Машинное обучение

Предмет изучения

Машинное обучение занимается построением математических моделей для исследования данных.

$$\text{Машинное обучение} \subset \text{Искусственный интеллект}$$

$$\text{Машинное обучение} \cap \text{Анализ данных} \neq 0$$

$$\text{Машинное обучение} \approx \text{Распознование образов}$$

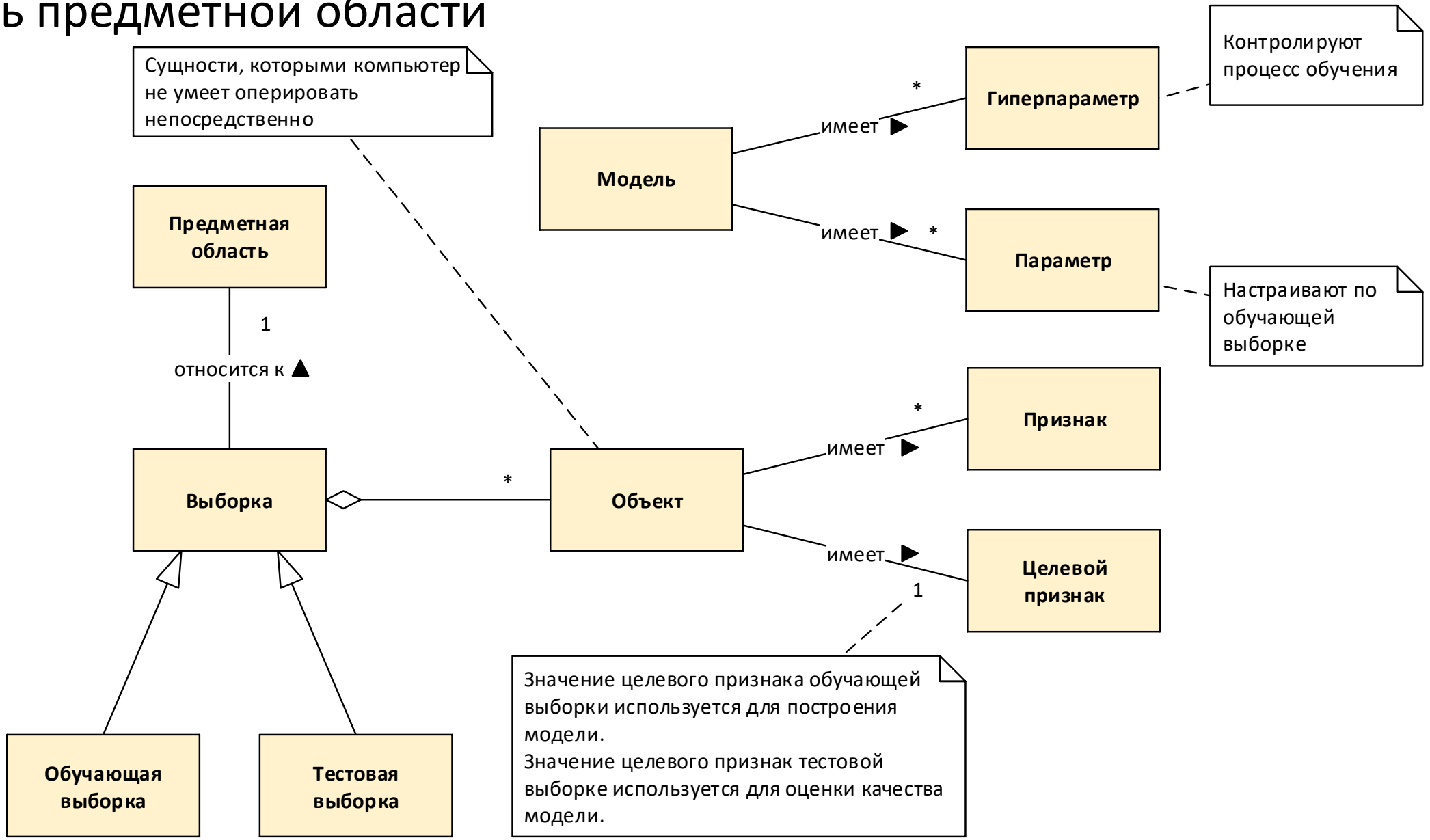
$$\text{Машинное обучение} \approx \text{Обучение по прецедентам}$$

Категории машинного обучения

- **Обучение с учителем** - построение модели на основе обучающих данных и заданного целевого признака
  - **Классификация** - предсказание значения целевого признака (дискретный случай)
  - **Регрессия** - предсказание значения целевого признака (непрерывный случай)
- **Обучение без учителя** - распознавание структуры немаркированных данных
  - **Кластеризация** - разделение данных на отдельные группы
  - **Преобразование данных** - поиск альтернативного представления данных
    - Визуализация
    - понижение размерности,
    - …
- **Частичное обучение** - обучение с учителем, но значение целевого признака известно не всегда

Обучение с учителем

Модель предметной области

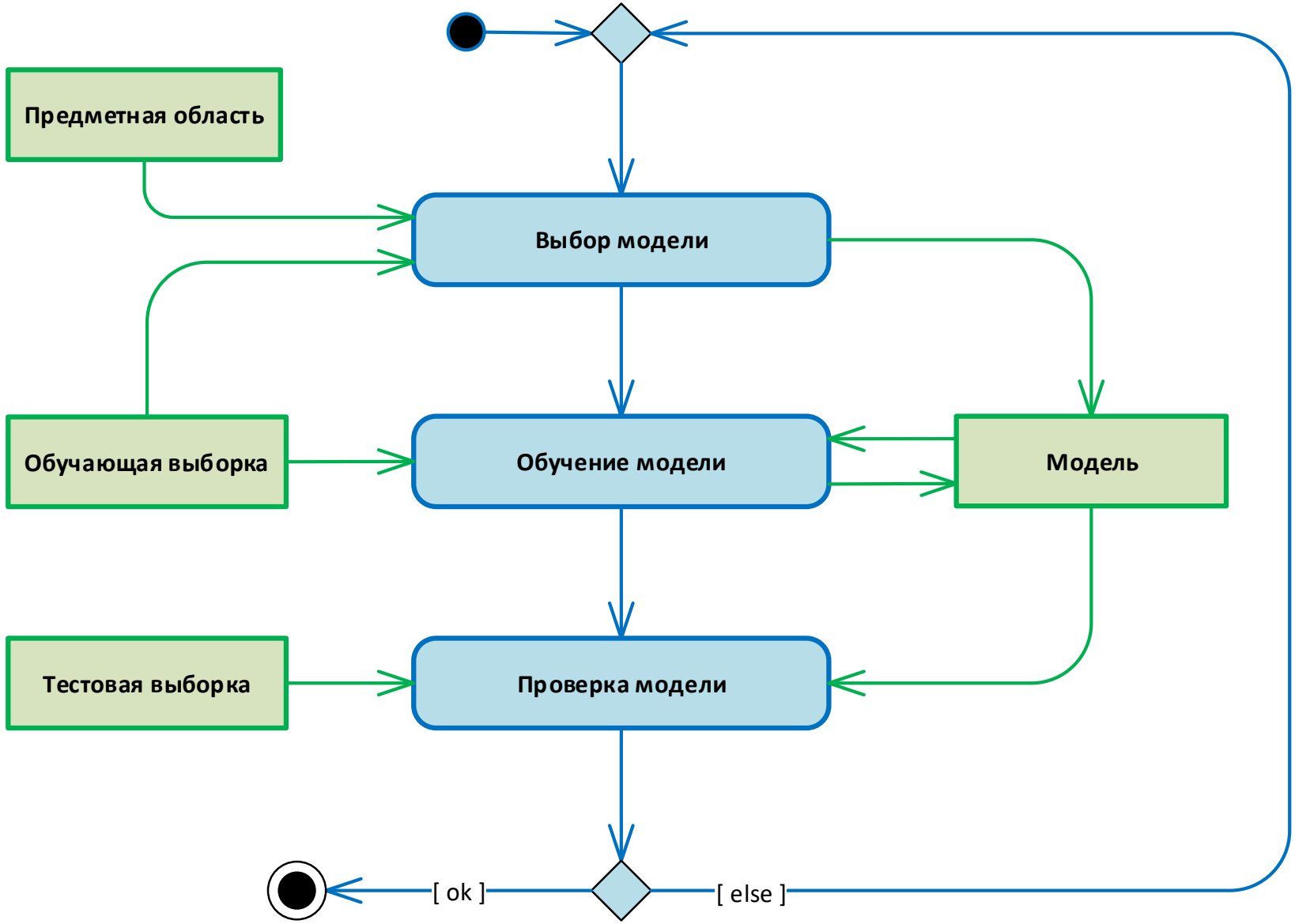


Терминология

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	
4.6	3.4	1.4	0.3	
5.0	3.4	1.5	0.2	
4.4	2.9	1.4	0.2	
4.9	3.1	1.5	0.1	

Обучение с учителем

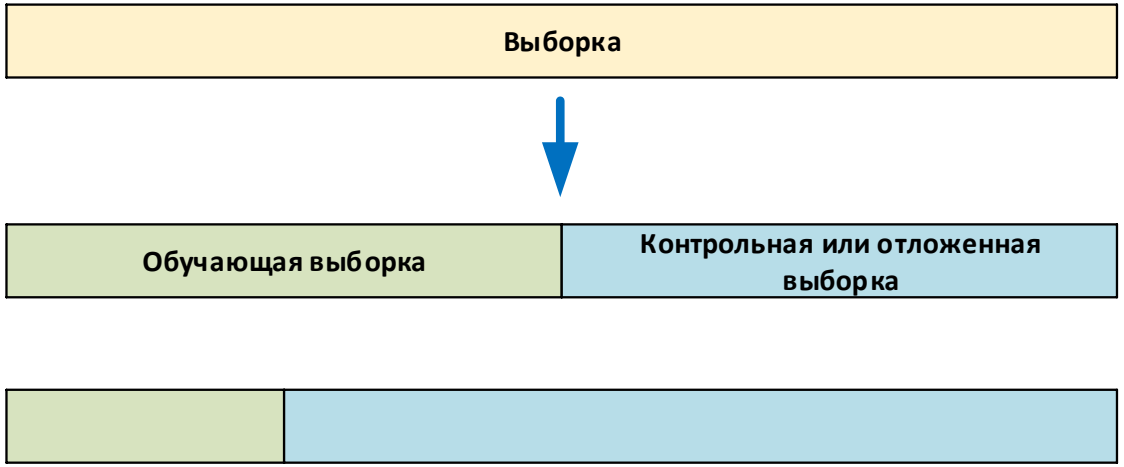
Алгоритм построения модели



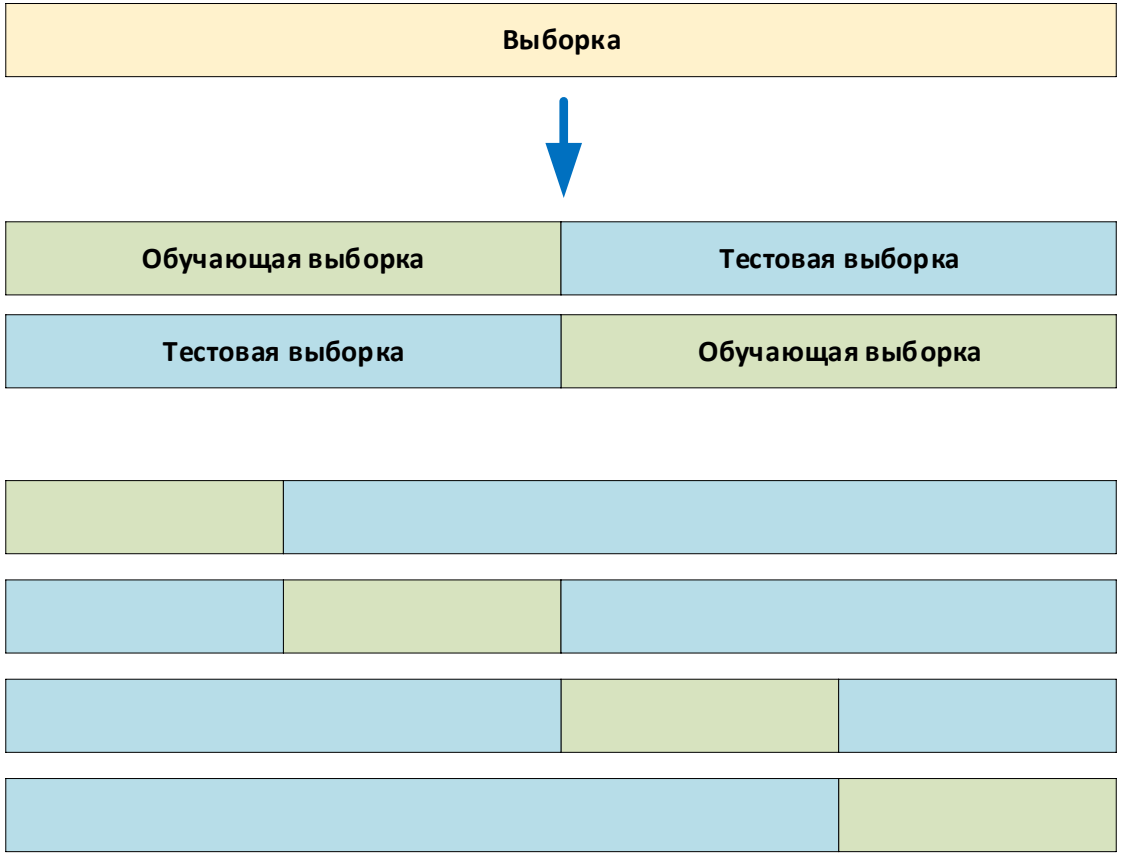
Обучение с учителем

Проверка модели

Отложенные данные



Перекрестная проверка



Обучение с учителем

- Мера качества (классификация)

$$\text{доля правильных ответов} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{точность} = \frac{TP}{TP + FP}$$

$$\text{полнота} = \frac{TP}{TP + FN}$$

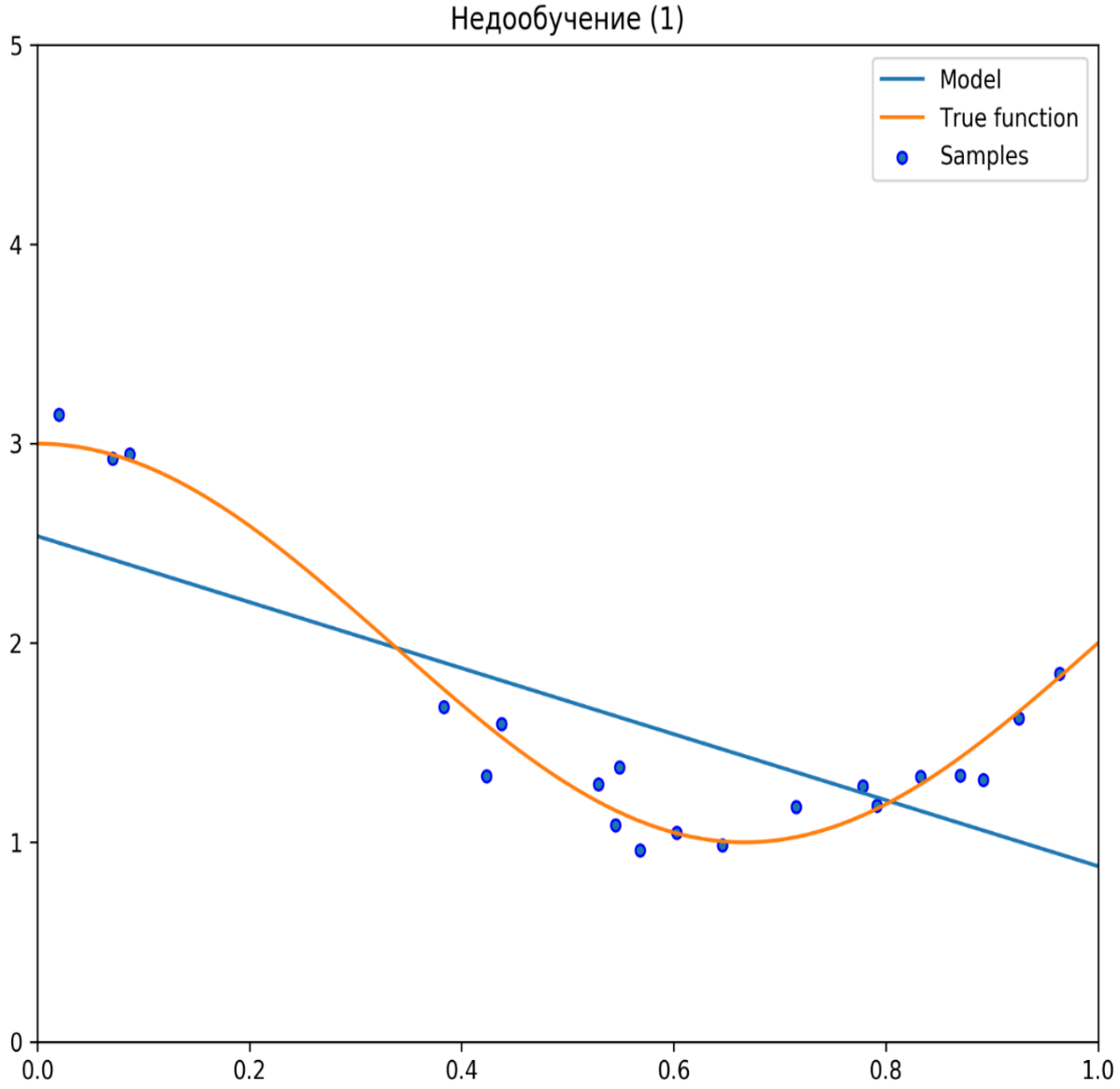
$$\text{F} = \frac{2 \cdot \text{точность} \cdot \text{полнота}}{\text{точность} + \text{полноста}}$$

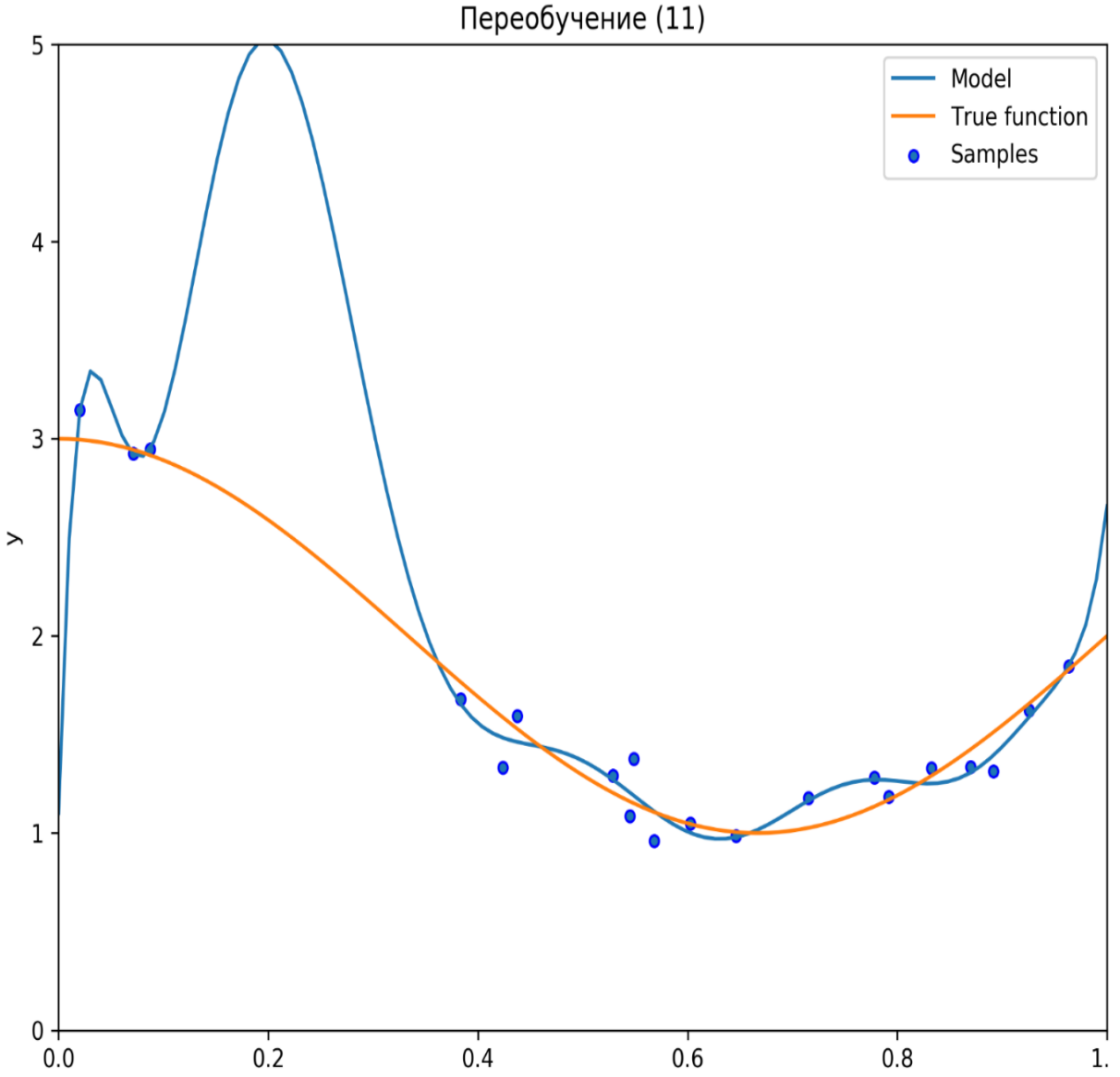
Матрица ошибок / несоответствий

	1	-1
1	TP	FP
-1	FN	TN

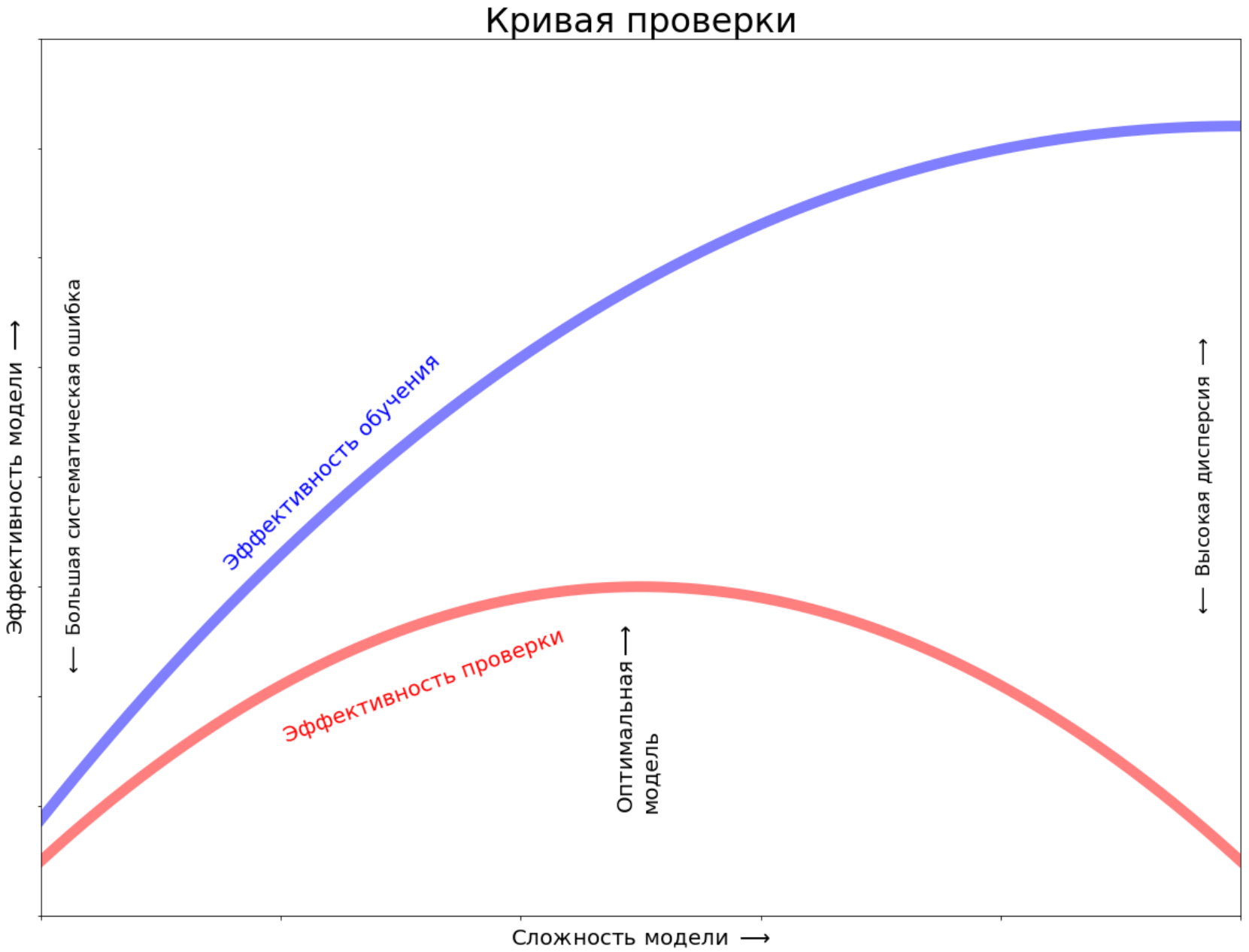
Обучение с учителем

Выбор модели

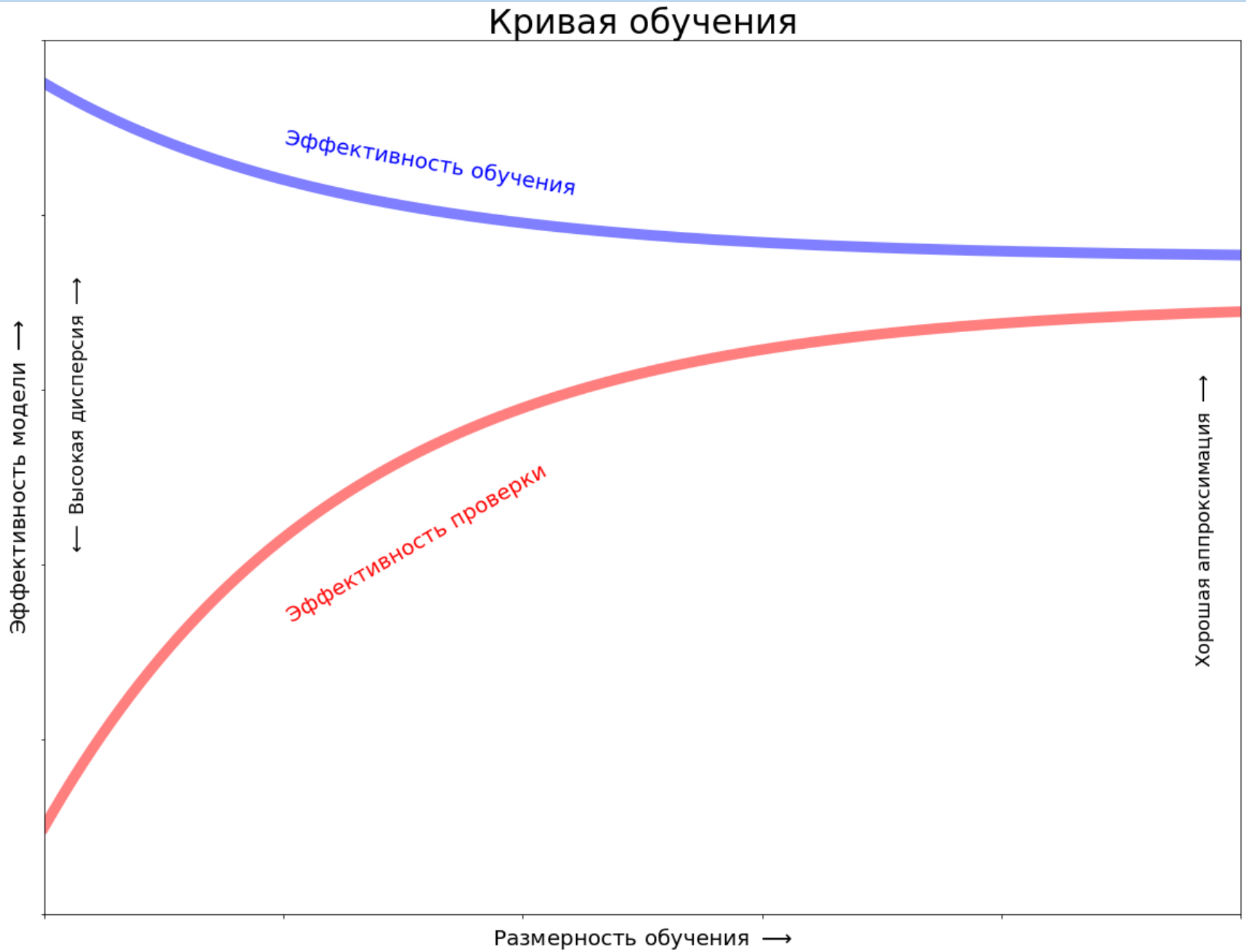




Обучение с учителем



Обучение с учителем



Обучение с учителем

- Мера качества (регрессия)

$$y = w_0 + w_1 x_1 + \cdots + w_m x_m$$

$$E = \frac{1}{n} \cdot \sum (y_i - \hat{y}_i)^2$$

- Регуляризация
  - Используется для того, чтобы избежать переобучения, путем введения штрафов за дополнительные признаки
  - L1-регуляризация

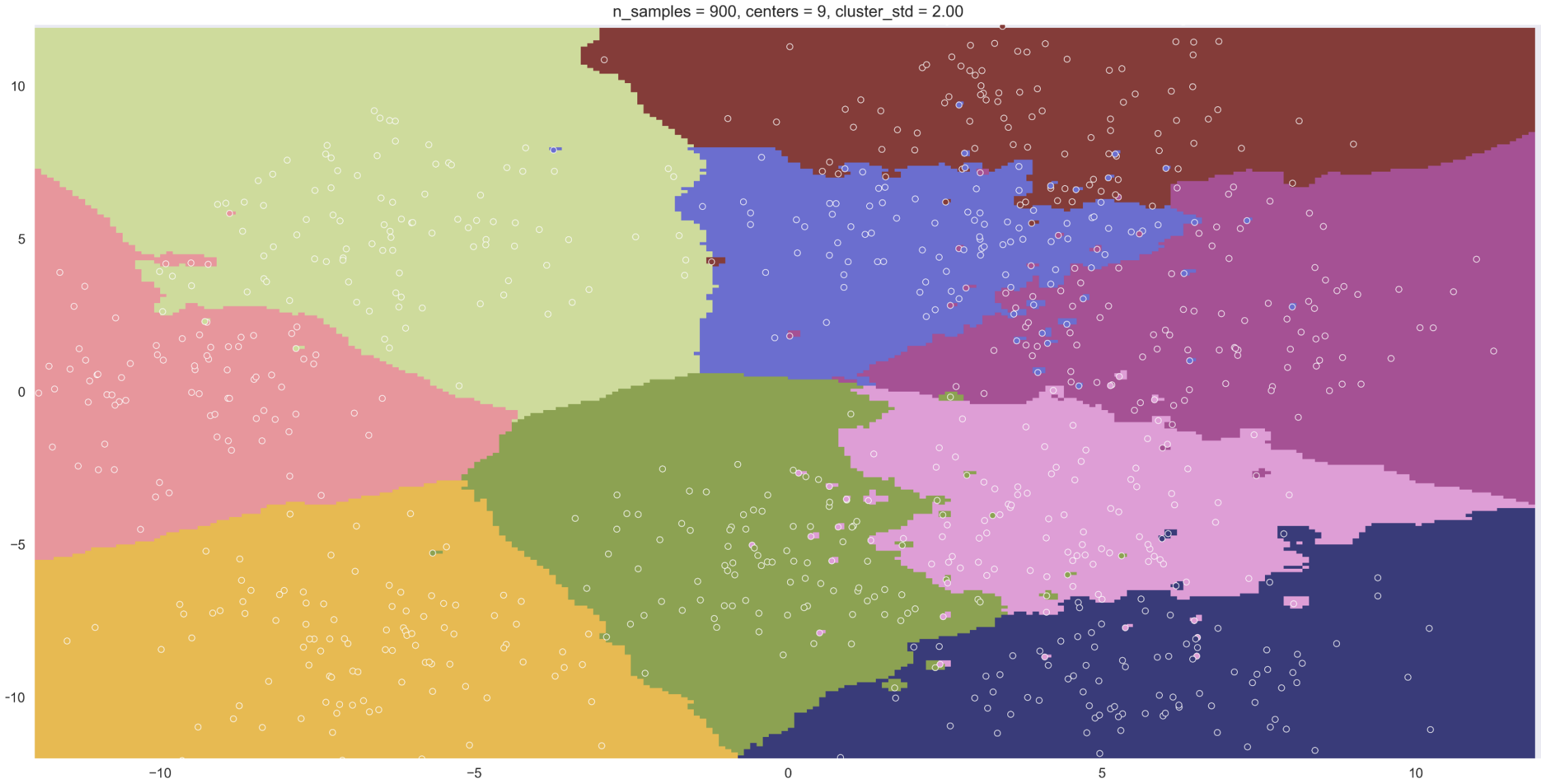
$$E = \frac{1}{n} \cdot \sum (y_i - \hat{y}_i)^2 + \alpha \sum |w|$$

  - L2-регуляризация

$$E = \frac{1}{n} \cdot \sum (y_i - \hat{y}_i)^2 + \alpha \sum w^2$$

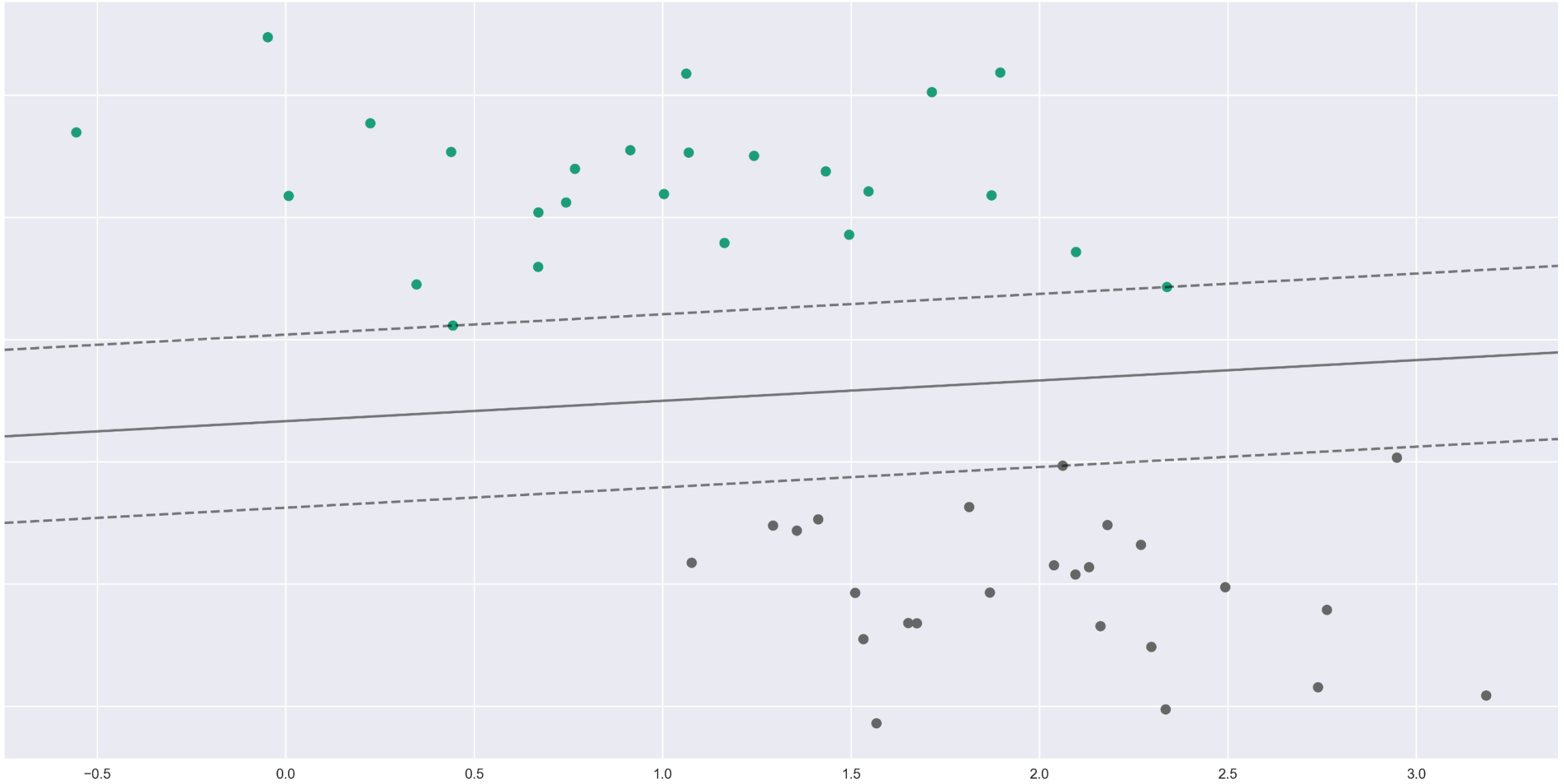
Обучение с учителем. Классификация

- Метод k-ближайших соседей



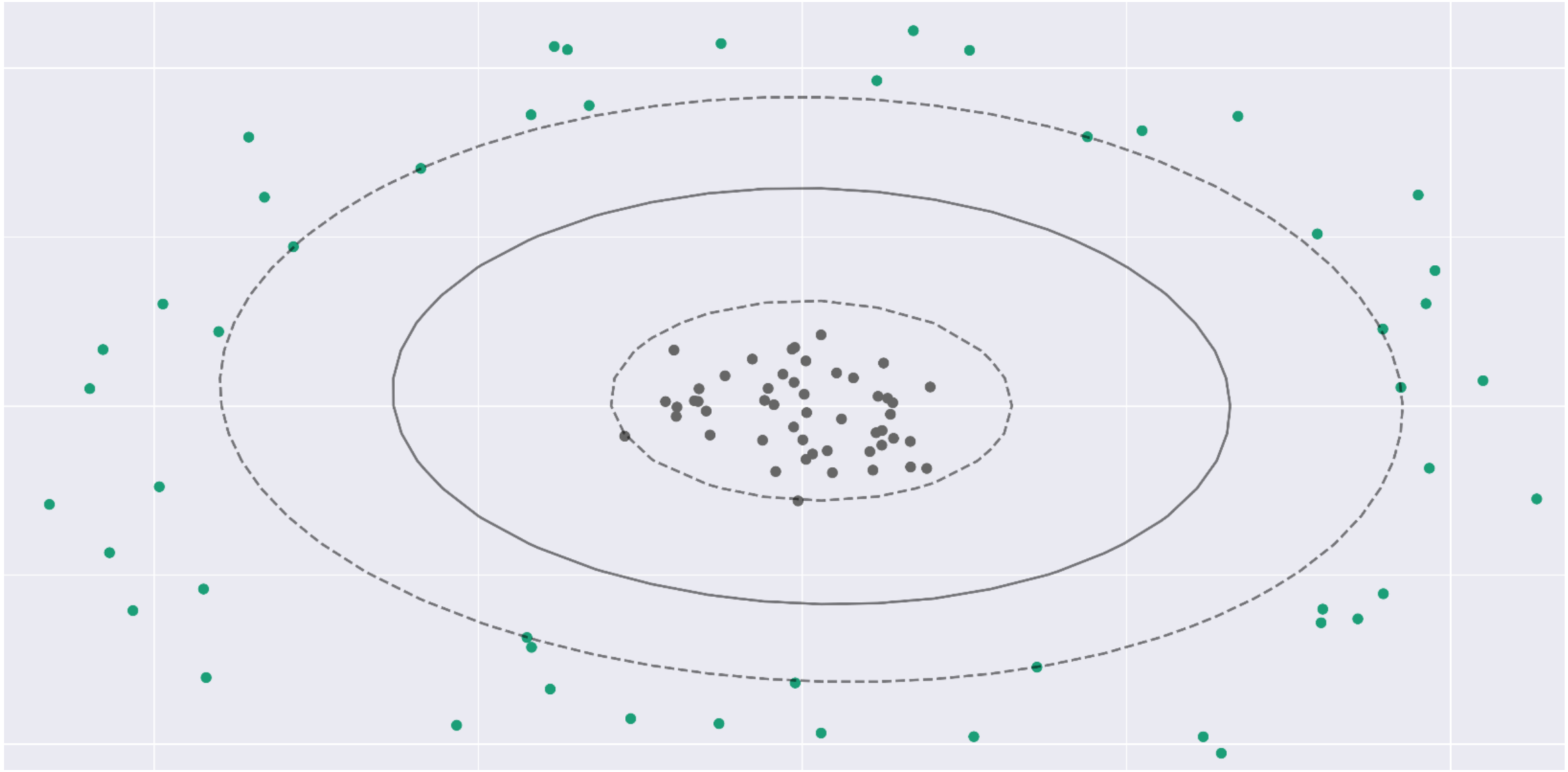
Обучение с учителем. Классификация

- Метод опорных векторов



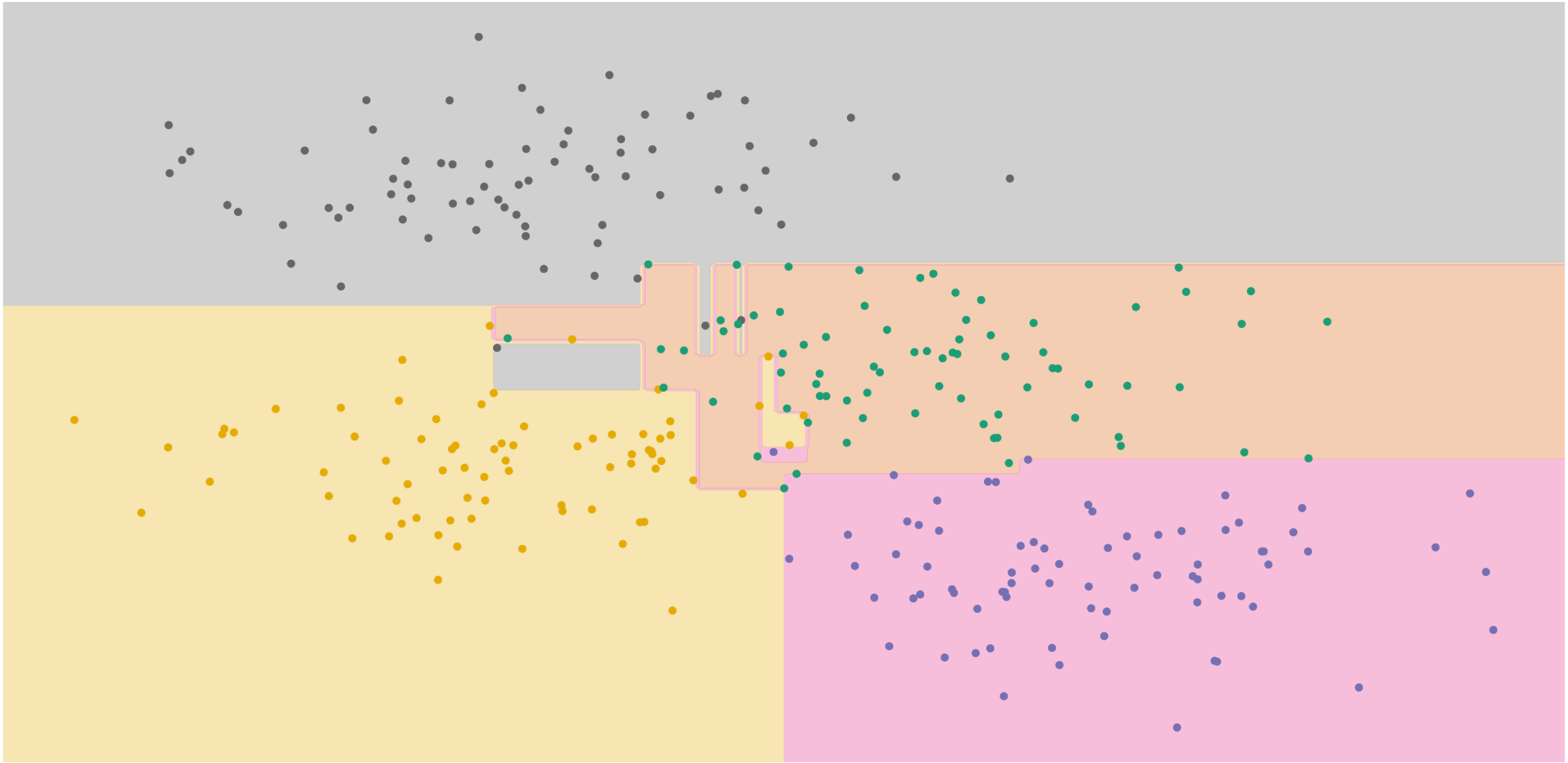
Обучение с учителем. Классификация

- Ядерное обобщение метода опорных векторов

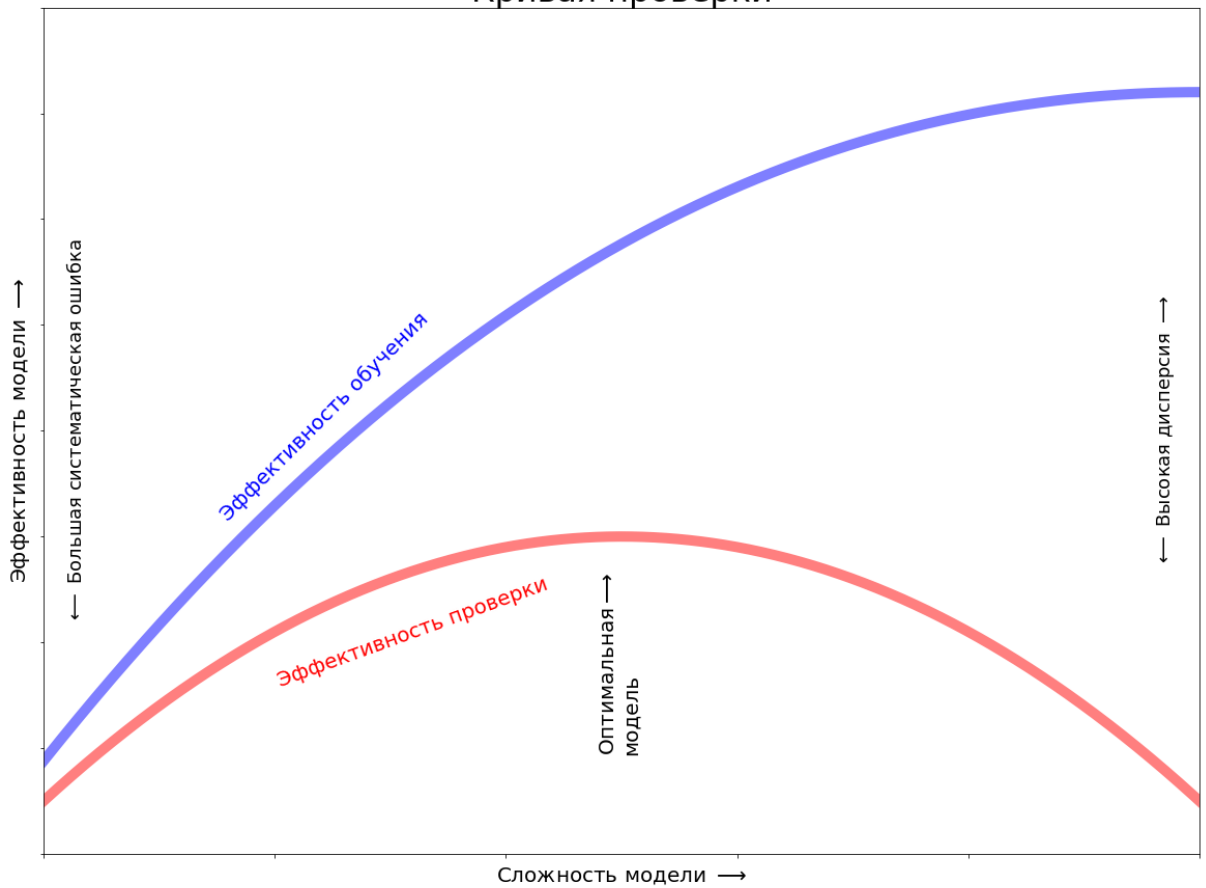


Обучение с учителем. Классификация

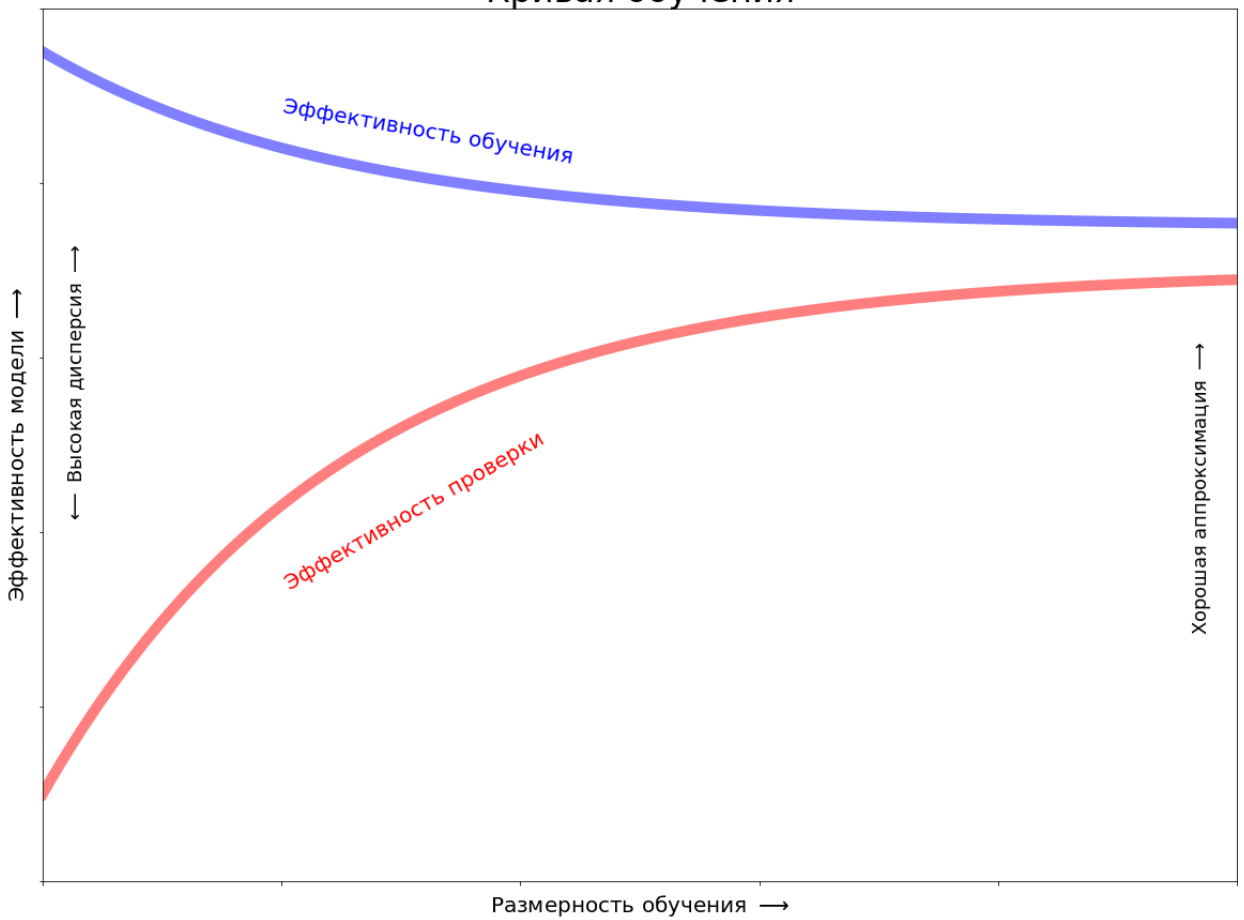
- Дерево принятия решений



Кривая проверки



Кривая обучения

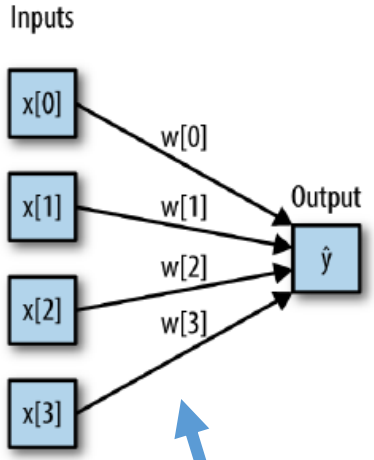


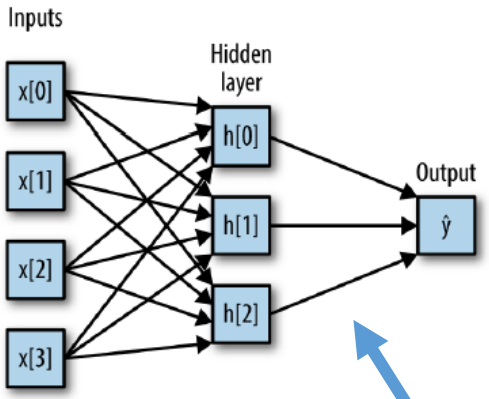
Обучение с учителем. Классификация

- Ансамбли деревьев решений
  - Bagging (**B**ootstrap **agg**regation)
    - Параллельное обучение + голосование
  - Градиентный бустинг
    - Последовательное обучение

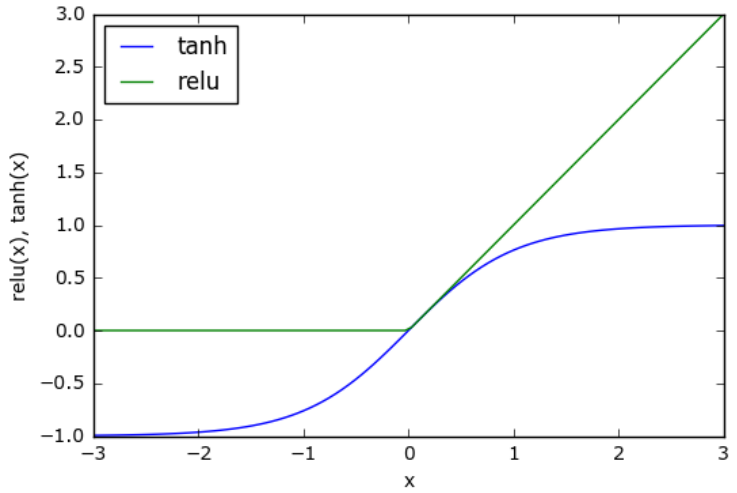
Обучение с учителем

- Многослойный персептрон





Функция активации



$$y = w[0]x[0] + w[1]x[1] + w[2]x[2] + w[3]x[3]$$

$$h[0] = \tanh(\, w[0,0]x[0] + w[1,0]x[1] + w[2,0]x[2] + w[3,0]x[3])$$

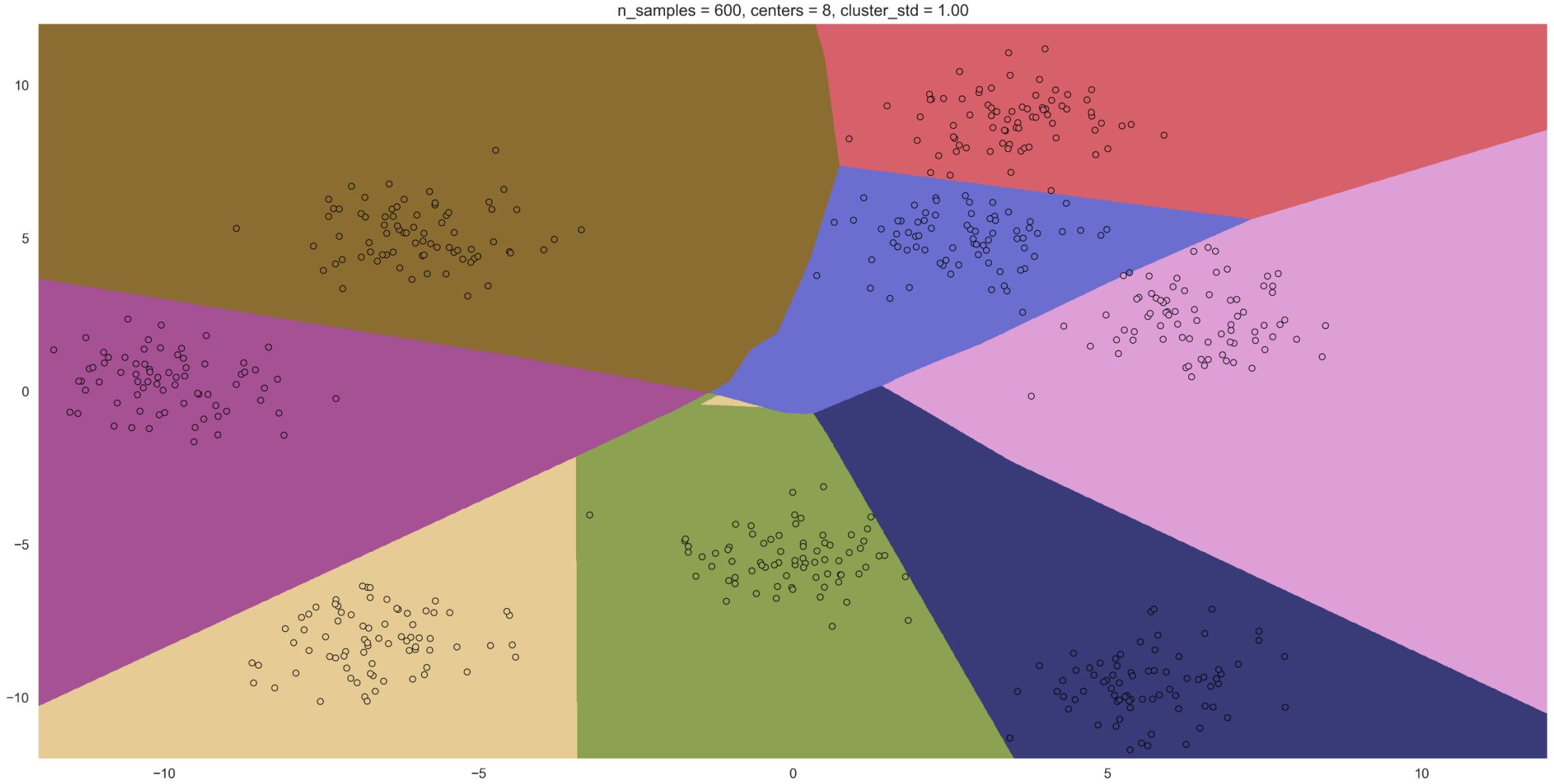
$$h[1] = \tanh(\, w[0,1]x[0] + w[1,1]x[1] + w[2,1]x[2] + w[3,1]x[3])$$

$$h[2] = \tanh(\, w[0,2]x[0] + w[1,2]x[1] + w[2,2]x[2] + w[3,2]x[3])$$

$$y = v[0]h[0] + v[1]h[1] + v[2]h[2]$$

Обучение с учителем. Классификация

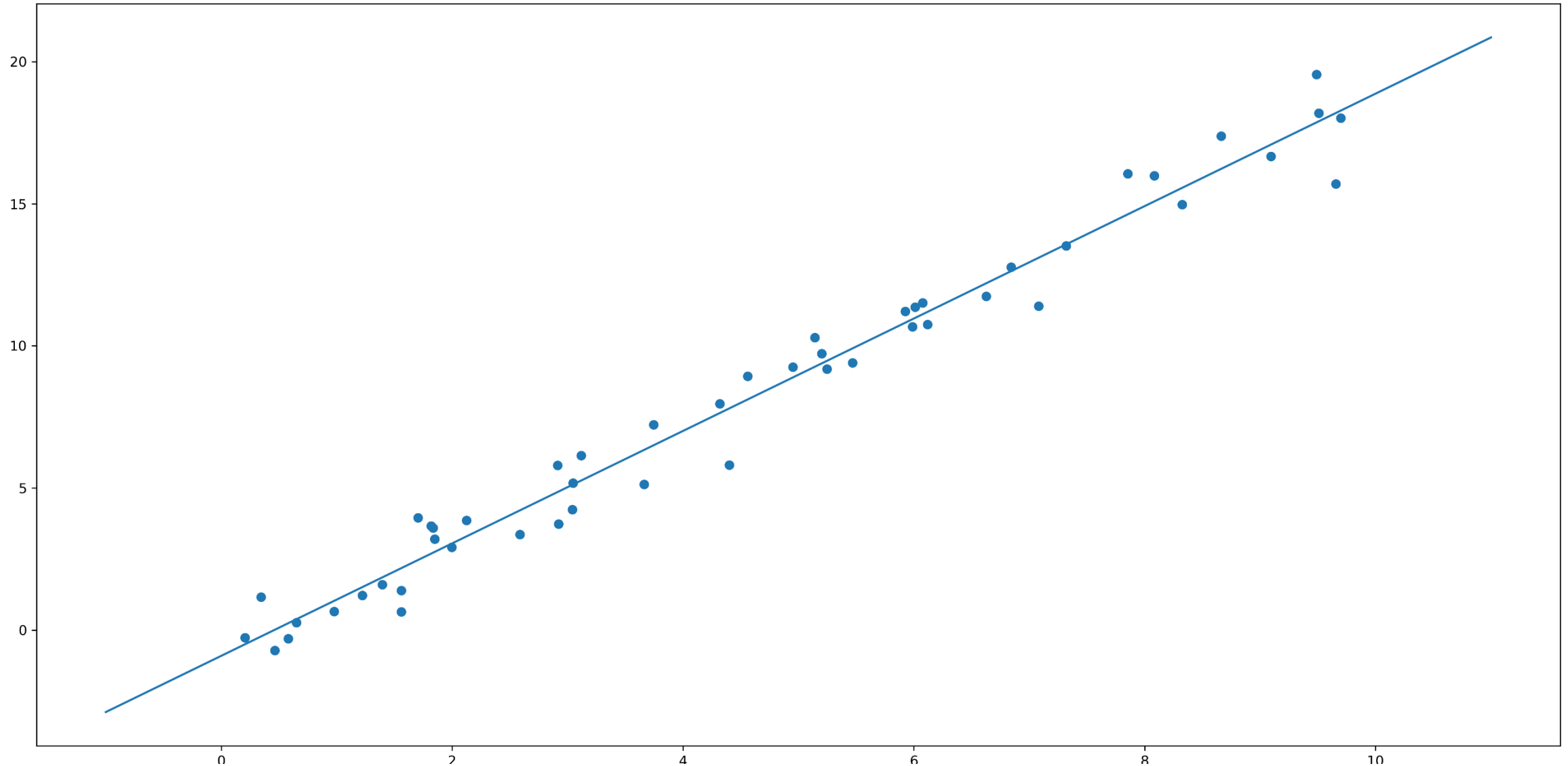
- Многослойный персептрон



Обучение с учителем. Регрессия

- Линейная регрессия

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$



Обучение с учителем. Регрессия

- Регрессия по комбинации базисных функций

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \ldots$$

Линейная регрессия

$$x_i = f_i(x) = x^i$$

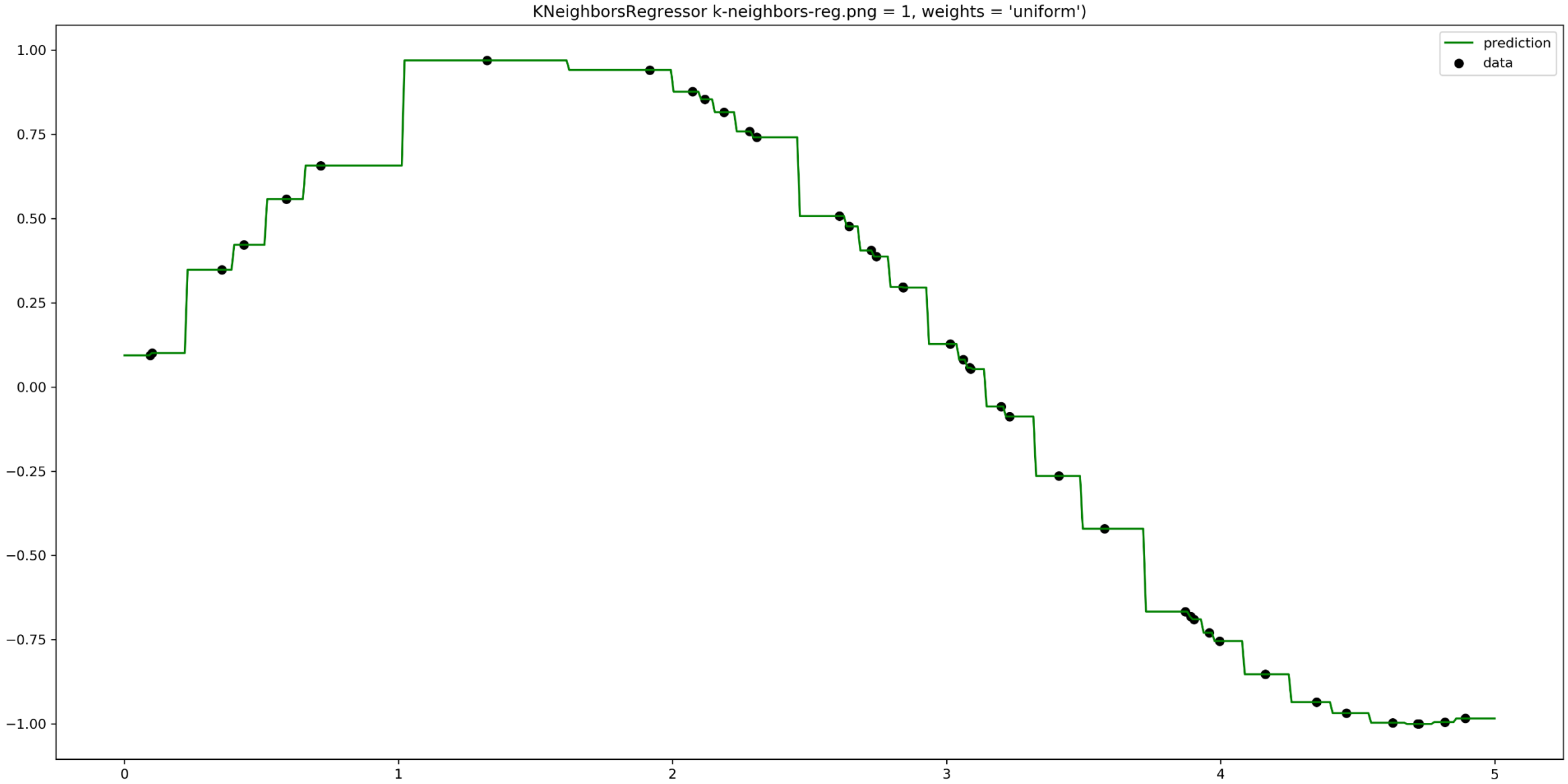
$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots$$

Полиномиальная регрессия

Модель остается линейной, так как a никогда не умножаются и не делятся друг на друга

Обучение с учителем. Регрессия

- Метод k-ближайших соседей



Обучение без учителя

- Метод главных компонент