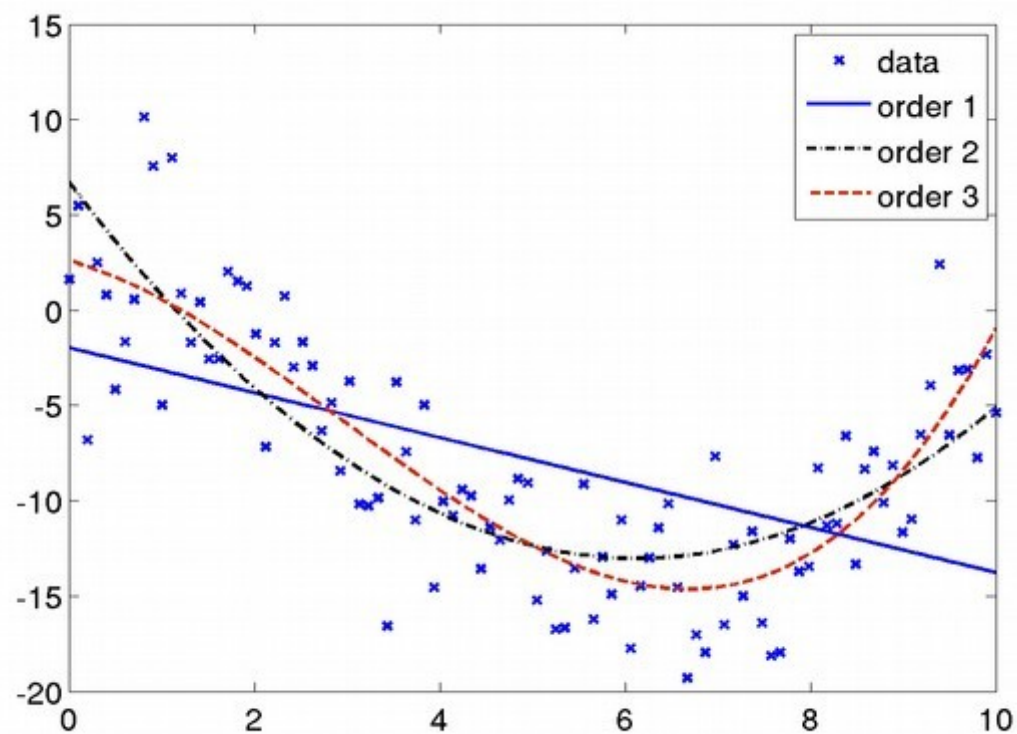


# Машинное обучение

## Лекция 2. Основные понятия



# Содержание лекции

- Задача обучения
- Матрица объектов-признаков
- Модель алгоритмов и метод обучения
- Функционал качества
- Вероятностная постановка задачи обучения
- Проблема переобучения

# Литература

- <http://www.machinelearning.ru>
- Курс К.В.Воронцова
- <https://www.kaggle.com/>

# Задача обучения

$X$  — множество объектов

$Y$  — множество ответов

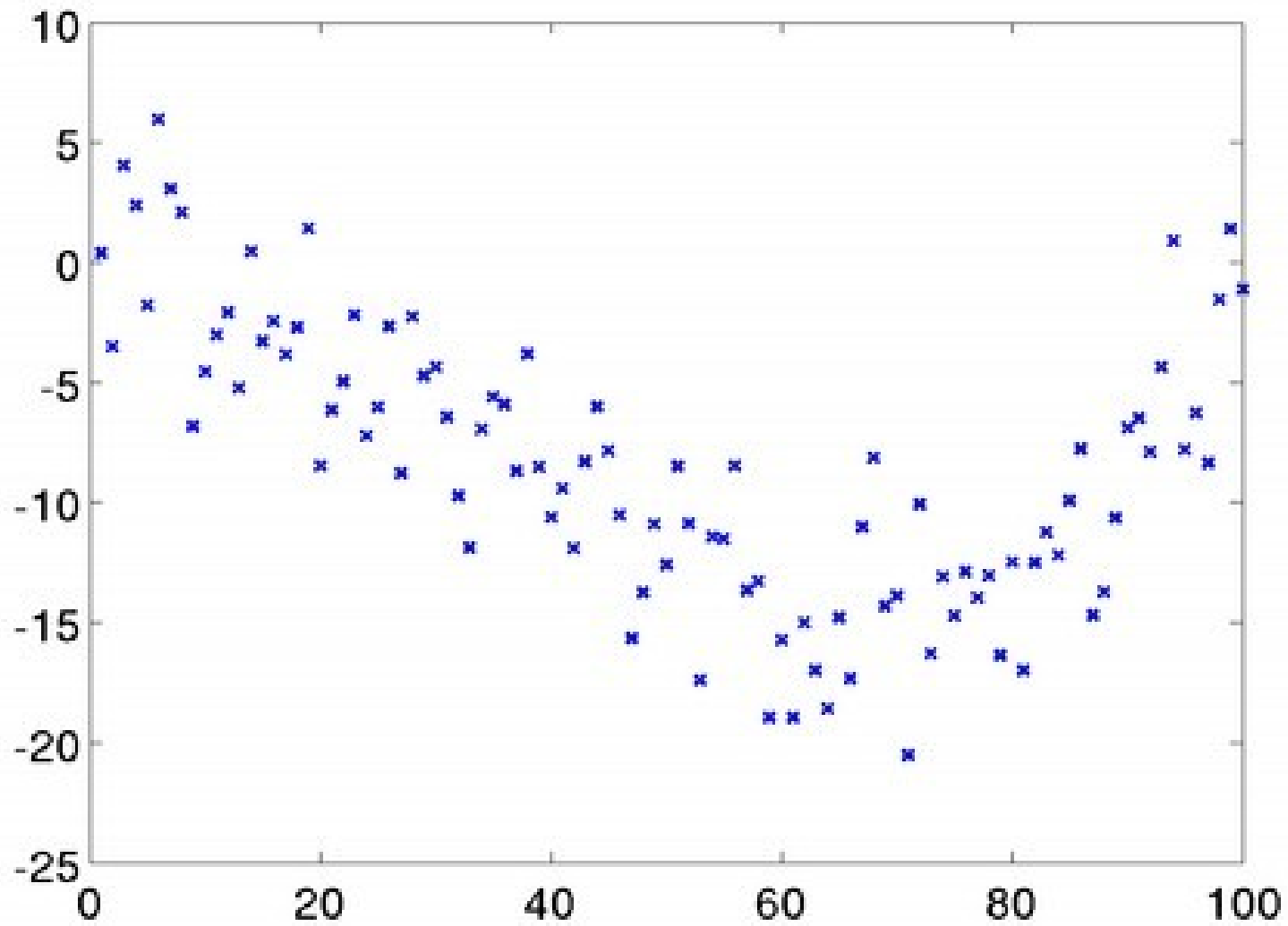
$y : X \rightarrow Y$  — неизвестная зависимость  
(target function)

**Дано:**

$\{x_1, \dots, x_\ell\} \subset X$  — обучающая выборка  
(training sample)

$y_i = y(x_i), i = 1, \dots, \ell$  — известные ответы

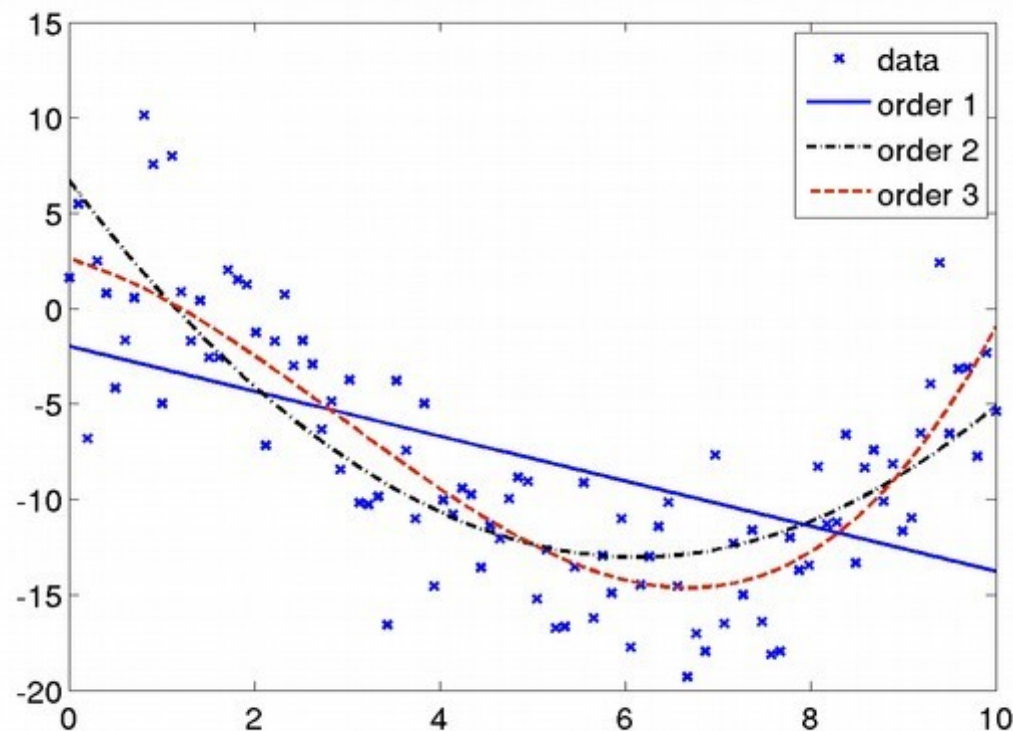
# Задача обучения



# Задача обучения

Найти:

$a : X \rightarrow Y$  — алгоритм, решающую функцию (decision function), приближающую  $y$  на всём множестве  $X$



# Типы задач

## Задачи классификации (classification):

$Y = \{-1, +1\}$  — классификация на 2 класса

$Y = \{1, \dots, M\}$  — на  $M$  непересекающихся классов (multi-class classification)

$Y = \{0, 1\}^M$  — на  $M$  классов, которые могут пересекаться (multi-label classification).

## Задачи восстановления регрессии (regression):

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

## Задачи ранжирования (ranking):

$Y$  — конечное упорядоченное множество

# Признаки

- Компьютер всегда имеет дело с признаковым описанием объектов. Например: пациента можно описать признаками: имя, возраст, номер полиса, жалобы, давление, температура, результаты анализов

- $f: X \rightarrow D_f$

- Типы признаков:

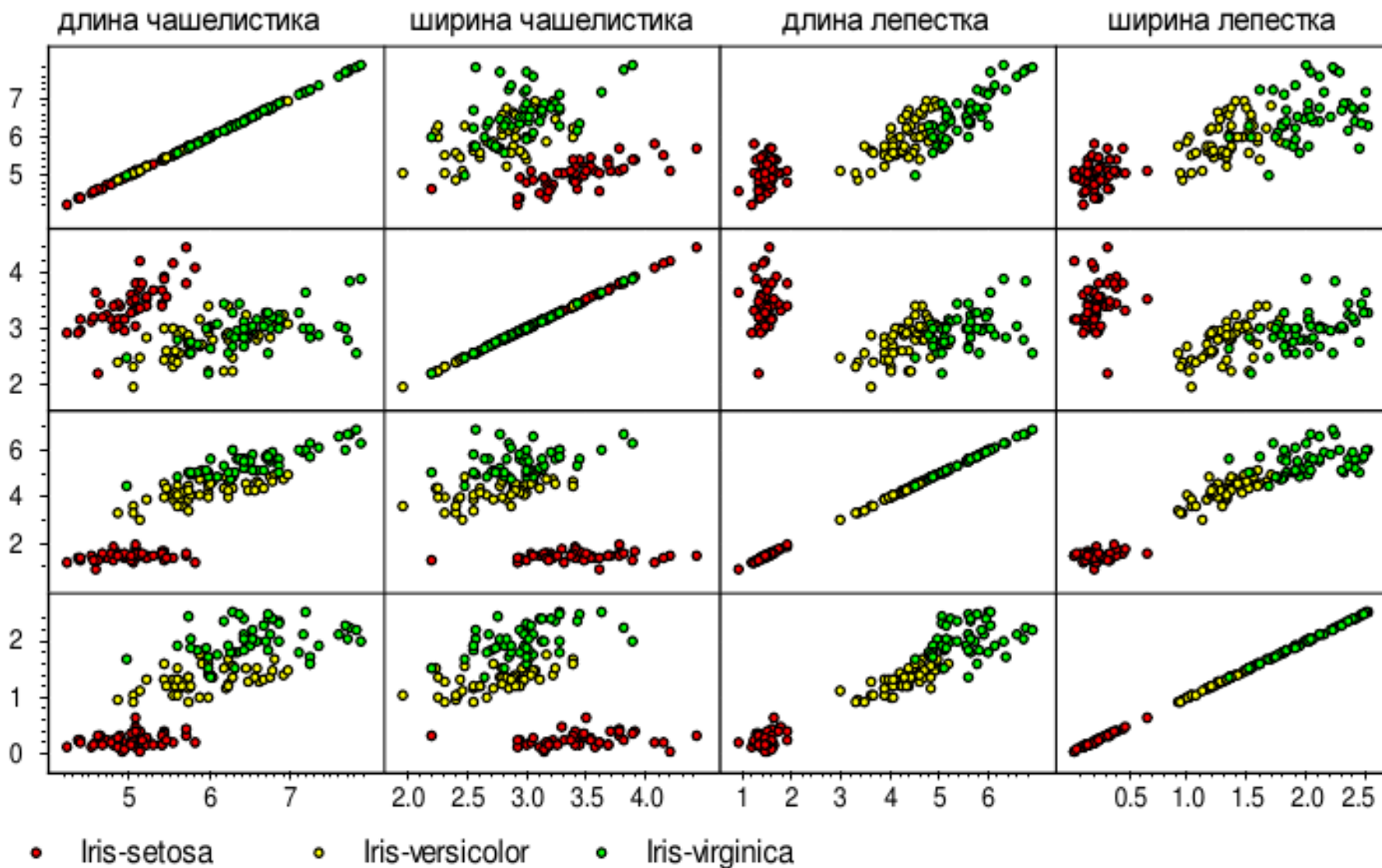
- бинарный
- номинальный
- порядковый
- количественный

Матрица объектов-признаков:

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$



# Пример. Задача классификации видов ириса (Фишер 1936)



# Модель и алгоритм обучения

- **Модель** – это семейство “гипотез”

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

одна из которых (как мы надеемся)  
хорошо приближает целевую функцию

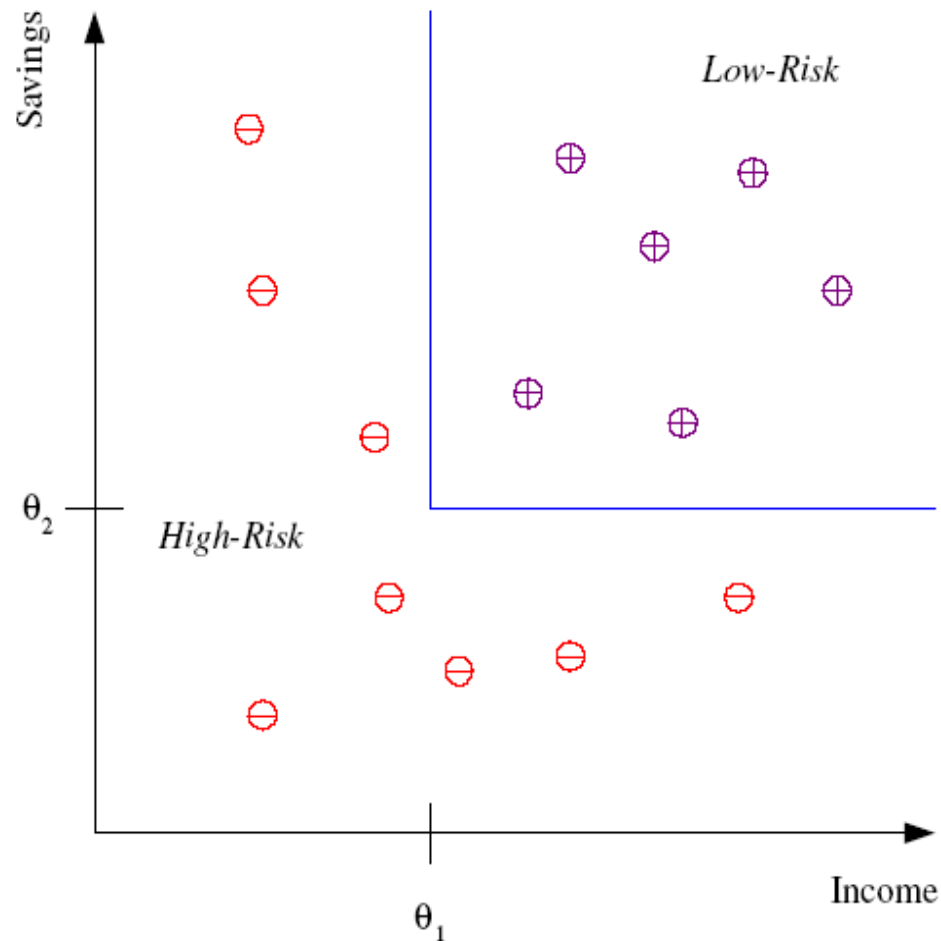
- **Алгоритм обучения**

$$\mu: (X \times Y)^\ell \rightarrow A$$

находит гипотезу в модели, которая  
наилучшим образом приближает  
целевую функцию, используя известные  
значения (обучающую выборку)

# Пример - классификация

- Кредитный скоринг
- Разделение клиентов на **low-risk** и **high-risk** по их зарплате и сбережениям

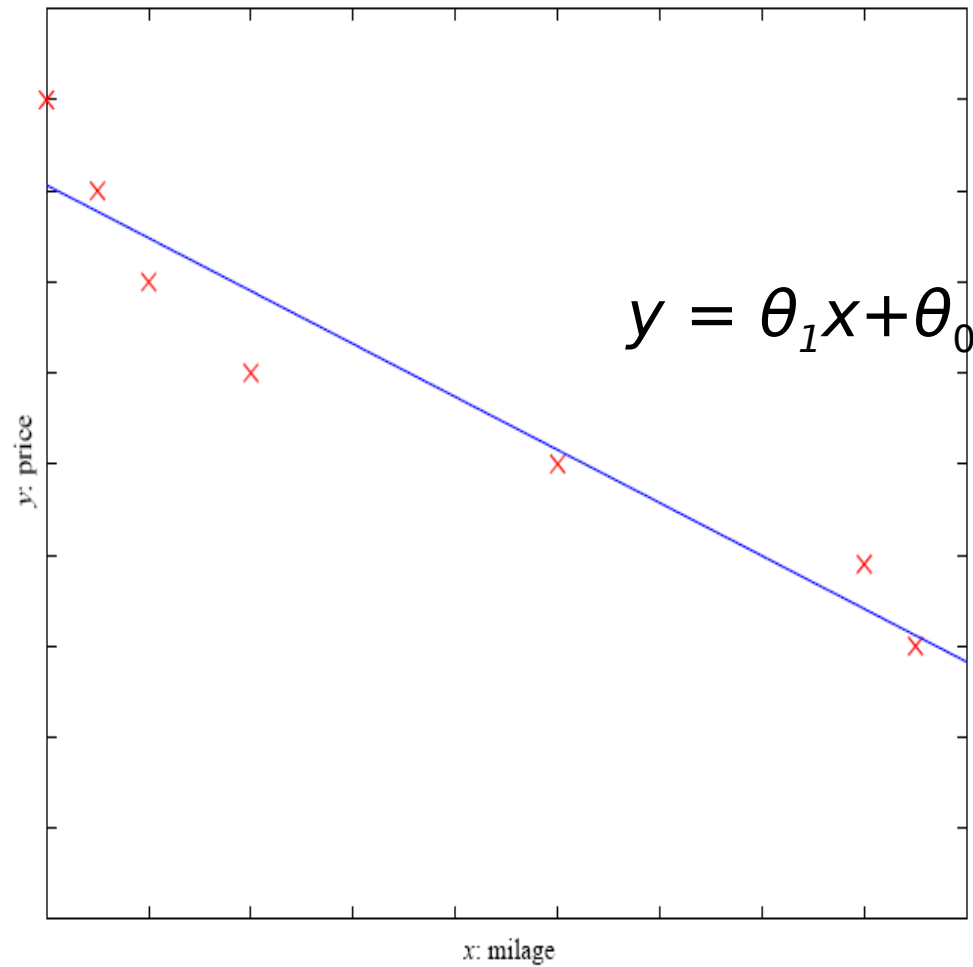


IF  $income > \theta_1$  AND  $savings > \theta_2$   
THEN **low-risk** ELSE **high-risk**

Модель

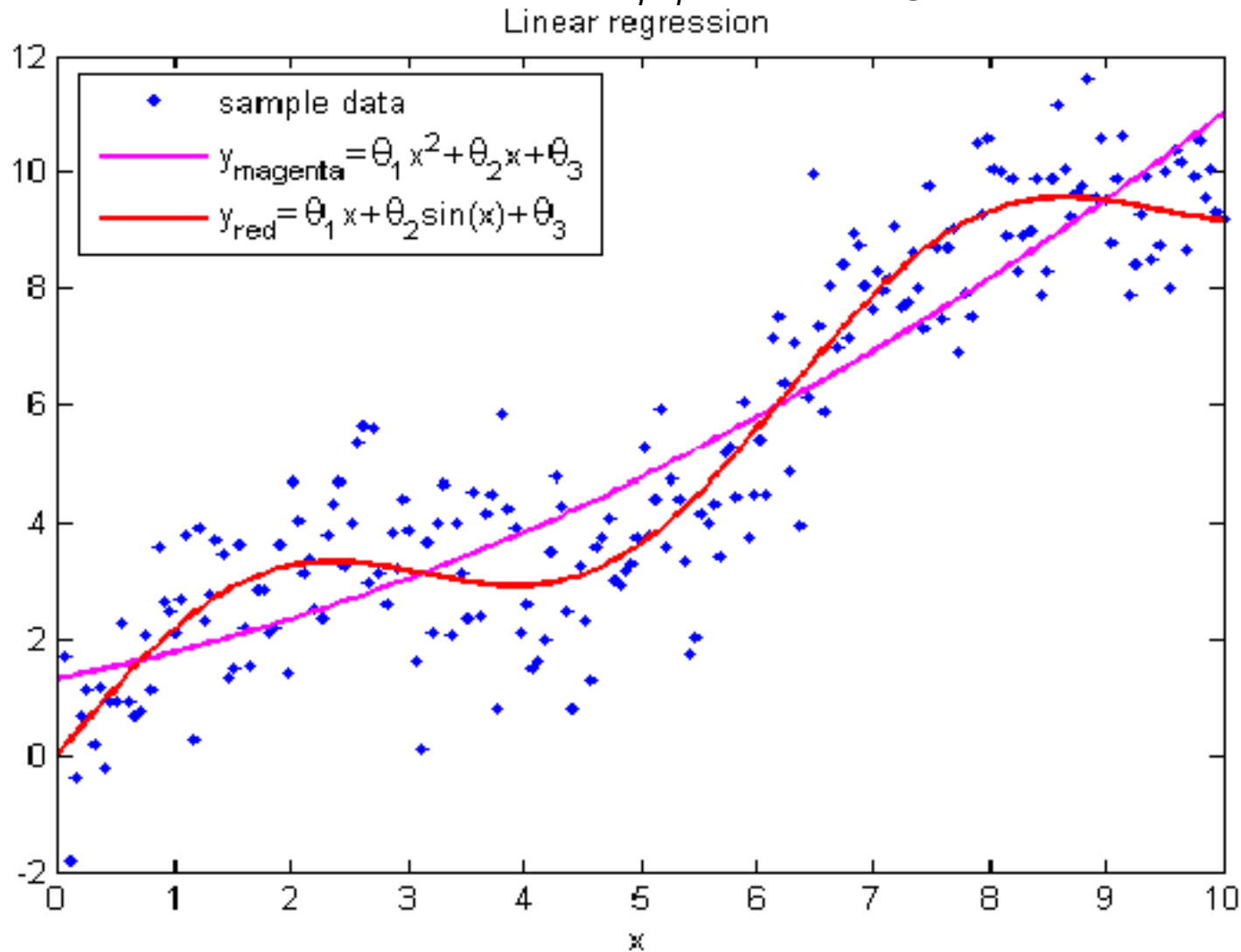
# Пример - регрессия

- $y$  - цена автомобиля
- $x$  - пробег
- $y = \theta_1 x + \theta_0$  - модель
- $\theta_0, \theta_1$  - параметры



# Пример – две точки зрения

1.  $x$  – один признак,  $\theta_1 x^2 + \theta_2 x + \theta_3$  и  $\theta_1 x + \theta_2 \sin(x) + \theta_3$  – две модели
2.  $\{x^2, x\}$ ,  $\{x, \sin(x)\}$  – два набора разных признаков, модель – одна (линейная  $\theta_1 f_1 + \theta_2 f_2 + \theta_3$ )



# Обучение на основе минимизации эмпирического риска

- Функция потерь  $\mathcal{L}(a, x)$  - величина ошибки гипотезы  $a$  на объекте  $x$ .

Примеры:

- бинарная (где используется?)

- $\mathcal{L}(a, x) = |a(x) - y^*(x)|$

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

- Эмпирический риск:  $Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(a, x_i)$

- Самый популярный алгоритм обучения – минимизация эмпирического риска:

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

# Проблемы реальных задач

- Одинаковые признаковые описания могут соответствовать разным объектам
- Объекты с похожими (даже одинаковыми) значениями признаков могут иметь различные значения целевой функции

# Вероятностная постановка задачи

- $p(x, y)$  – неизвестная точная плотность распределения на  $X \times Y$
- $X^\ell$  - выборка из случайных, независимых и одинаково распределенных прецедентов
- $p(X^\ell) = p((x_1, y_1), \dots, (x_\ell, y_\ell)) = p(x_1, y_1) \times \dots \times p(x_\ell, y_\ell)$
- $\varphi(x, y, \theta)$  - модель
- Принцип максимума правдоподобия:

$$L(\theta, X^\ell) = \prod_{i=1}^{\ell} \varphi(x_i, y_i, \theta) \rightarrow \max_{\theta}$$



# Decision function

- Предположим, что мы нашли вероятность  $p(y|x) = p(x,y)/p(x)$ . Какое значение  $y$  нужно предсказать для заданного  $x$  ?

- Минимизация среднего риска:

$$a(x) = \arg \min_s E_y \mathcal{L}(s, y)$$

- Упражнение:

$y$	2	3	4	5
$p(y x)$	0.1	0.2	0.3	0.4

примите правильные решения  $a(x)$  для каждой функции потерь со слайда 14

# Несимметричные потери

- Пример: нужно предсказать, сколько фирма потратит на рекламу (**m** - мало 50 т.р, **B** - много 200 т.р.)
- $\mathcal{L}(a=m, y=B) = ???$
- $\mathcal{L}(a=B, y=m) = ???$

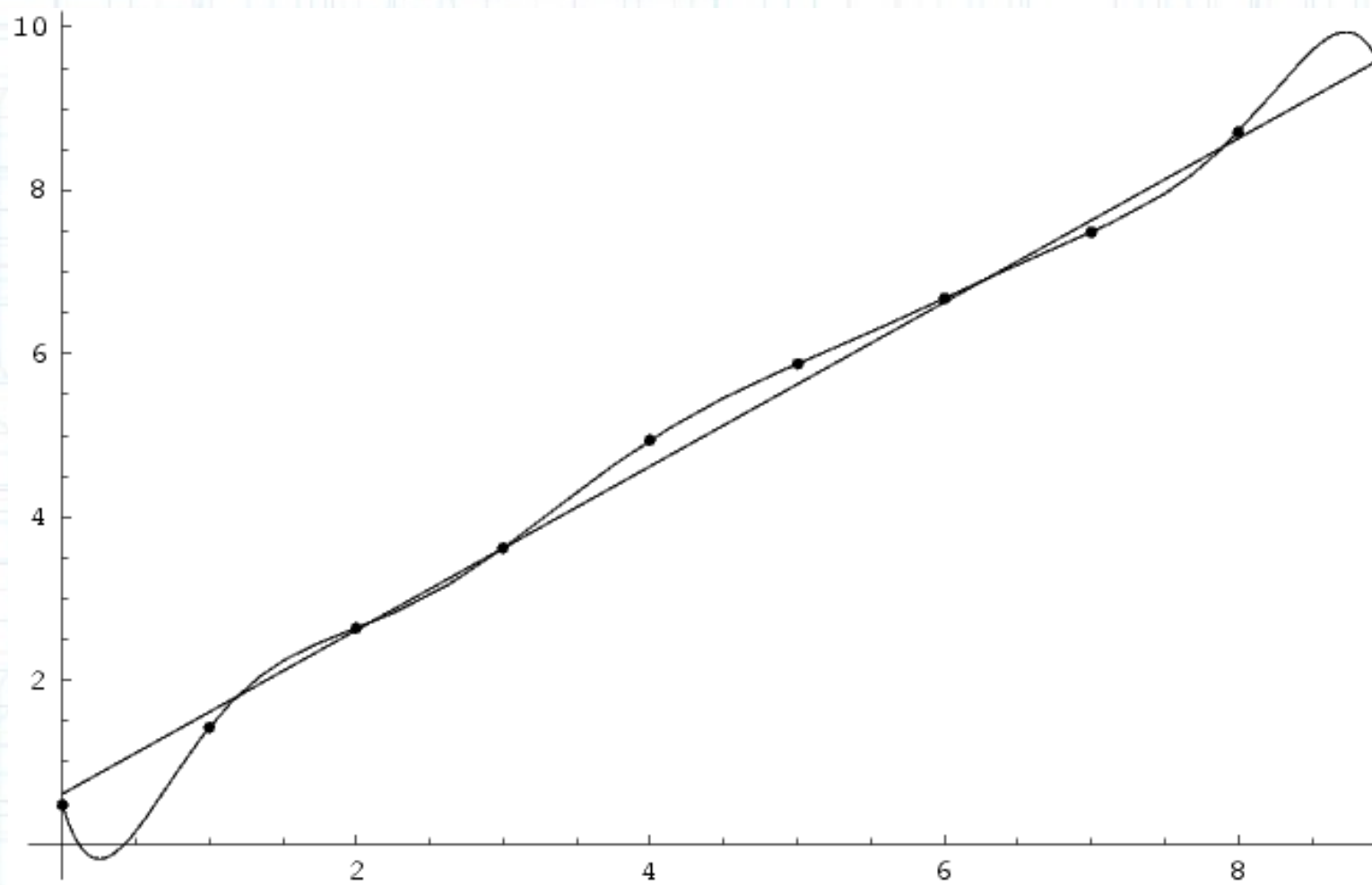
# Несимметричные потери

- Пример: нужно предсказать, сколько фирма потратит на рекламу (**m** - мало 50 т.р, **B** - много 200 т.р.)
- $\mathcal{L}(a=m, y=B)$  — не сделаем подарочную скидку и от нас могут отказаться и не заказать много рекламы  
 $\mathcal{L}(m, B) = 200\text{т.р.} * 0.8$  (вер-ть отказа)
- $\mathcal{L}(a=B, y=m)$  — сделаем большую скидку, а закажут мало рекламы по низкой цене  
 $\mathcal{L}(B, m) = 50\text{т.р.} * 0.3$  (скидка)
- Допустим, что метод МО предсказал  $p(m|x) = 0.9$     $p(B|x) = 0.1$   
Скидку делать?

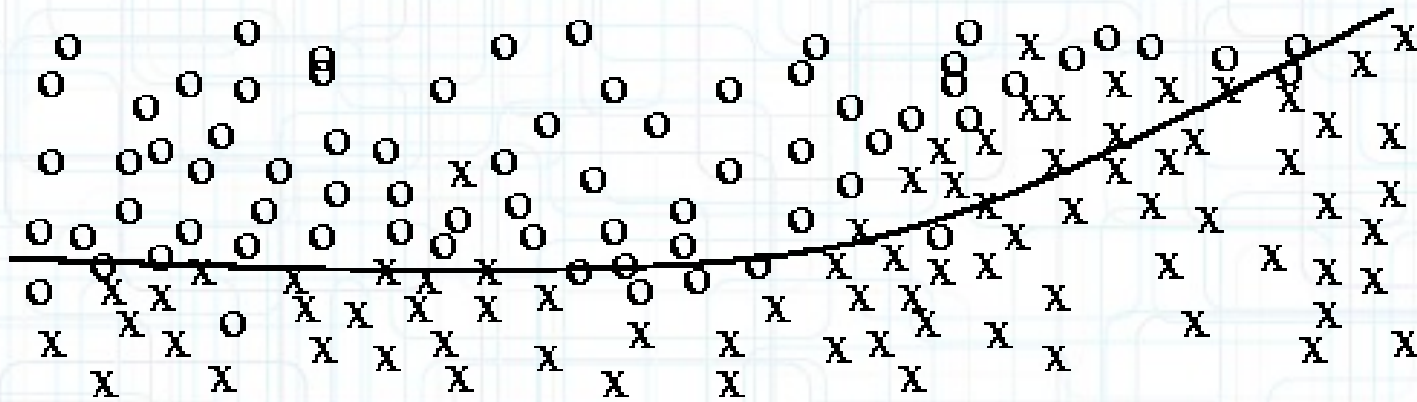
# Степени обученности модели

- Недообученная модель
  - Модель, слишком сильно упрощающая закономерность  $X \rightarrow Y$ .
- Переобученная модель
  - Модель, слишком сильно настроенная на особенности обучающей выборки (на шум в наблюдениях), а не на реальную закономерность  $X \rightarrow Y$ .

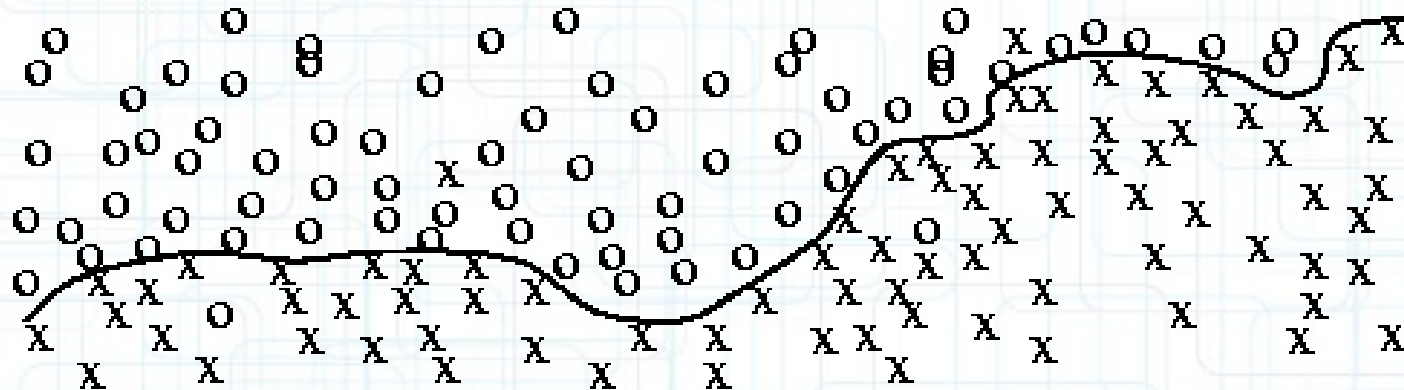
# Переобучение



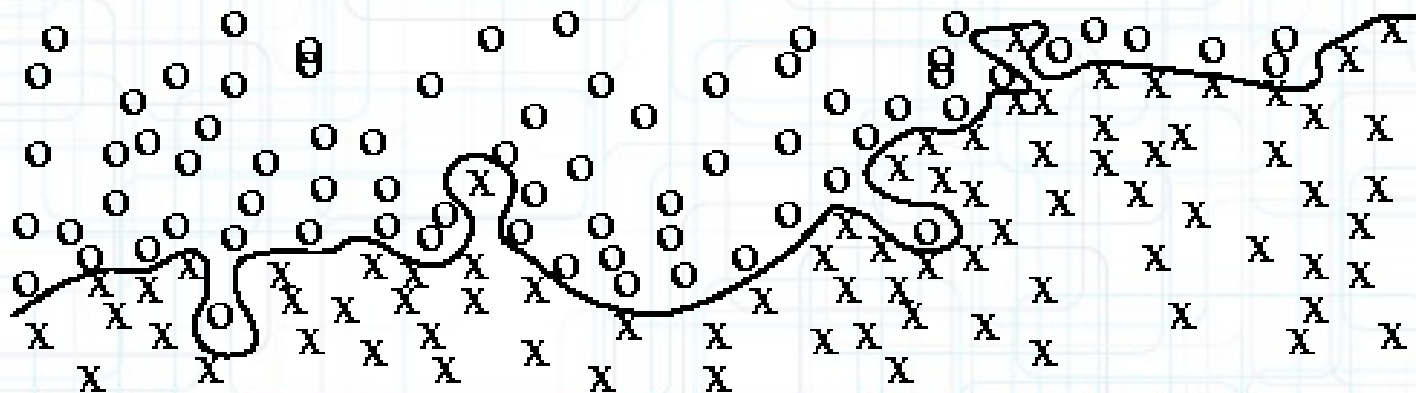
# Переобучение



Under-Trained

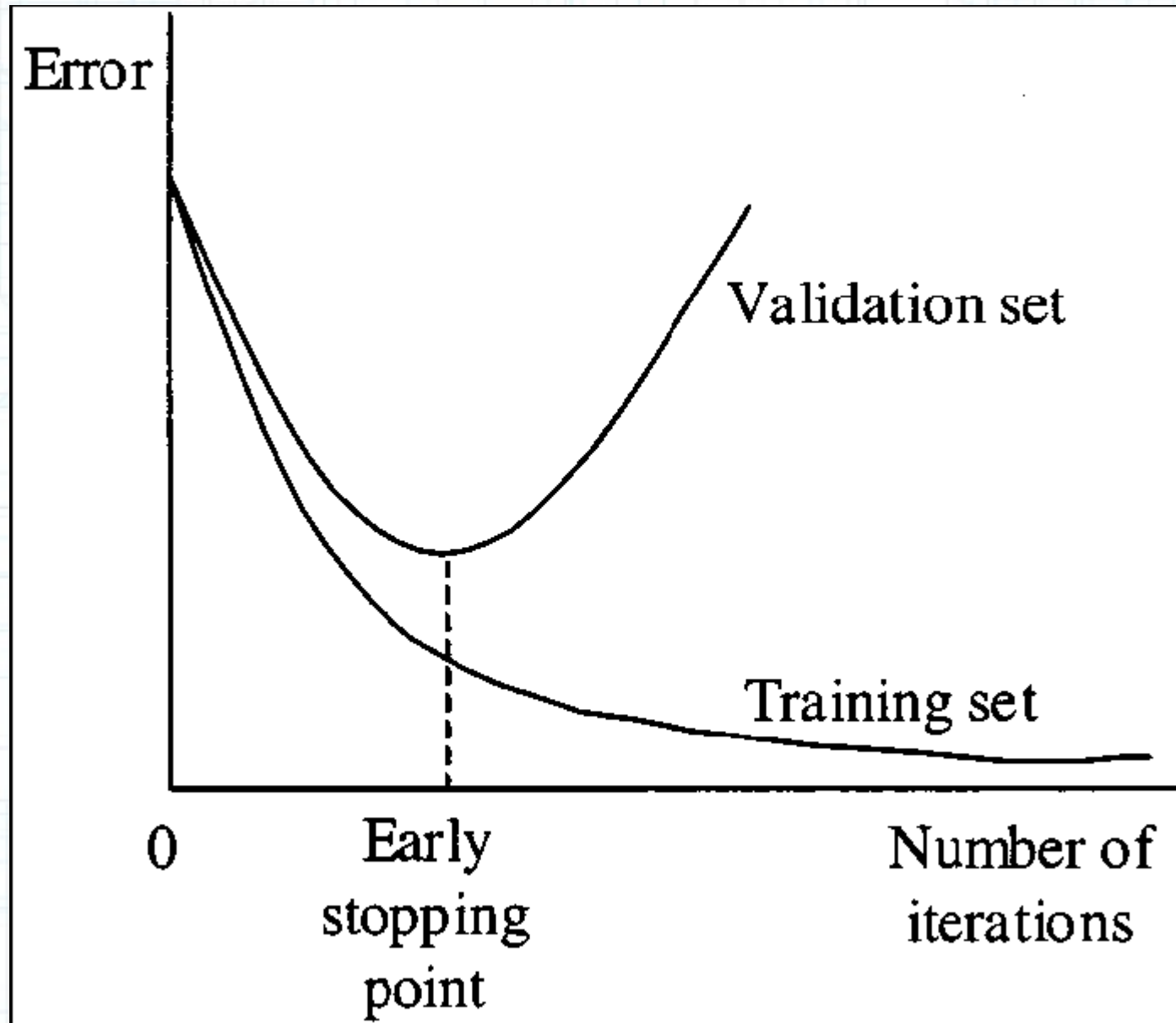


Well-Trained



Overfitted

# Когда нужно заканчивать обучение?



# Контроль переобучения

- Для оценки обобщающей способности алгоритма обучения  $\mu$  используют:
  - Эмпирический риск на тестовых данных (hold-out):

$$\text{HO}(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \rightarrow \min$$

- Скользящий контроль (leave-one-out),  $L=l+1$ :

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

- Кросс-проверка (cross-validation):

$$\text{CV}(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} Q(\mu(X_n^\ell), X_n^k) \rightarrow \min$$

- Оценка вероятности переобучения:

$$Q_\varepsilon(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} \left[ Q(\mu(X_n^\ell), X_n^k) - Q(\mu(X_n^\ell), X_n^\ell) \geq \varepsilon \right] \rightarrow \min$$