Машинное обучение в гидрометеорологии

Лекция №6.

Оценка неопределенностей в машинном обучении.

Кросс-валидация.

Михаил Иванович Варенцов (mikhail.varentsov@srcc.msu.ru)

Михаил Алексеевич Криницкий (krinitsky@sail.msk.ru)

ml4hydromet@ml4es.ru

ИНТЕЛЛЕКТ

ФОНД РАЗВИТИЯ НАУКИ И ОБРАЗОВАНИЯ

Ранее в ML4hydromet…

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

1. формулировка задачи:
   - какой тип (классификация, регрессия, другой)? Или переформулировать в легко решаемый тип!
   - определиться, что есть объекты (события)
   - определиться, что есть целевая переменная
   - определить признаковое описание объектов (событий)
   - определить критерии качества решения задачи (MSE, MAE, pattern correlation, etc.)
2. сформулировать модель:
   - вид модели (линейная регрессия, дерево решений, композиционный алгоритм, нейронная сеть, etc.)
   - определиться с функцией потерь (MSE, MAE, BCE, CCE, etc., комбинации)
   - сложность модели (задается гиперпараметрами – настройками модели)

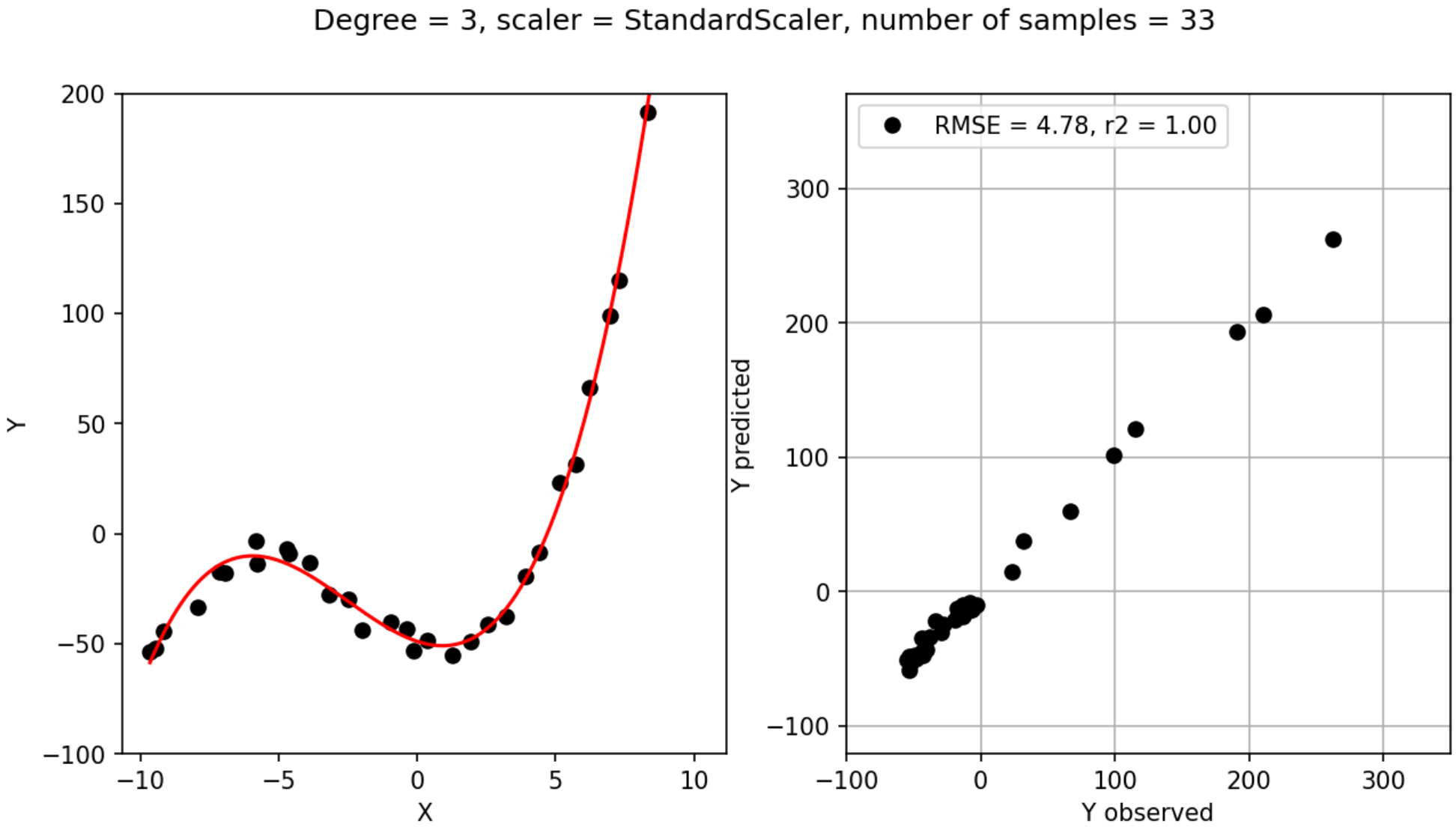
ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

3. подготовить данные или генератор данных:
   - стандартизировать данные (если нужно)
   - обработка пропусков, категориальных значений, кодирование текста, понижение размерности данных
   - оставить часть данных для проверки качества (train-validation-test split)
   - подготовить генератор данных с учетом стратегии скользящего контроля (cross-validation quality estimation)
4. оптимизировать модель на обучающей выборке:
   - $\hat{p} = \underset{\mathbb{P}}{\operatorname{argmin}}\big(L(\vec{p}, \mathcal{T})\big)$
5. оптимизация гиперпараметров модели и отбор моделей. Провизодится по значениям метрик качества на контрольной(контрольных) выборке(выборках)
6. оценка модели:
   - оценить качество по метрикам, определенным на этапе 1. на тестовой выборке
   - оценить неопределенность параметров модели (если возможно)
   - оценить неопределенность оценок целевой переменной

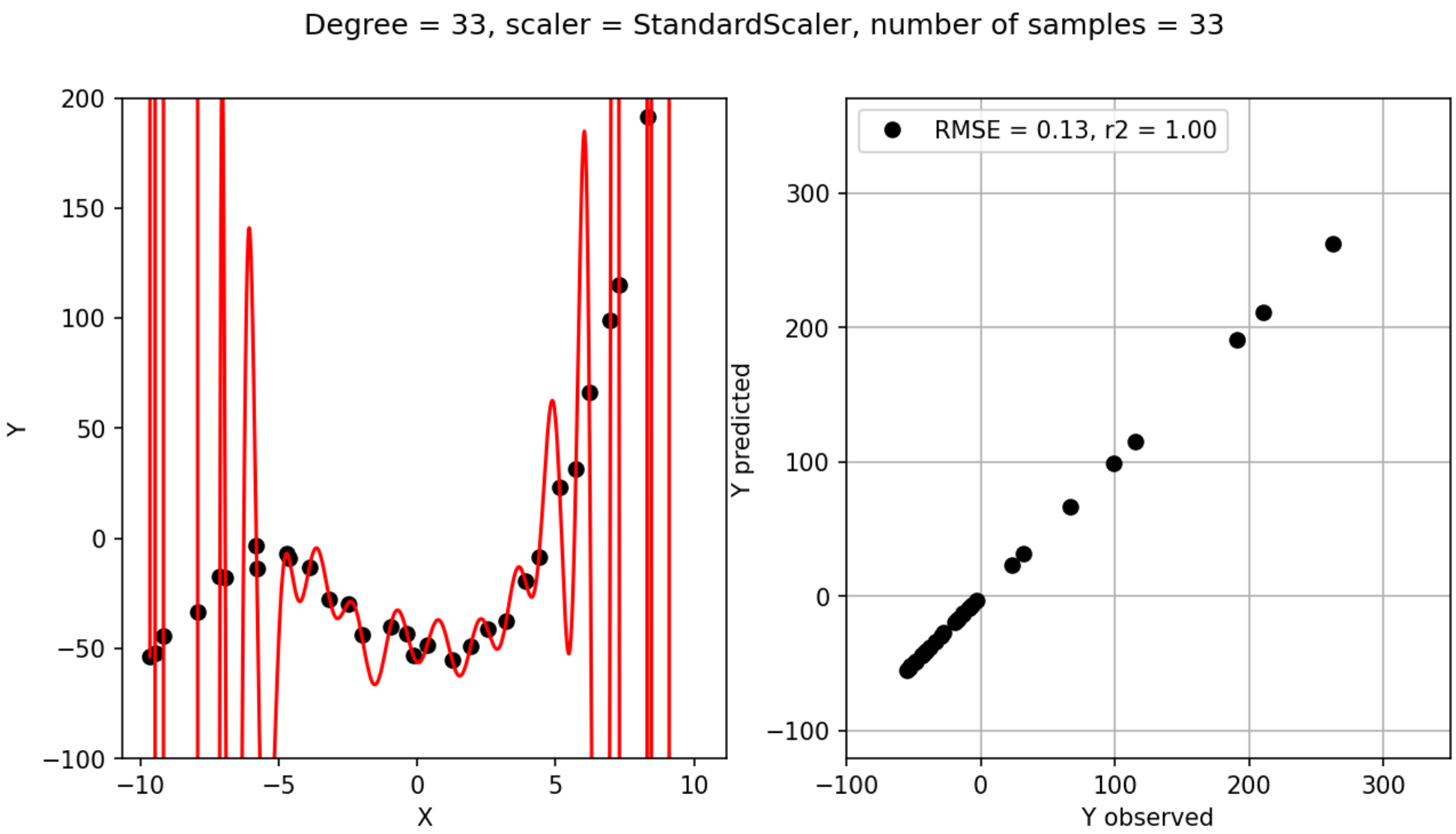
ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

6. применение модели на вновь получаемых данных:
   - оценка распределения вновь получаемых данных: генерируются ли они из того же распределения, что и обучающая выборка?
   - предобработка новых данных идентично п.3 с точностью до коэффициентов стандартизации и деталей способов предобработки
   - применение модели к предобработанным новым данным для получения значений целевой переменной
   - построение научных выводов, описание их в виде статей, получение наград в виде Нобелевских премий, etc.

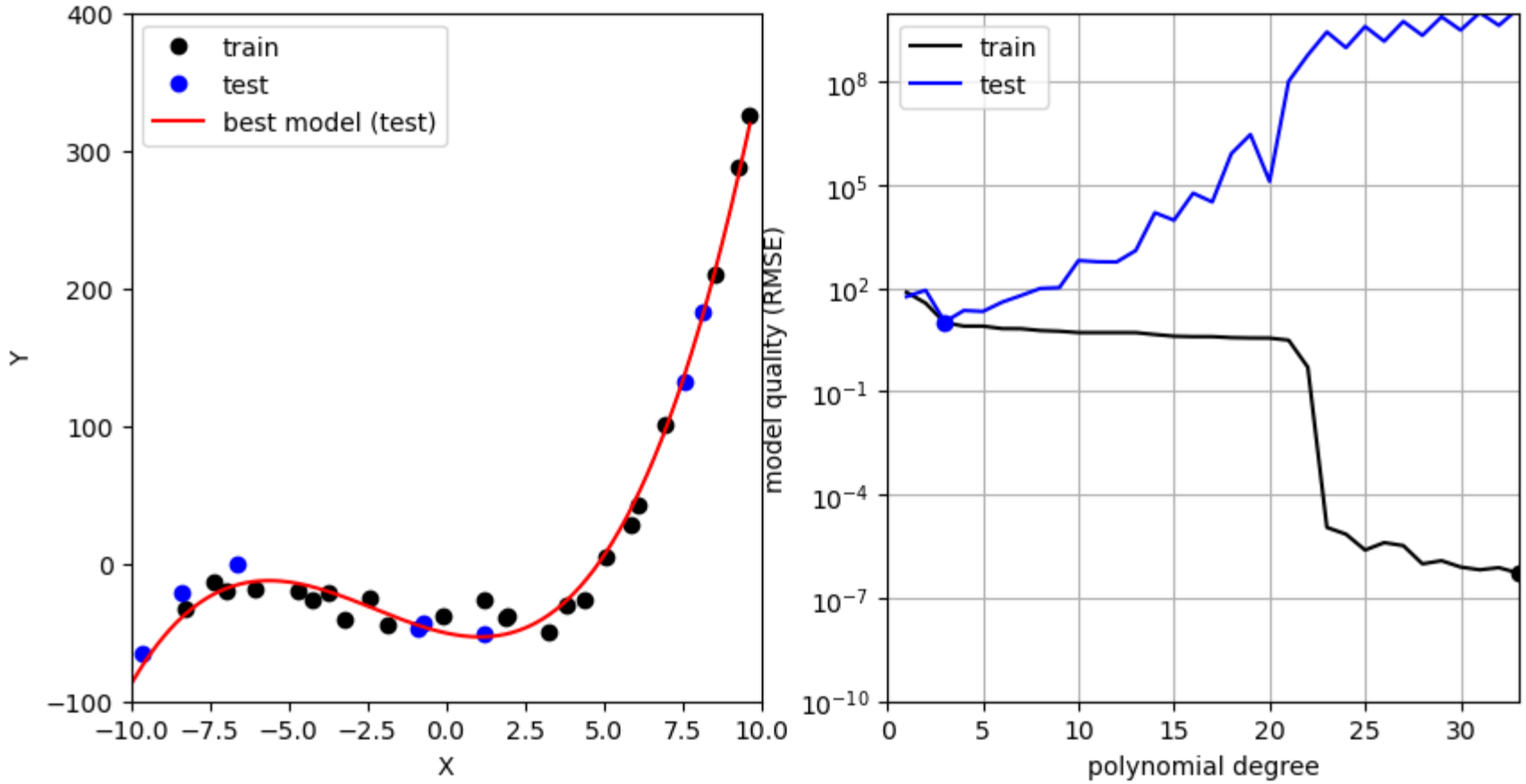
Ранее в ML4hydromet…



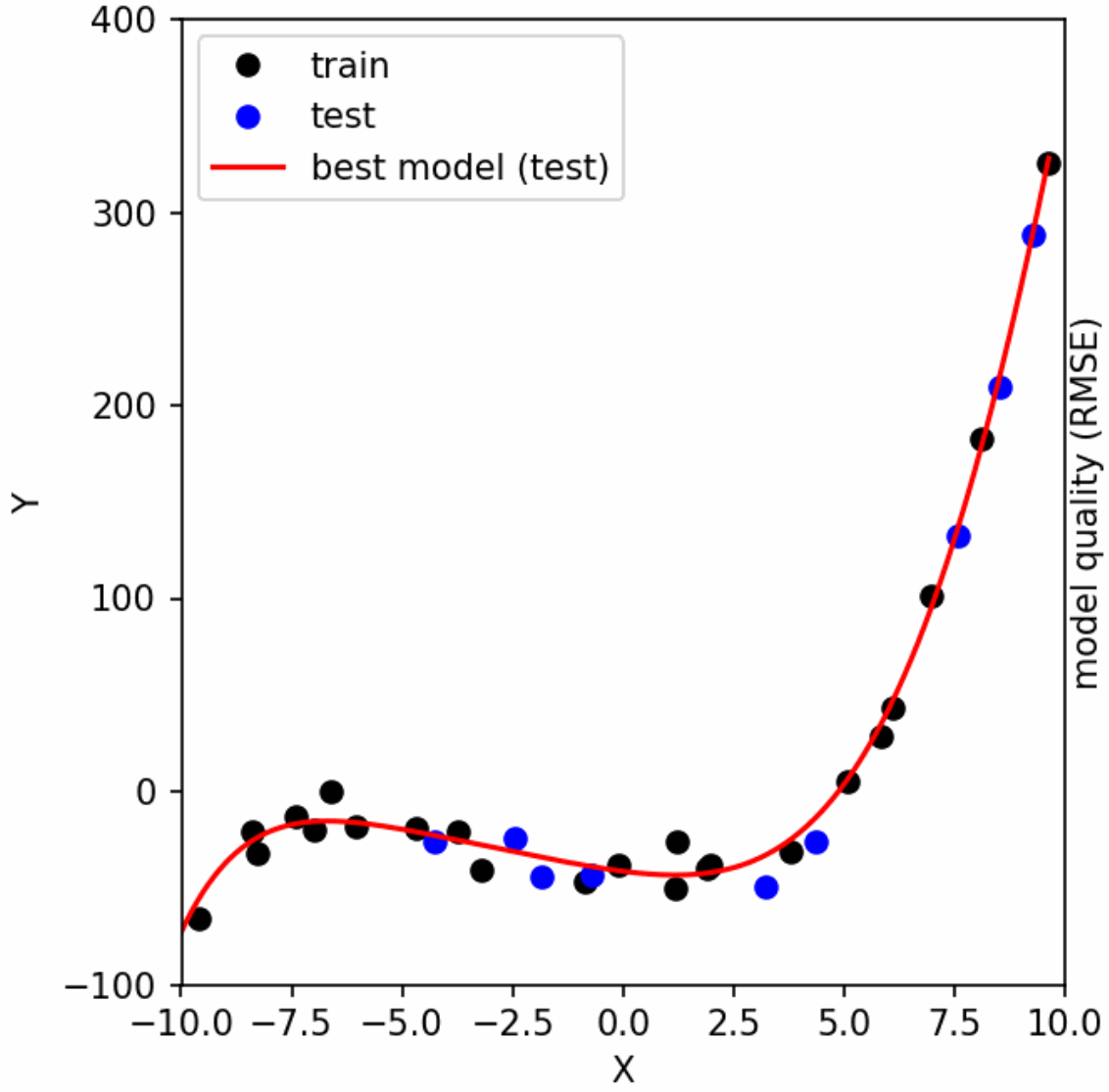
Переобучение

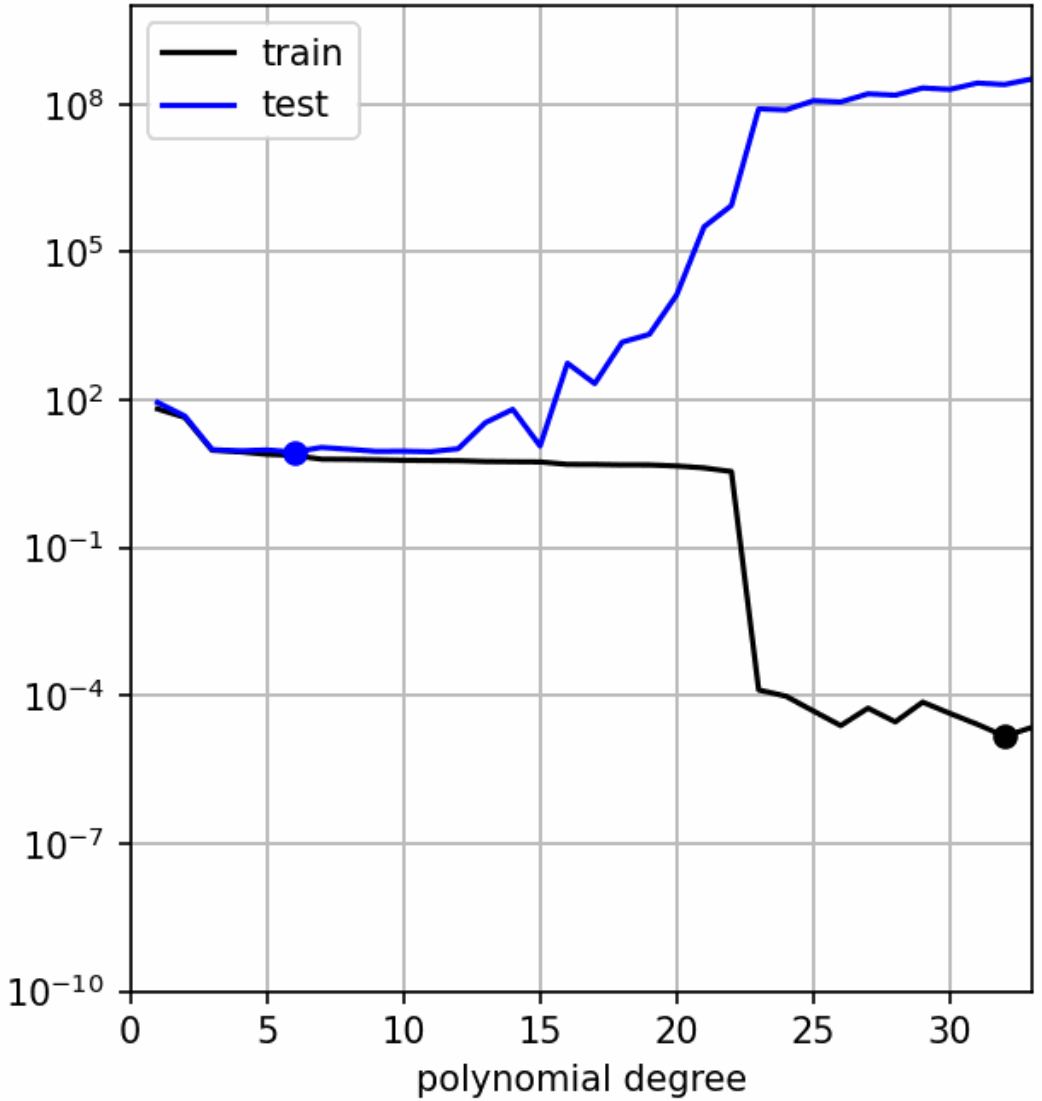


Тестовая и тренировочная выборки



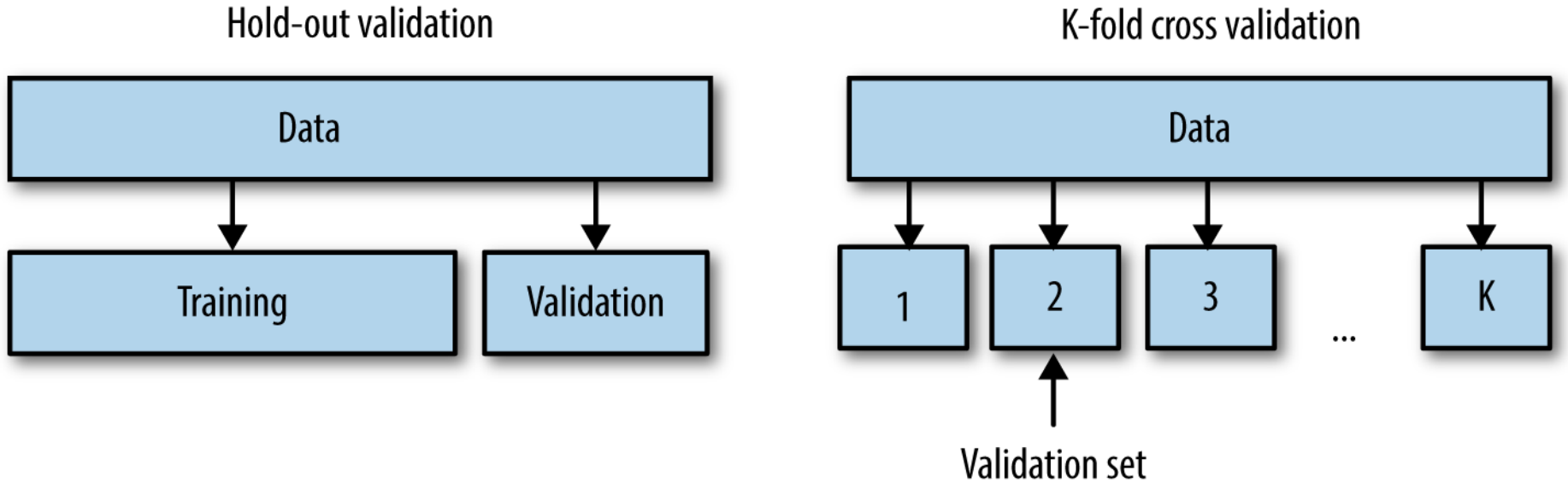
Неопределенность параметров и оценок



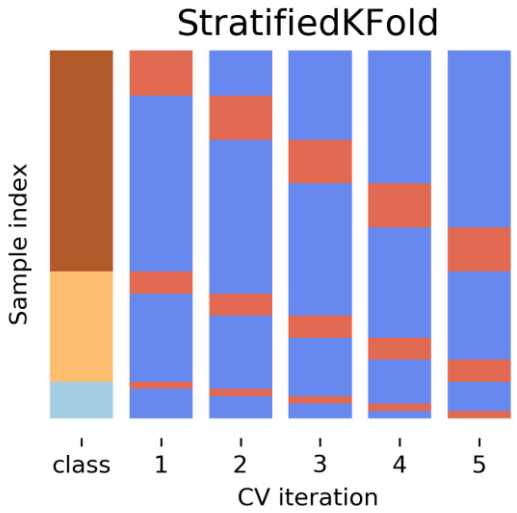


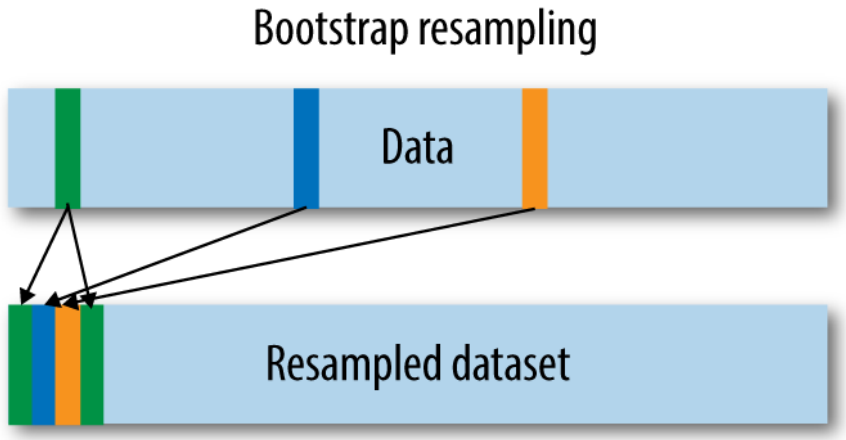
Кросс-валидация

Скользящий контроль или **кросс-проверка** или **кросс-валидация** (cross-validation, CV)



- Hold-out
- k-Fold
- Stratified k-Fold
- Blocked k-Fold
- Leave-one-out (Jacknife)
- Bootstrap
- …





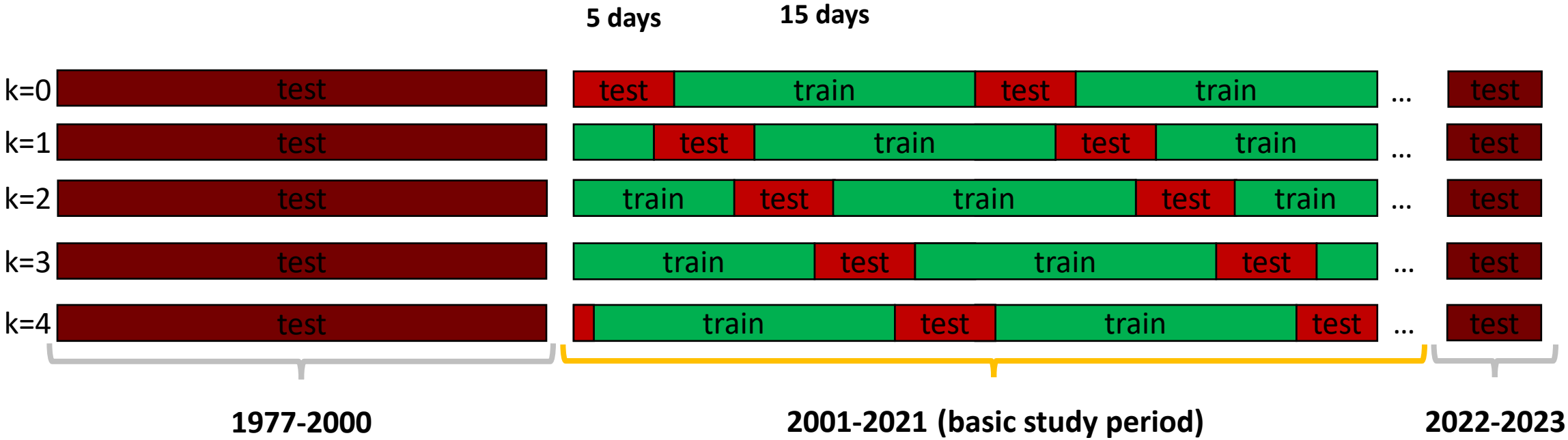
Кросс-валидация

Особенности кросс-валидации в гидрометеорологии:

- Автокорреляция (зависимость последовательных событий)
- Суточный и сезонный ход
- Нестационарность условий
  (изменения климата, антропогенное воздействие, урбанизация)
- Экстремальные события
- **…**

Пример: аппроксимация острова тепла

- **The dataset:**
  - time series with 3-hour step for 47 years (1977-2021)
  - 21 years (2001-2021) are used to train the models and for the major part of analysis
- **Train-test split using blocked k-fold method with 5-days blocks and 4:1 train-to-test ratio**
- **Model quality metrics: RMSE, ME, R^2**



Varentsov M., Krinitskiy M., Stepanenko V. Machine learning for simulation of urban heat island dynamics based on large-scale meteorological conditions // *CLIMATE*. — 2023. — Vol. 11, no. 10. — P. 200.

Домашнее задание №5

- Внедрить подход кросс-валидации в решение выбранной вами задачи машинного обучения на примере модели линейной регрессии.
- Оценить неопределенность:
  - параметров модели (наиболее важных)
  - метрик качества
  - результатов расчета целевой переменной
- Сравнить метрики качества при использовании как минимум двух разных методов кросс-валидации (например, hold-out и Blocked k-Fold). Сделать выводы о том, какой их них больше подходит к выбранной задаче.
- С помощью кросс-валидации исследовать влияние на качество модели следующих этапов подготовки данных:
  - Масштабирование (стандартизации/нормализация)
  - Порождение новых признаков
- При необходимости: реализовать способ разбиения на train-test с размером блока, заданным в единицах времени