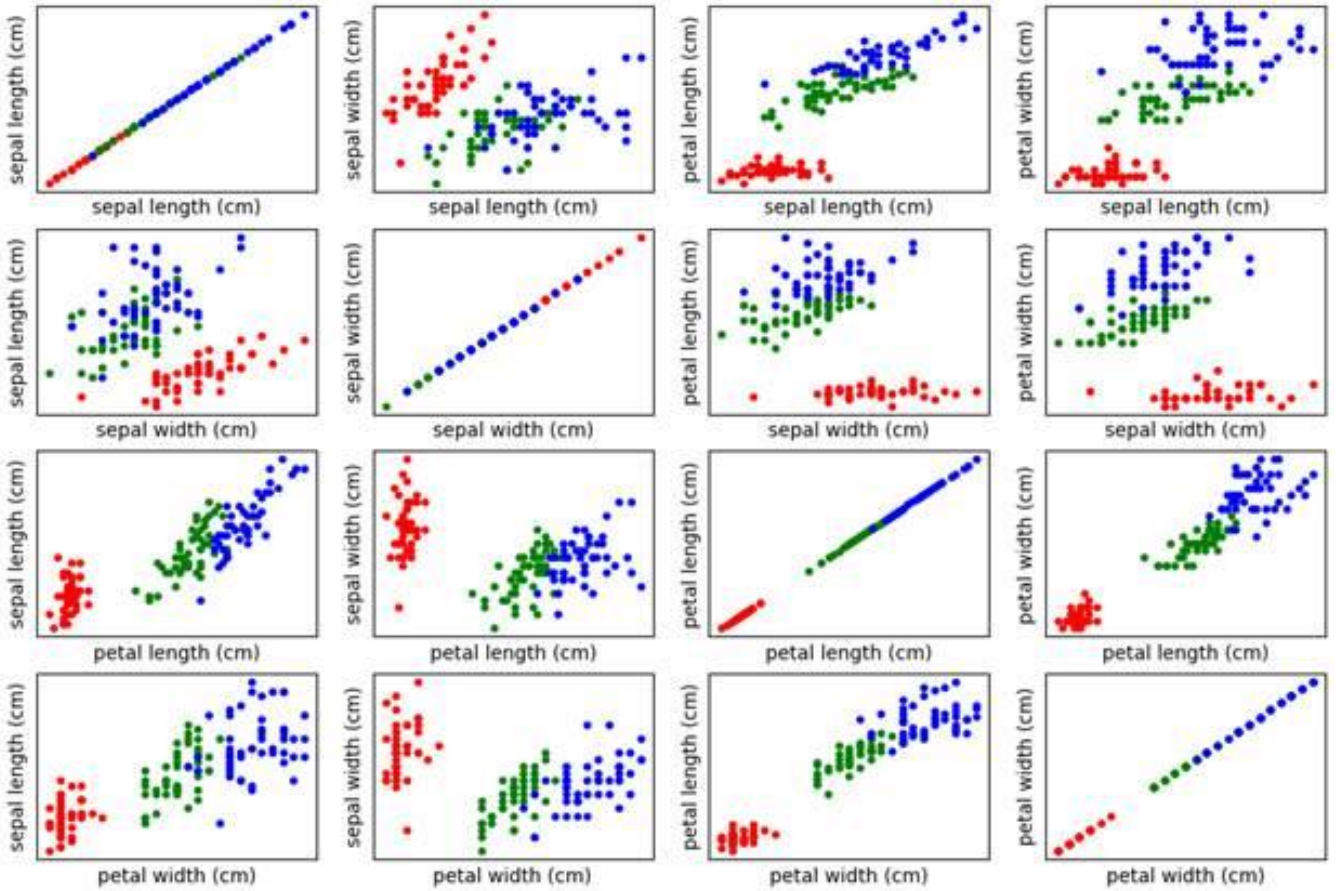


О МЕТРИЧЕСКИХ МЕТОДАХ

- ирис щетинистый (iris setosa);
- ирис виргинский (iris virginica);
- ирис разноцветный (iris versicolor).

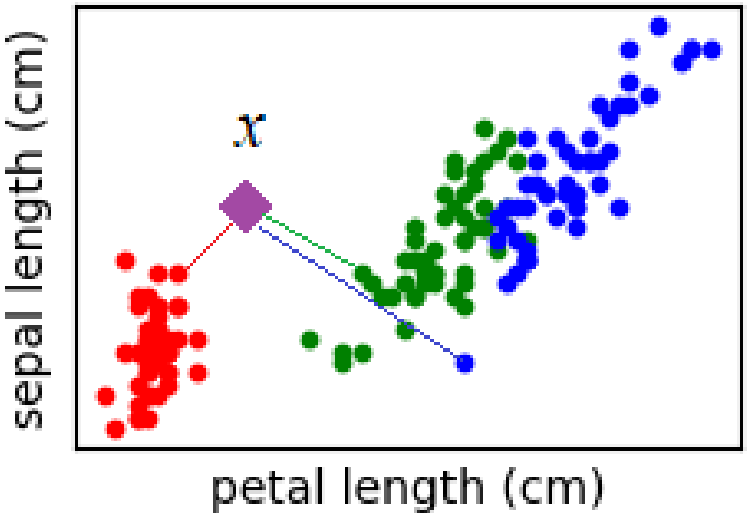
- длина наружной доли околоцветника (sepal length);
- ширина наружной доли околоцветника (sepal width);
- длина внутренней доли околоцветника (petal length);
- ширина внутренней доли околоцветника (petal width).

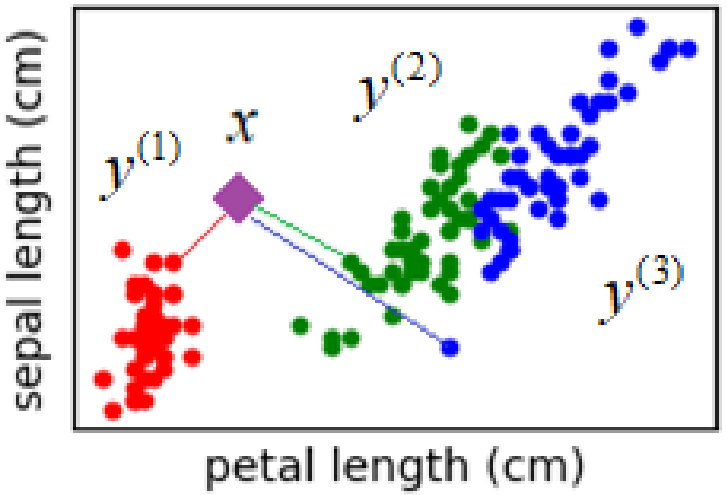


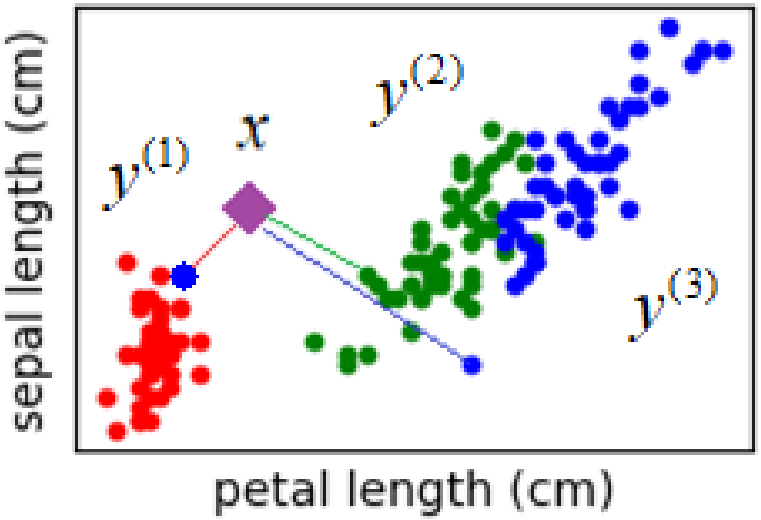


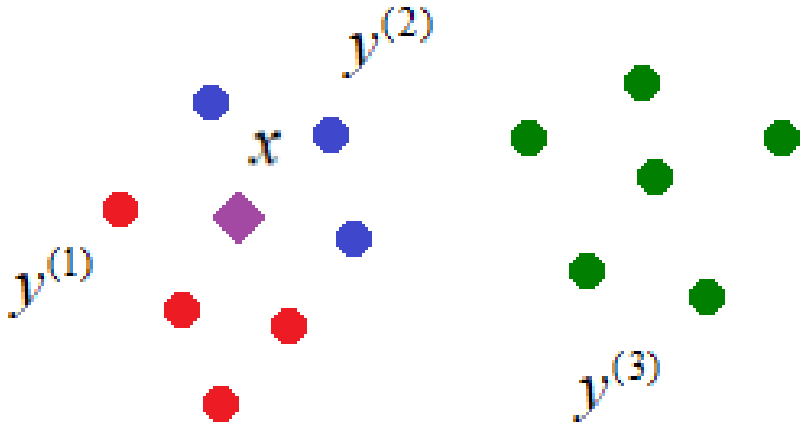
О МЕТРИЧЕСКИХ МЕТОДАХ

Метод ближайших соседей (kNN)

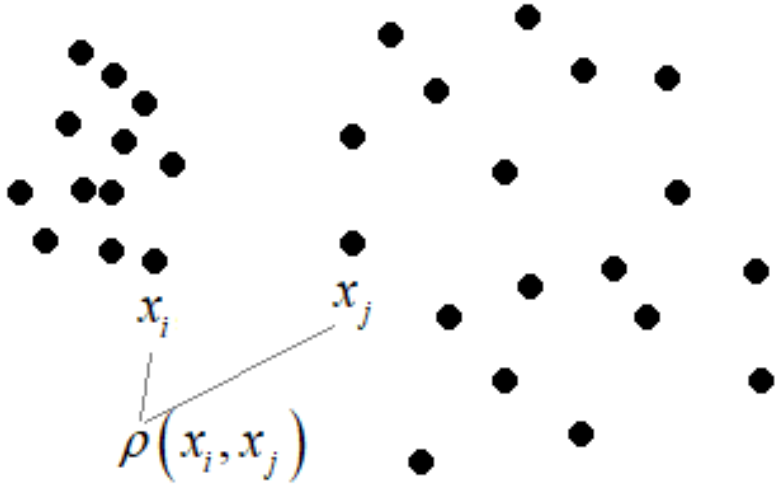








НЕМНОГО ПРО КЛАСТЕРИЗАЦИЮ



Внутриклассовые расстояния, в среднем, меньше, чем межклассовые

Ленточные кластеры

Кластеры с ярко выраженными центрами

Кластеры с перемычками

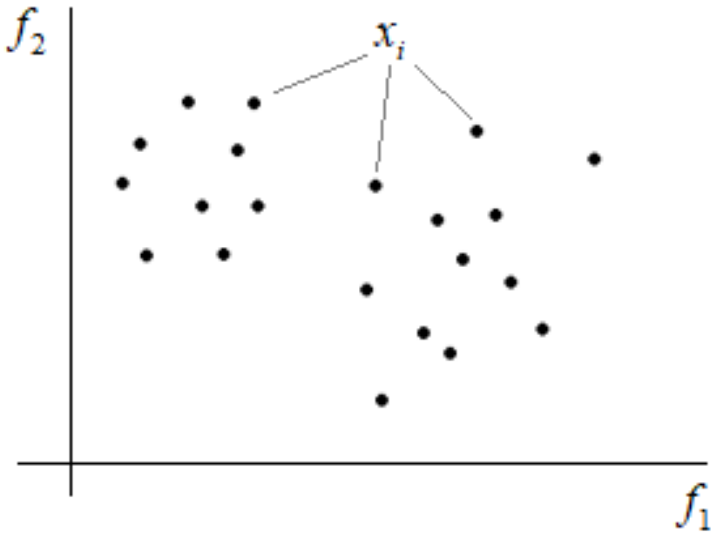
Кластеры на фоне шумов

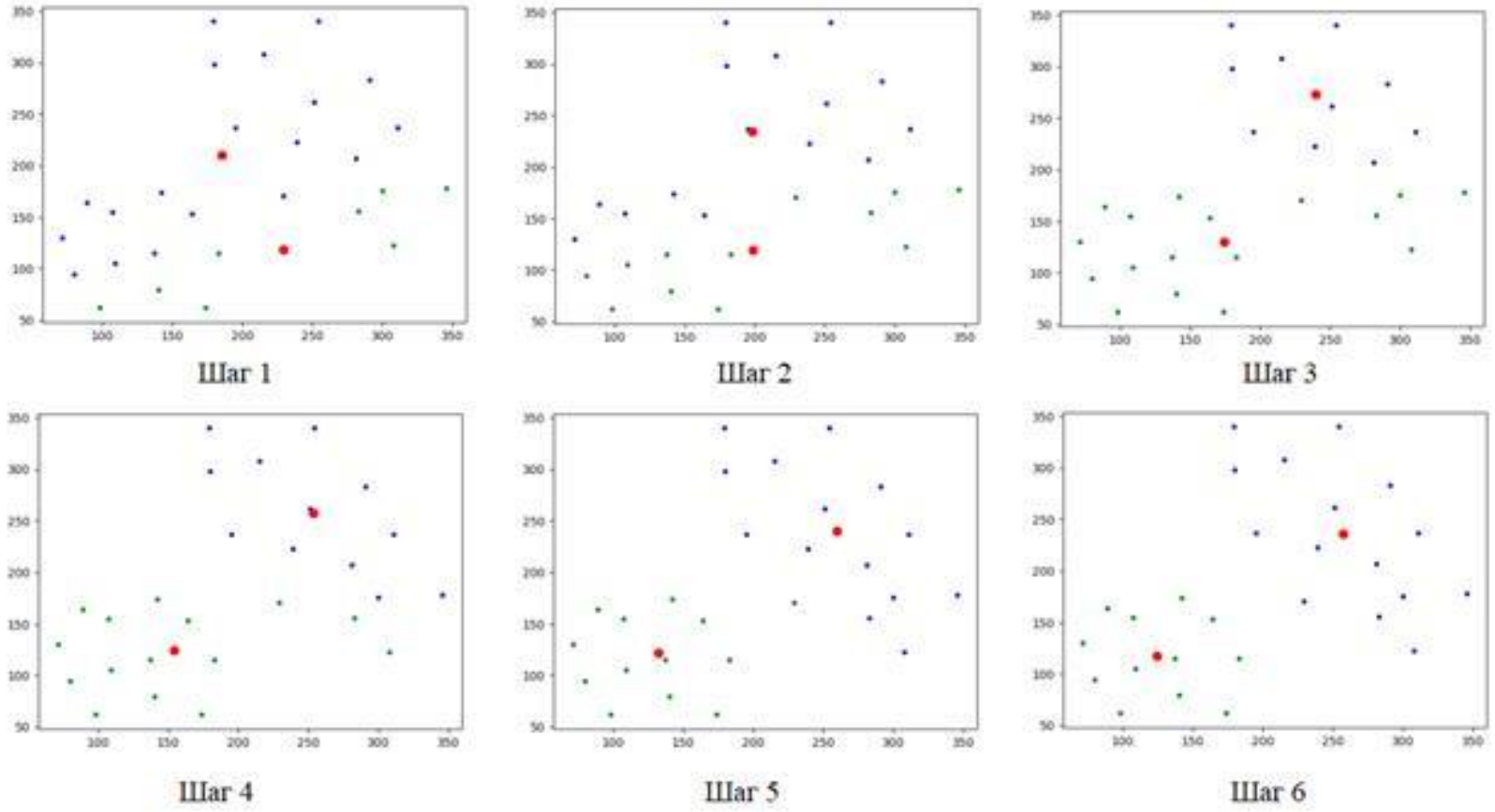
Кластеры в виде геометрических фигур

Один кластер

НЕМНОГО ПРО КЛАСТЕРИЗАЦИЮ

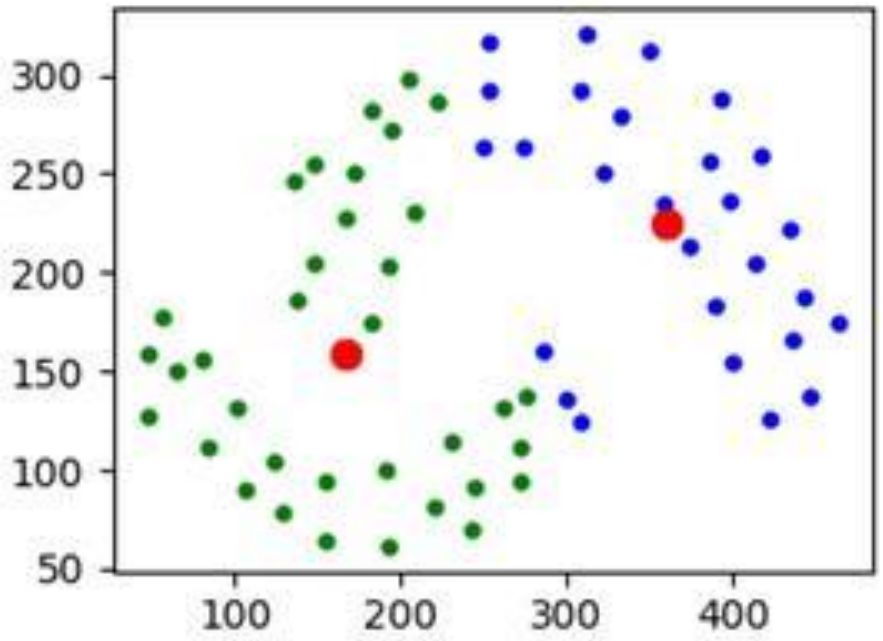
K-means

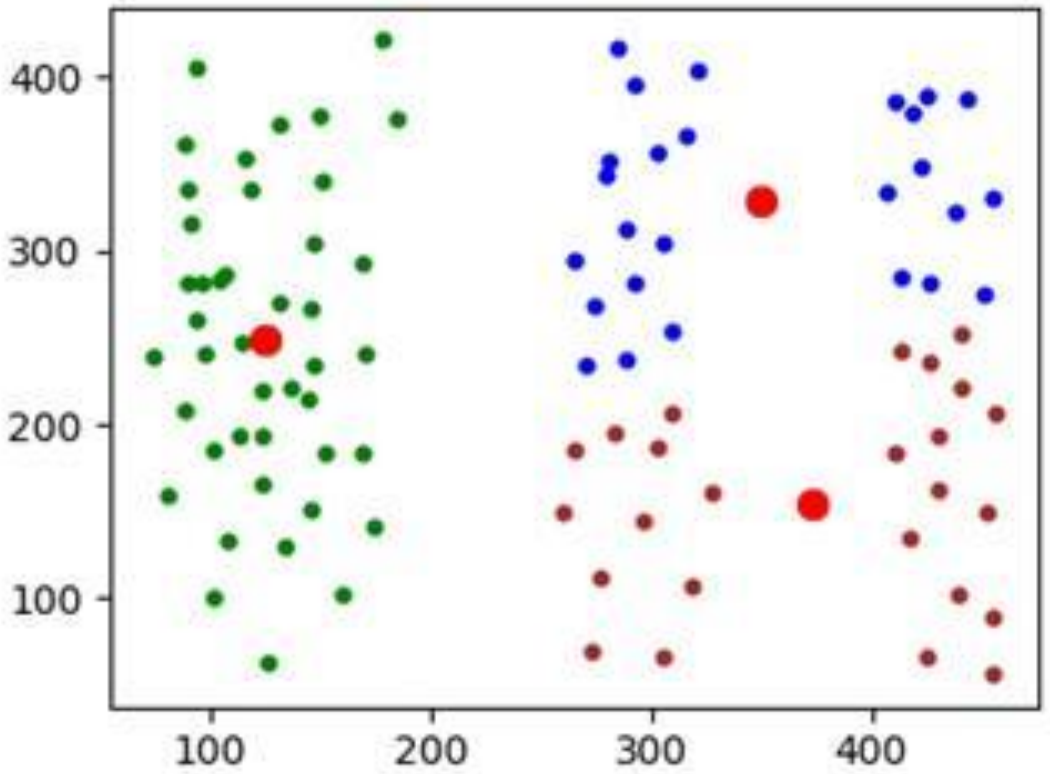




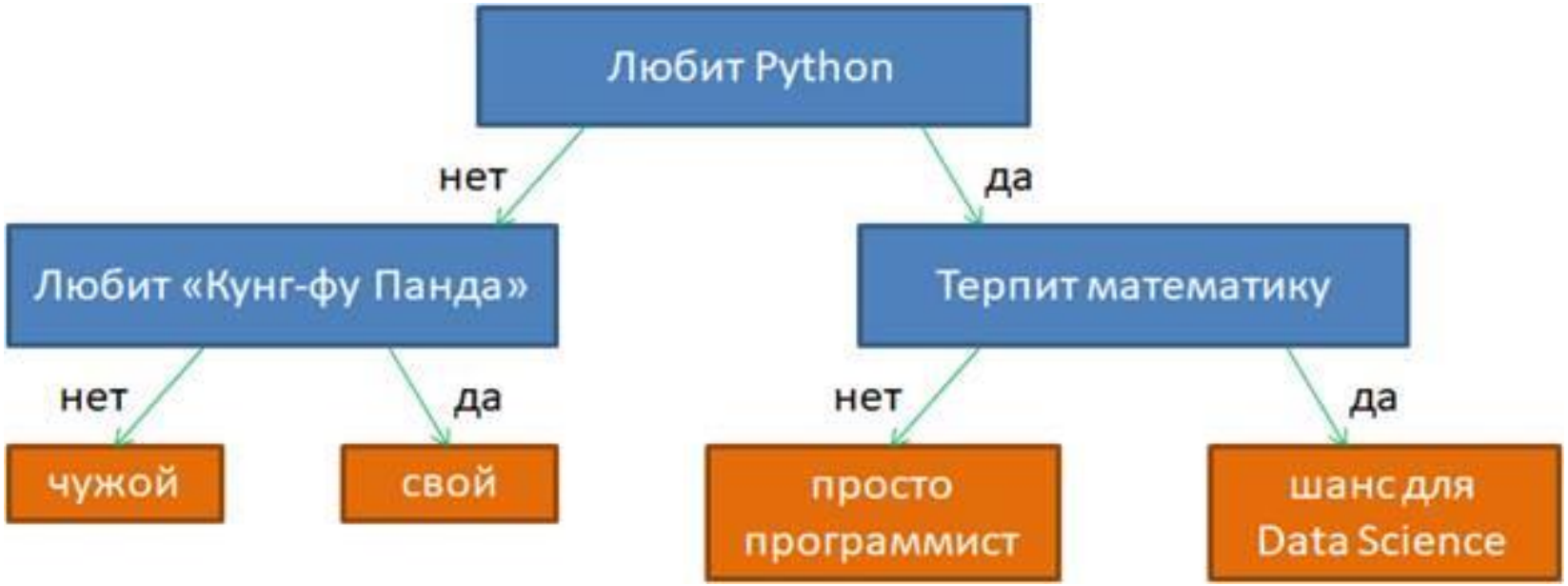
НЕМНОГО ПРО КЛАСТЕРИЗАЦИЮ

K-means





РЕШАЮЩИЕ ДЕРЕВЬЯ



$$f_1(x) = \begin{cases} 0 - \text{не любит Python} \\ 1 - \text{любит Python} \end{cases}$$

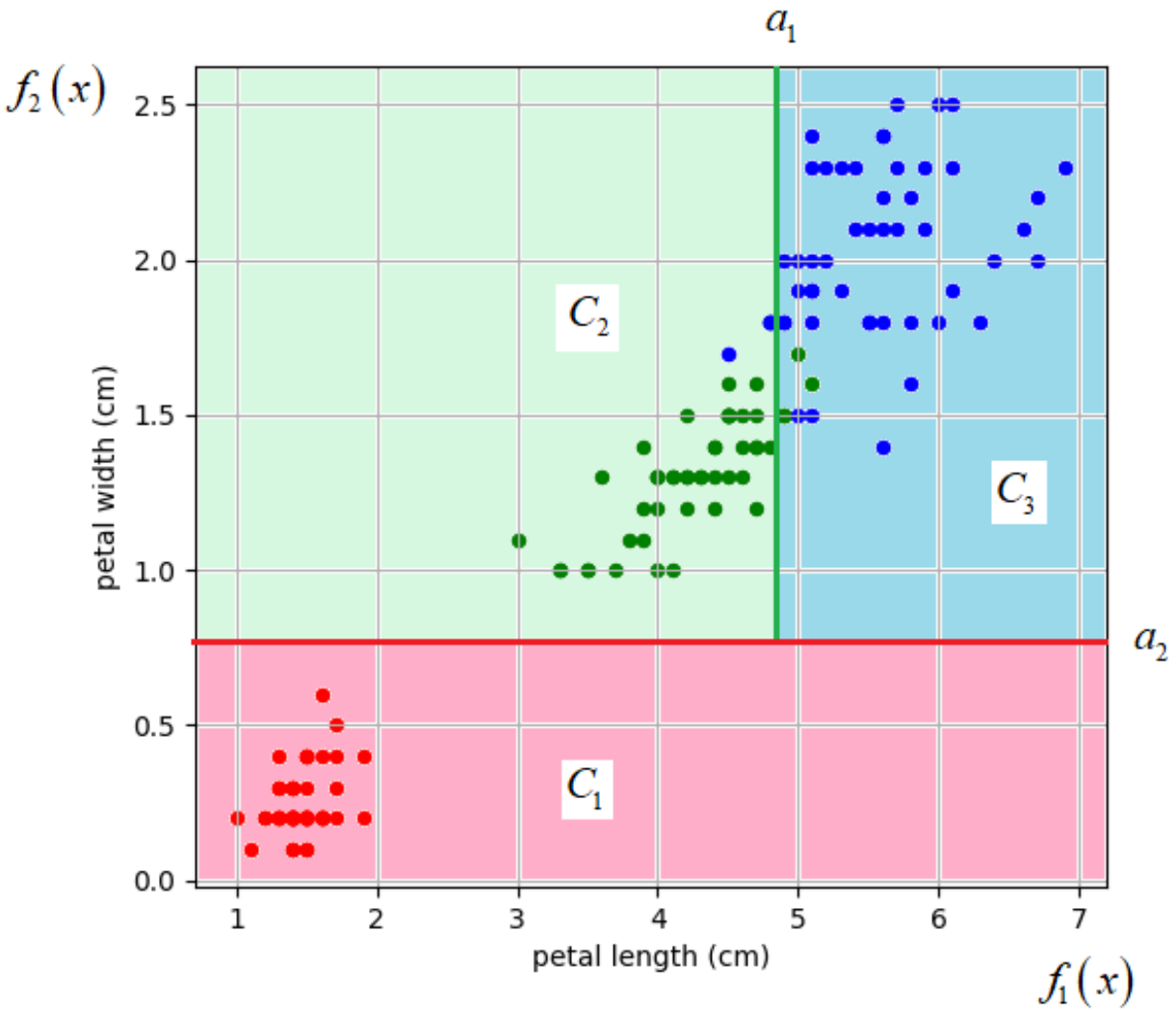
$$f_2(x) = \begin{cases} 0 - \text{не любит "Кунг-фу Панда"} \\ 1 - \text{любит "Кунг-фу Панда"} \end{cases}$$

$$f_3(x) = \begin{cases} 0 - \text{не терпит математику} \\ 1 - \text{терпит математику} \end{cases}$$

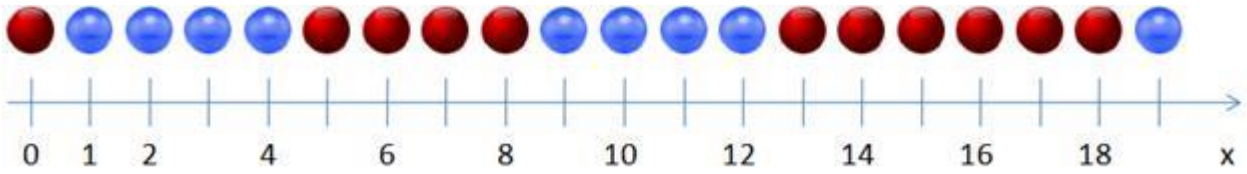
$$R(x) = f_1(x) \wedge \overline{f_3(x)} \rightarrow \text{просто программист}$$

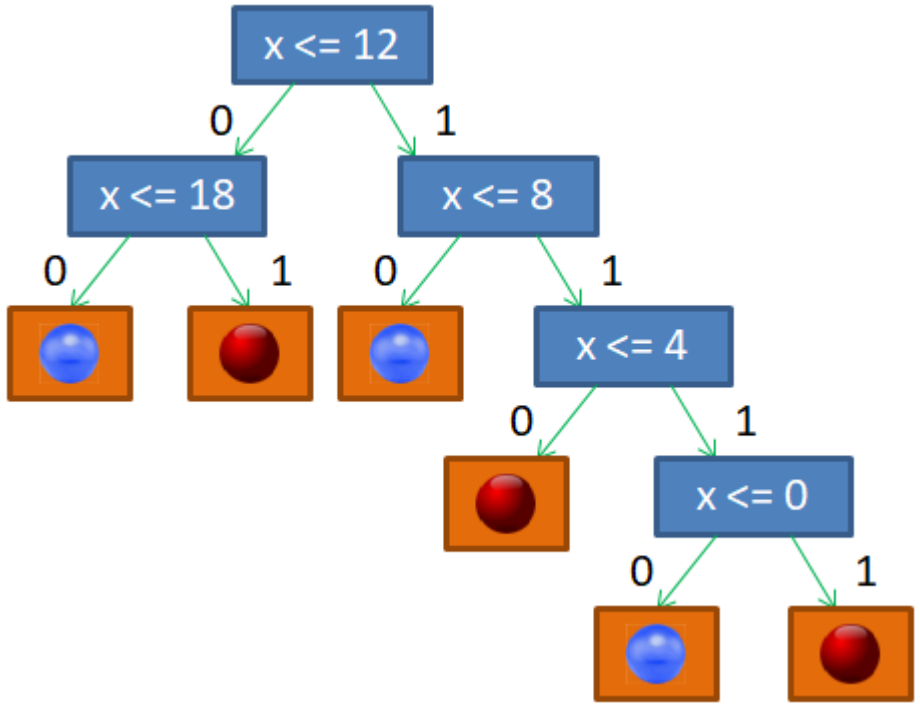
$$R(x) = \overline{f_1(x)} \wedge f_2(x) \rightarrow \text{свой}$$

РЕШАЮЩИЕ ДЕРЕВЬЯ

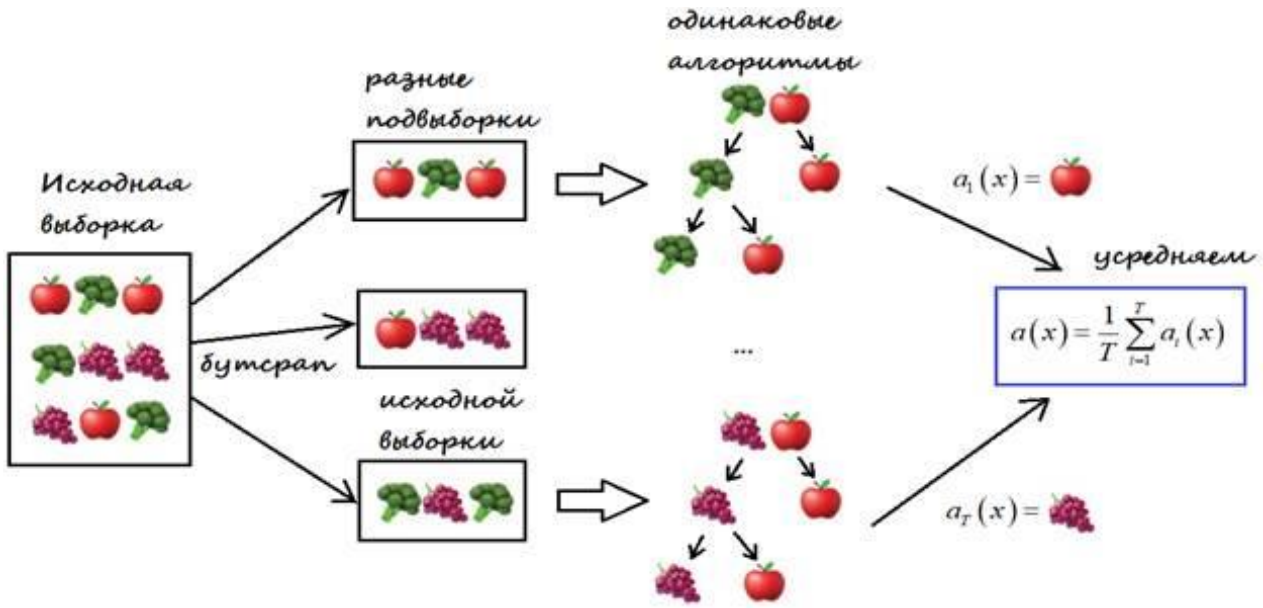


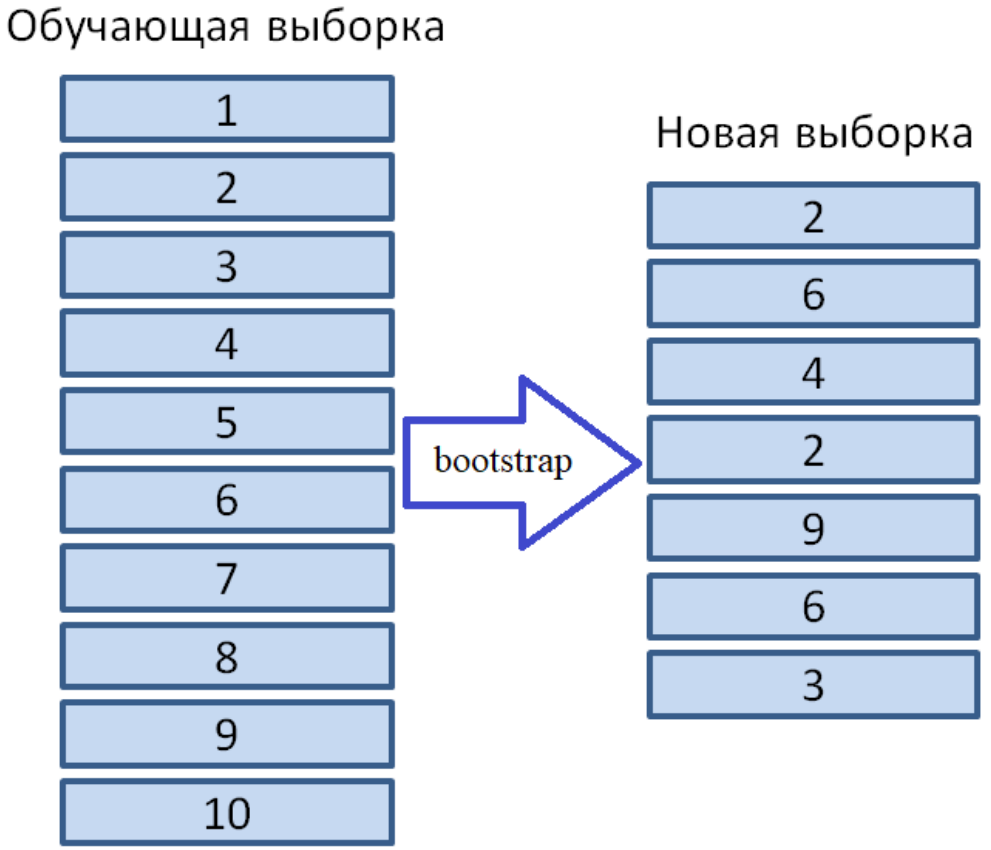
РЕШАЮЩИЕ ДЕРЕВЬЯ



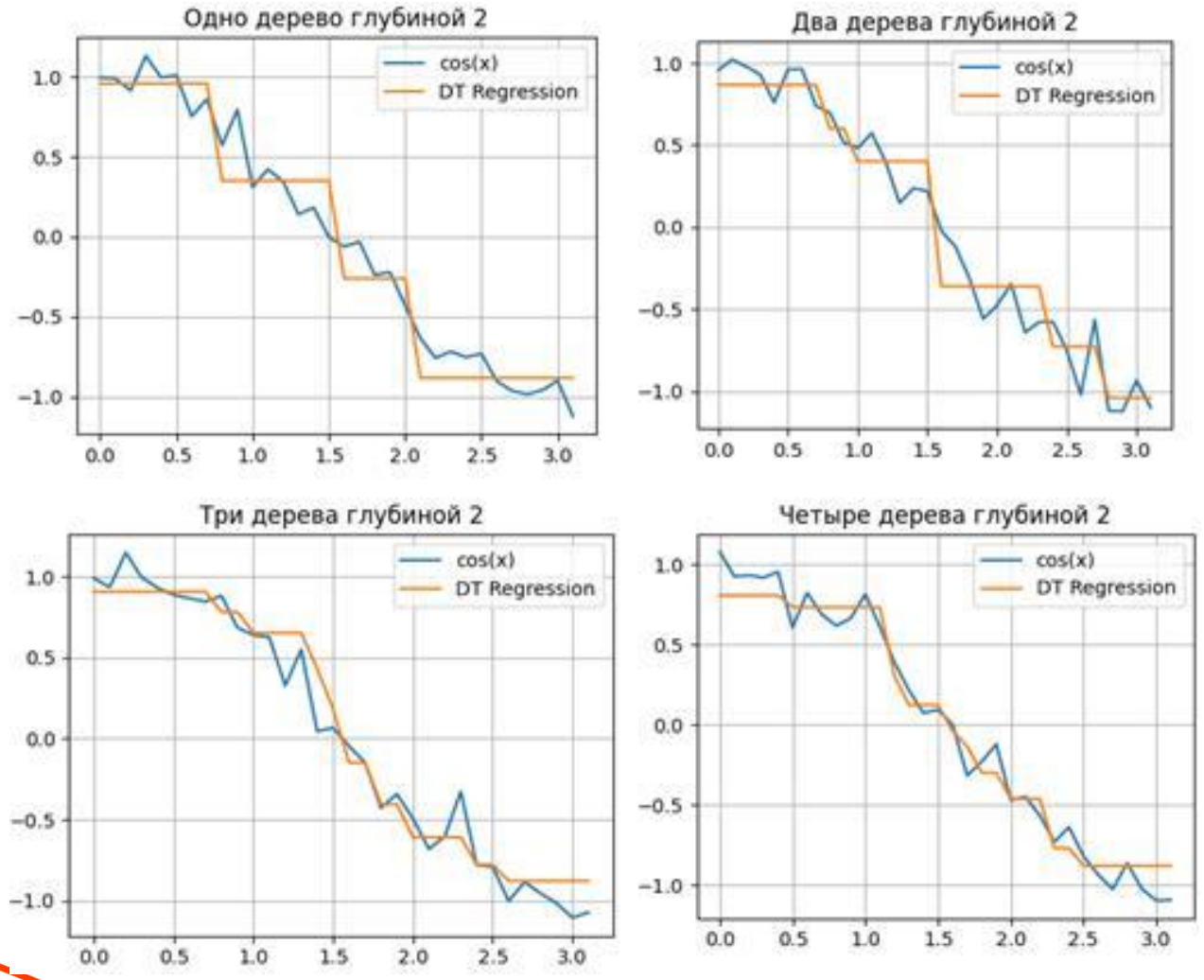


РЕШАЮЩИЕ ДЕРЕВЬЯ – БУТСТРЭП И БЭГГИНГ

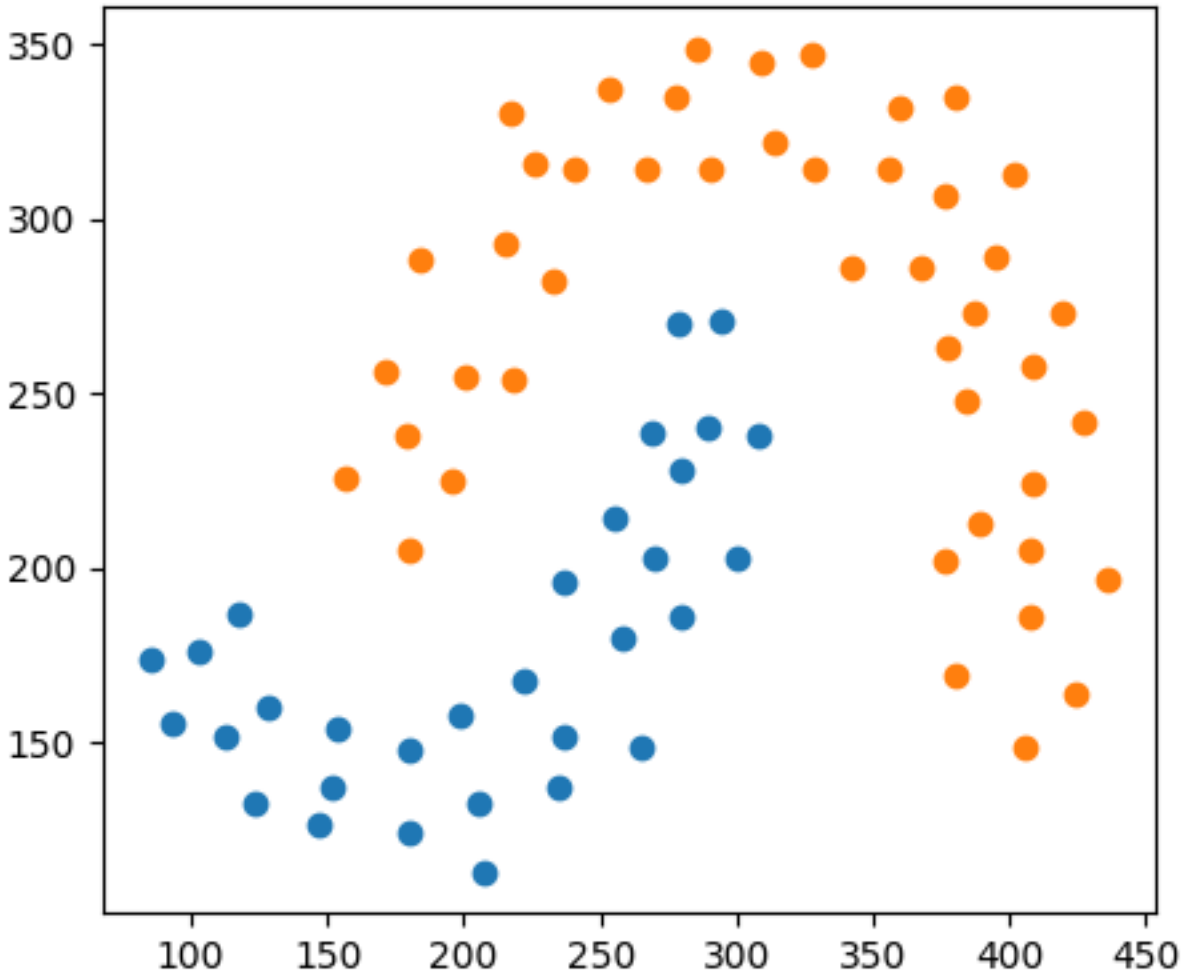


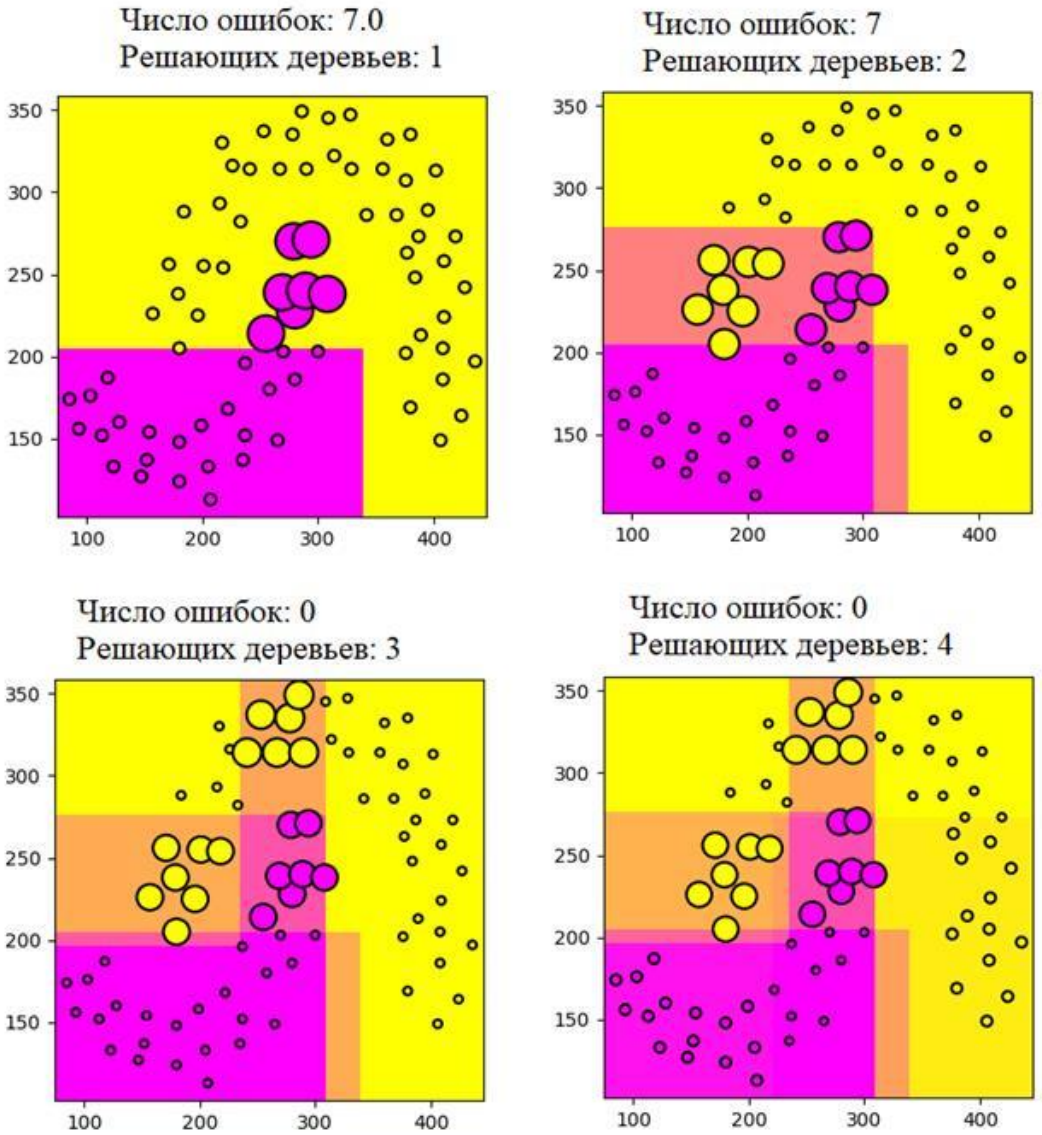


РЕШАЮЩИЕ ДЕРЕВЬЯ – БУТСТРЭП И БЭГГИНГ

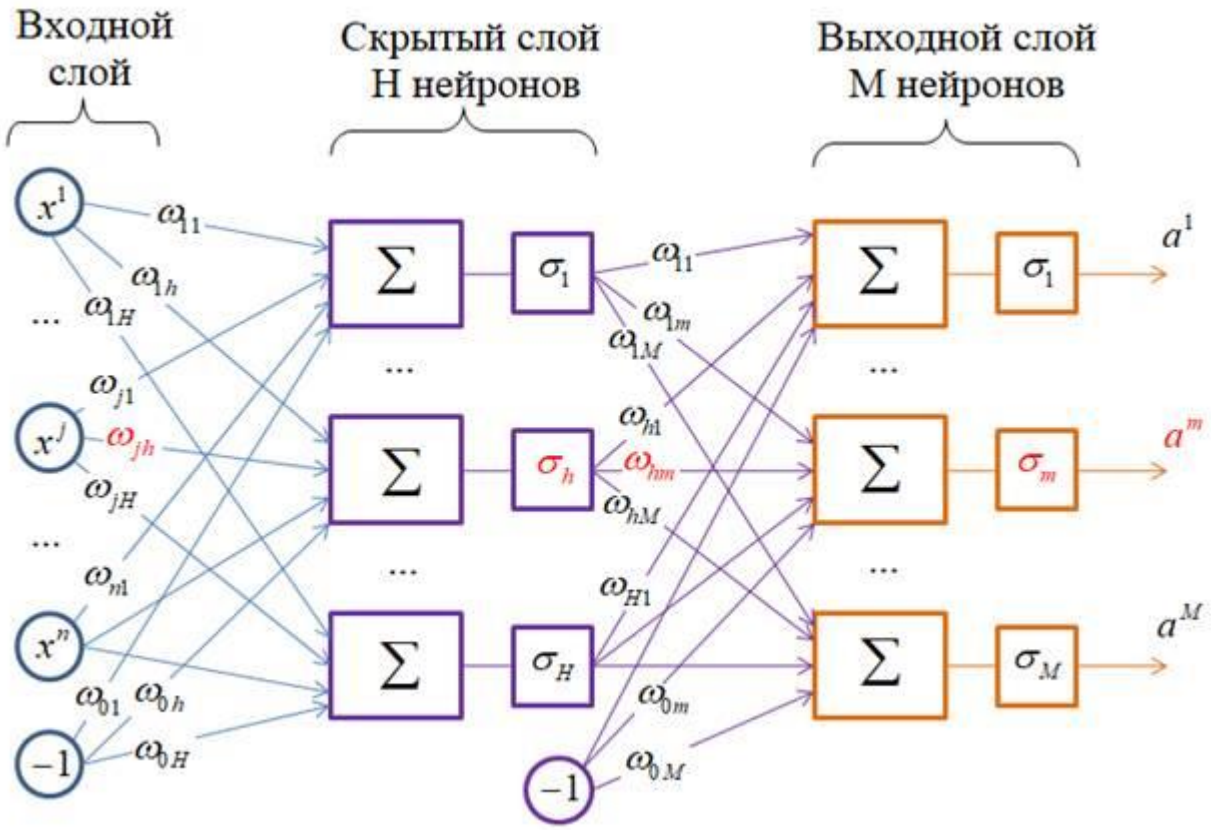


РЕШАЮЩИЕ ДЕРЕВЬЯ – БУСТИНГ



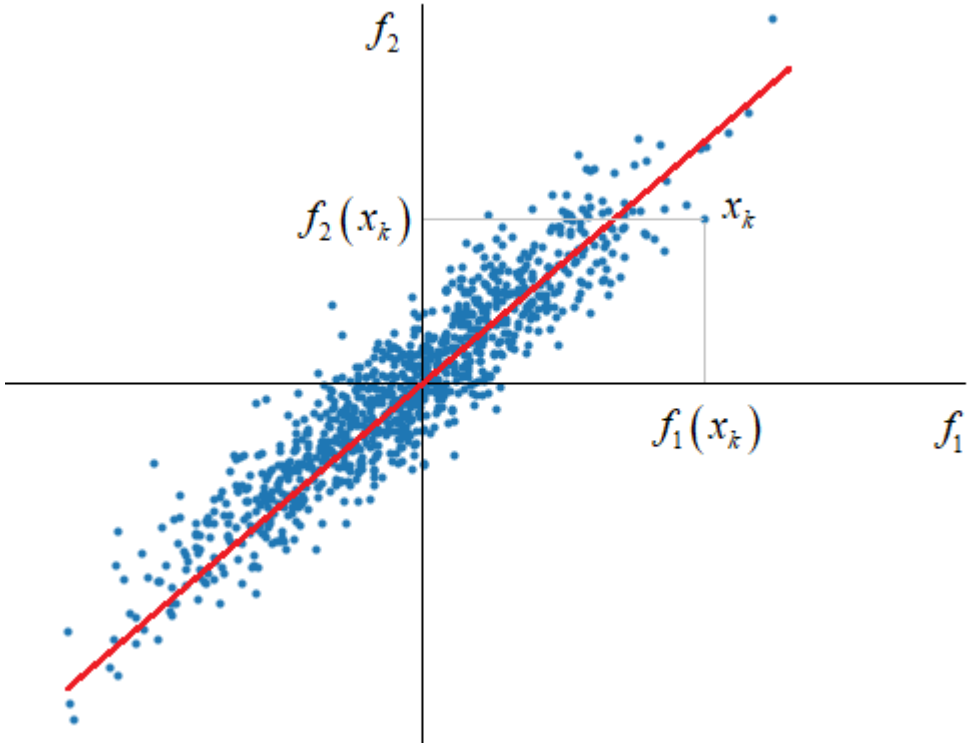


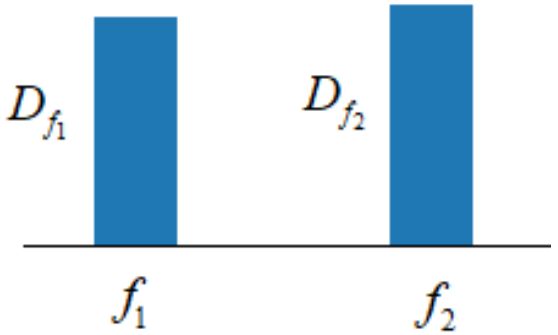
НЕЙРОННЫЕ СЕТИ



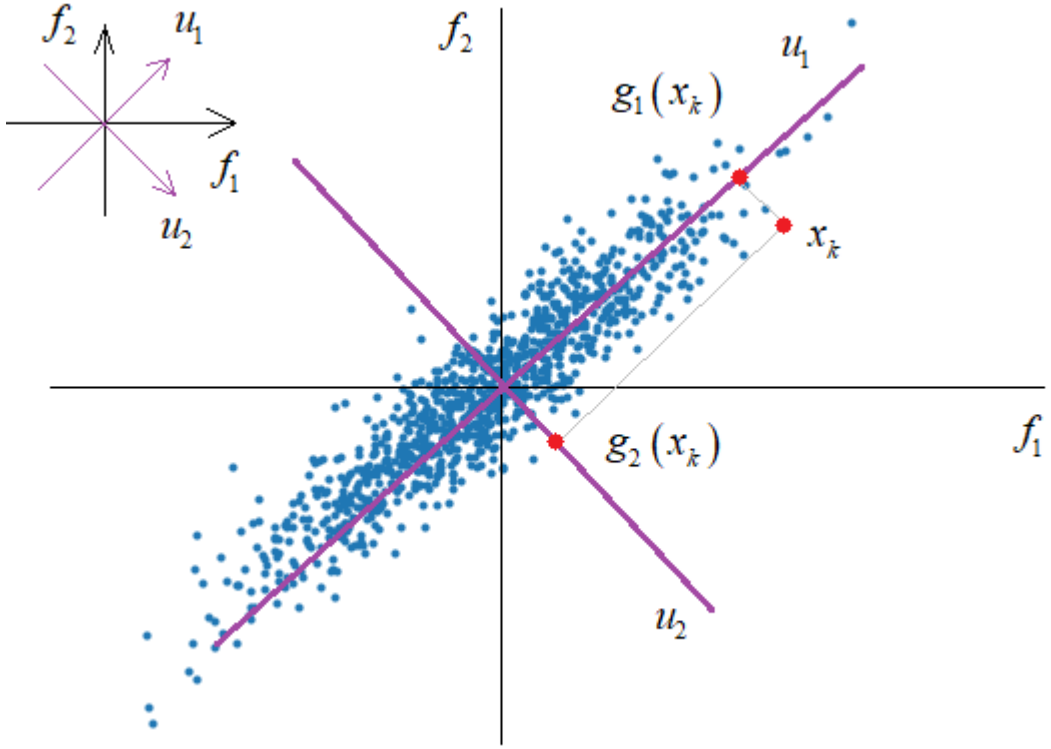
Нейросеть – нелинейная композиция линейных моделей

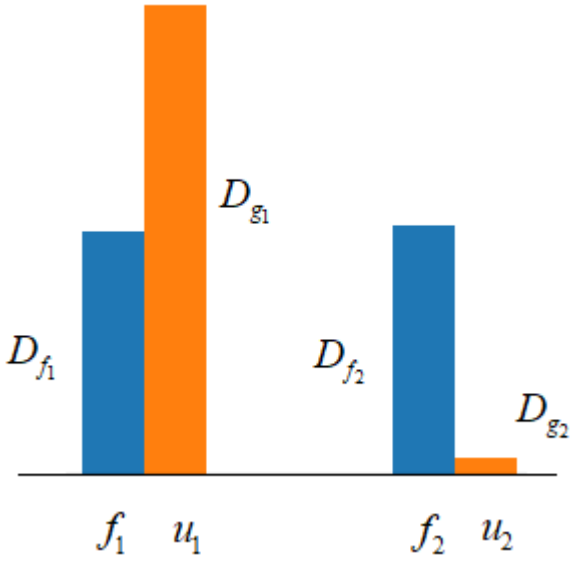
НЕМНОГО ПРО PCA





НЕМНОГО ПРО PCA





ХОРОШИЕ ИСТОЧНИКИ ДЛЯ САМОСТОЯТЕЛЬНОГО ИЗУЧЕНИЯ

https://proproprogs.ru/ml

https://stepik.org/course/209247/

https://girafe.ai/main

https://github.com/girafe-ai/ml-course

https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

Книга Mathematics for machine learning (Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong)