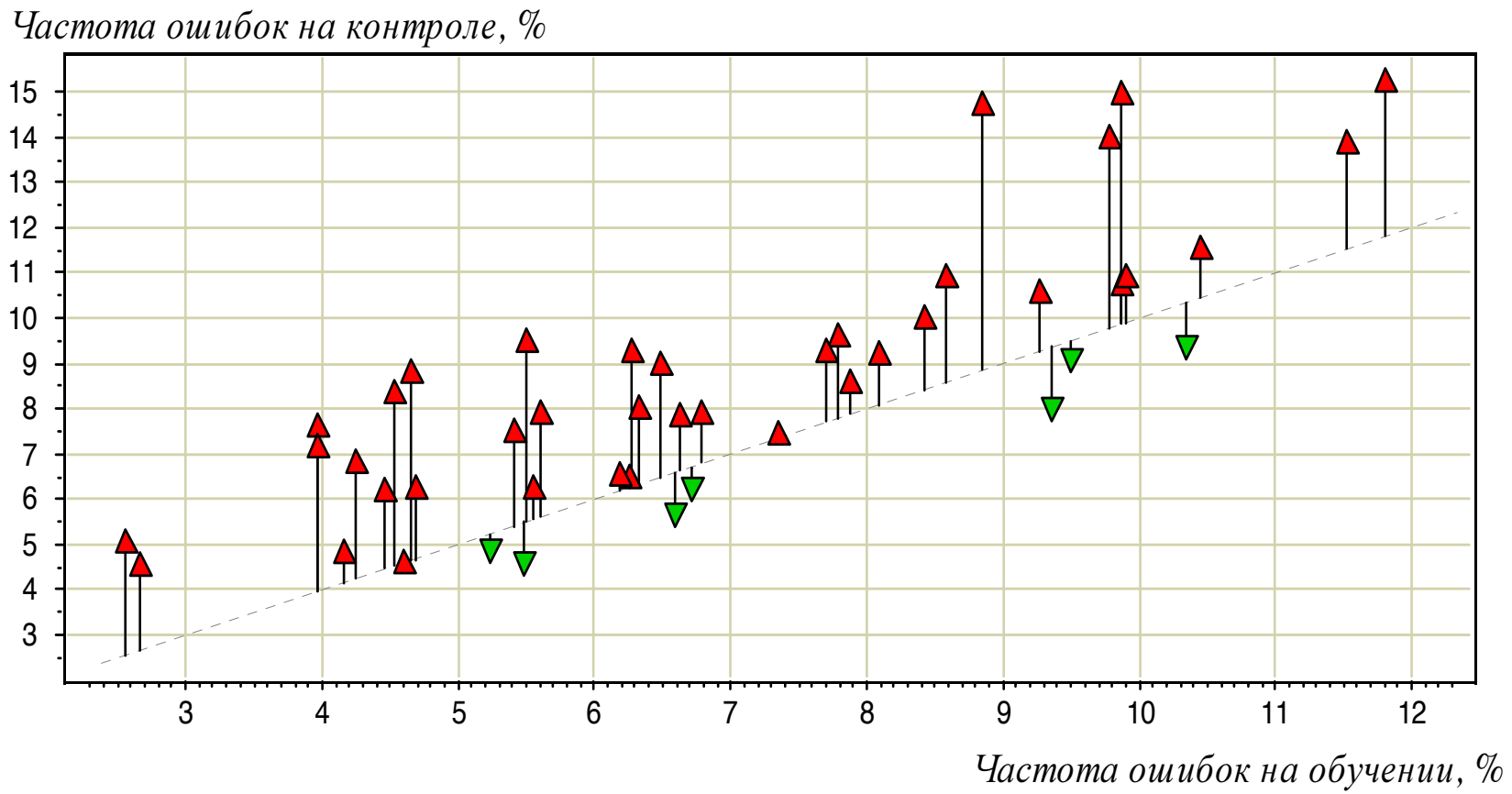
Пример. Переобучение в задаче медицинской диагностики

Задача предсказания отдалённого результата хирургического лечения атеросклероза. Точки — различные алгоритмы.



Пример: переобучение полиномиальной регрессии

Зависимость $y(x) = \dfrac{1}{1+25x^2}$ на отрезке $x \in [-2, 2]$.

Признаковое описание $x \mapsto (1, x^1, x^2, \ldots, x^n)$.

Модель полиномиальной регрессии

$$a(x, \theta) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n \quad \text{— полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \cdots + \theta_n x_i^n - y_i)^2 \to \min_{\theta_0, \ldots, \theta_n}.$$

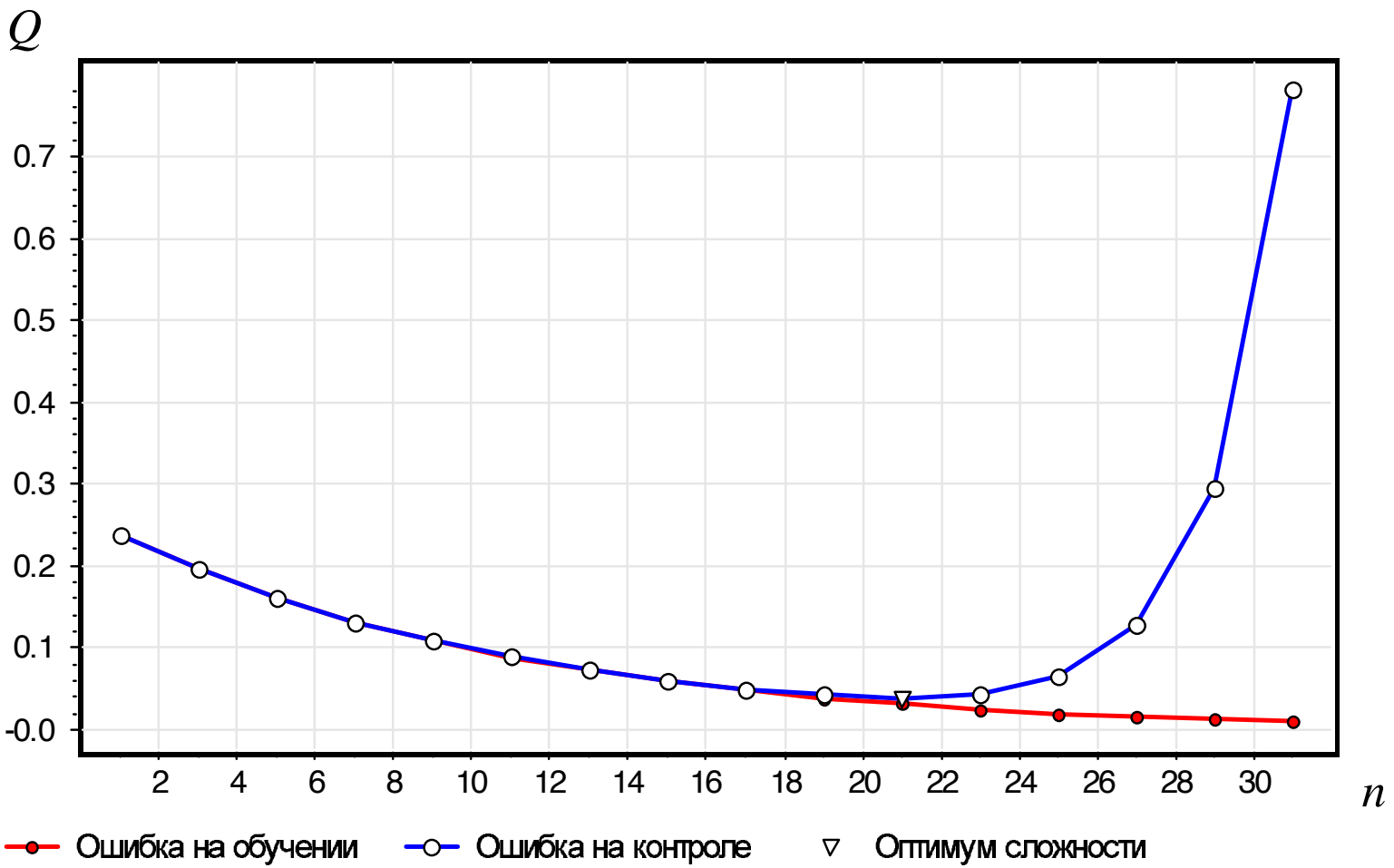
Обучающая выборка: $X^\ell = \left\{ x_i = 4\frac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell \right\}$.

Контрольная выборка: $X^k = \left\{ x_i = 4\frac{i-0.5}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell - 1 \right\}$.

Что происходит с $Q\left(a, X^\ell\right)$ и $Q\left(a, X^k\right)$ при увеличении n?

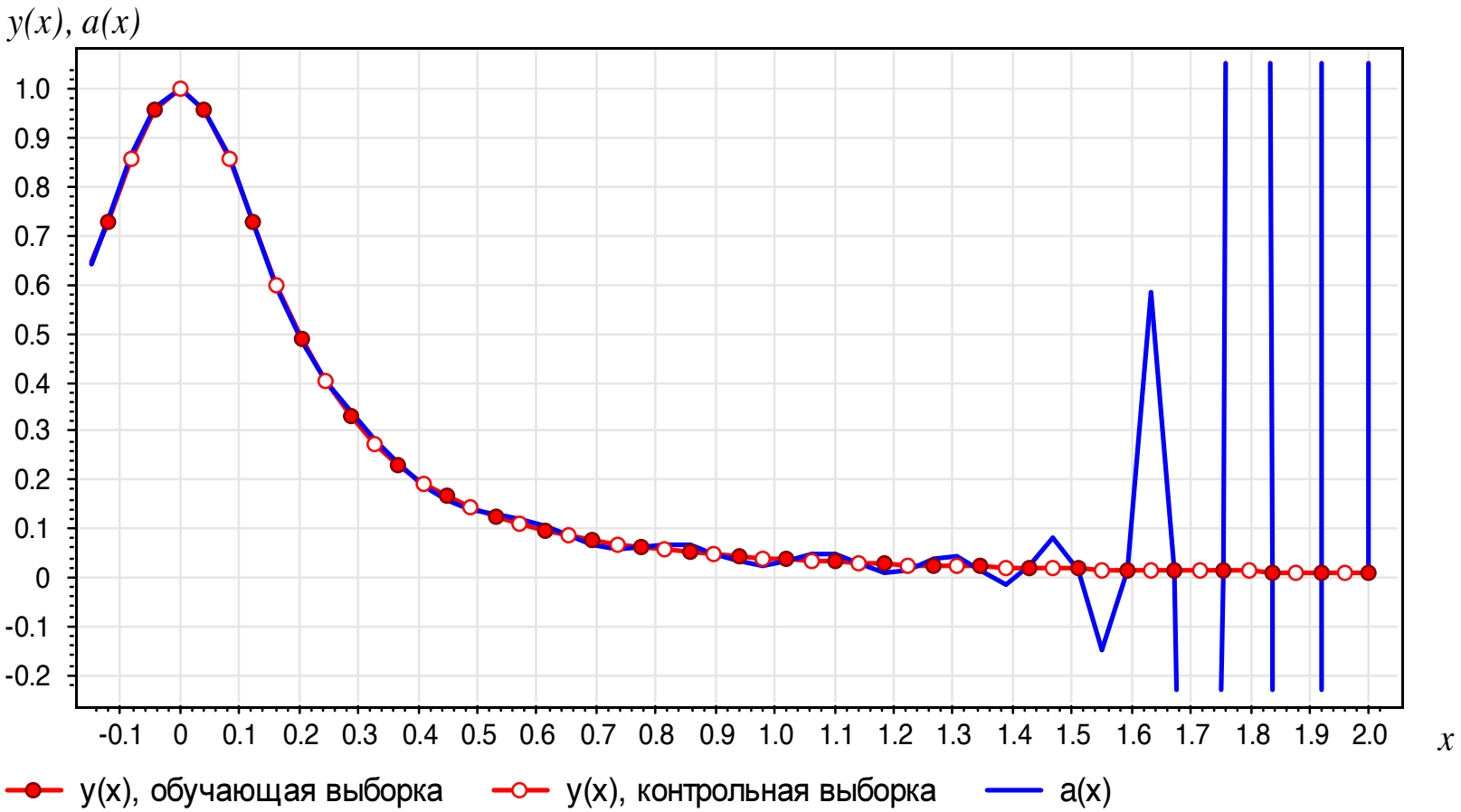
Пример переобучения: эксперимент при $\ell = 50$, $n = 1, \ldots, 31$

Переобучение — это когда $Q\big(\mu(X^\ell), X^k\big) \gg Q\big(\mu(X^\ell), X^\ell\big)$:



Пример переобучения: эксперимент при $\ell = 50$, $n = 38$

$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) \text{ — полином степени } n = 38$$



Эмпирические оценки обобщающей способности

- Эмпирический риск на тестовых данных (hold-out):

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

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

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

- Кросс-проверка (cross-validation) по N разбиениям, $X^L = X_n^\ell \sqcup X_n^k$, $L = \ell + k$:

$$\mathrm{CV}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} Q(\mu(X_n^\ell), X_n^k) \to \min$$

Эксперименты на реальных данных

Эксперименты на конкретной прикладной задаче:

- цель — решить задачу как можно лучше
- важно понимание задачи и данных
- важно придумывать информативные признаки
- конкурсы по анализу данных: http://www.kaggle.com

Эксперименты на наборах прикладных задач:

- цель — протестировать метод в разнообразных условиях
- нет необходимости (и времени) разбираться в сути задач :(
- признаки, как правило, уже кем-то придуманы
- репозиторий UC Irvine Machine Learning Repository http://archive.ics.uci.edu/ml (308 задач, 09-02-2015)

Эксперименты на модельных (синтетических) данных

Используются для тестирования новых методов обучения.
Преимущество — мы знаем истинную $y(x)$ (ground truth)

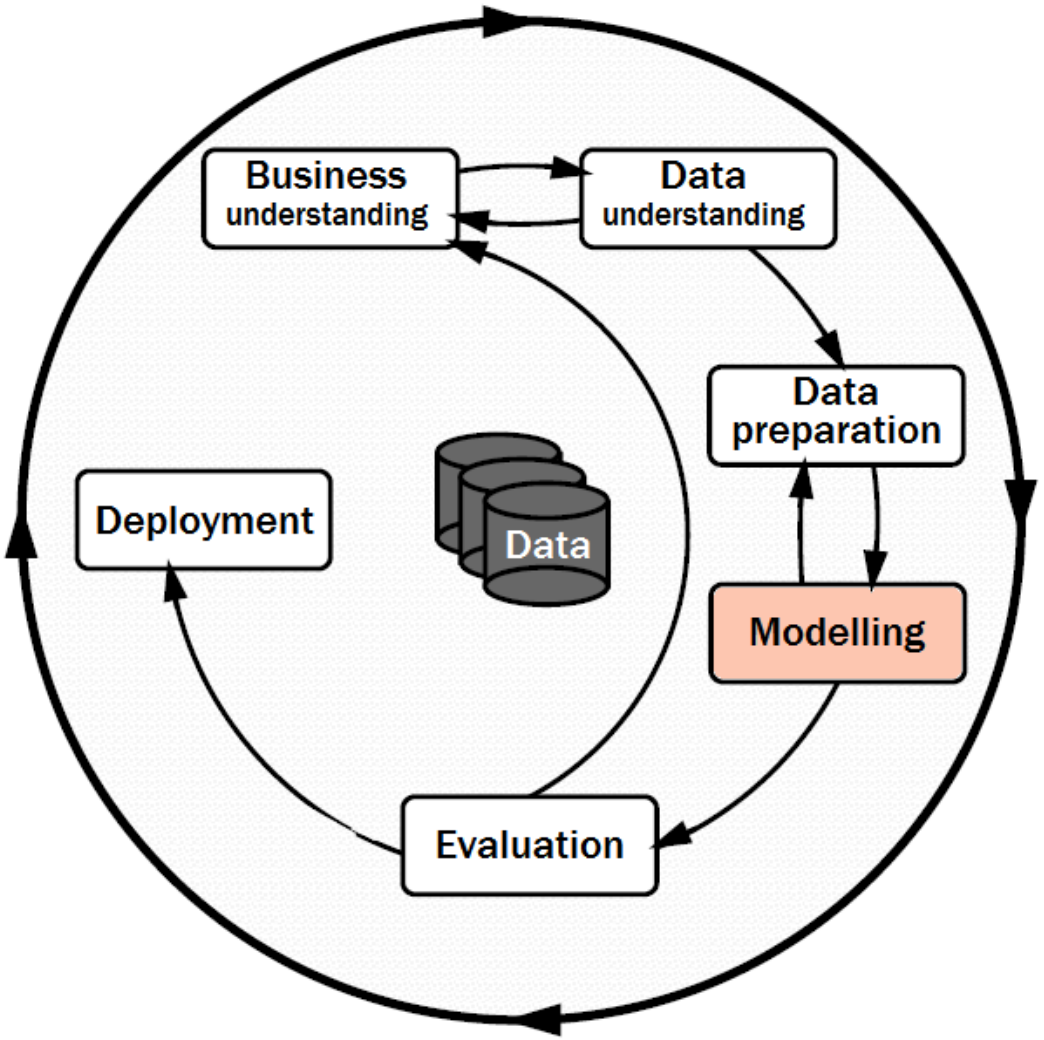
Эксперименты на модельных (synthetic) данных:

- цель — отладить метод, выявить границы применимости
- объекты x_i из придуманного распределения (часто 2D)
- ответы $y_i = y(x_i)$ для придуманной функции $y(x)$
- двумерные данные + визуализация выборки

Эксперименты на полумодельных (semi-synthetic) данных:

- цель — протестировать помехоустойчивость модели
- объекты x_i из реальной задачи (+ шум)
- ответы $y_i = a(x_i)$ для полученного решения $a(x)$ (+ шум)

CRISP-DM — межотраслевой стандарт решения задач интеллектуального анализа данных



Резюме

Этапы решения задач машинного обучения:

- понимание задачи и данных;
- предобработка данных и изобретение признаков;
- построение модели;
- сведение обучения к оптимизации;
- решение проблем оптимизации и переобучения;
- оценивание качества решения;
- внедрение и эксплуатация.