Машинное обучение

Лекция 6:
Модели классического машинного обучения

Докладчик: Тима Бовт

Что рассмотрим сегодня?

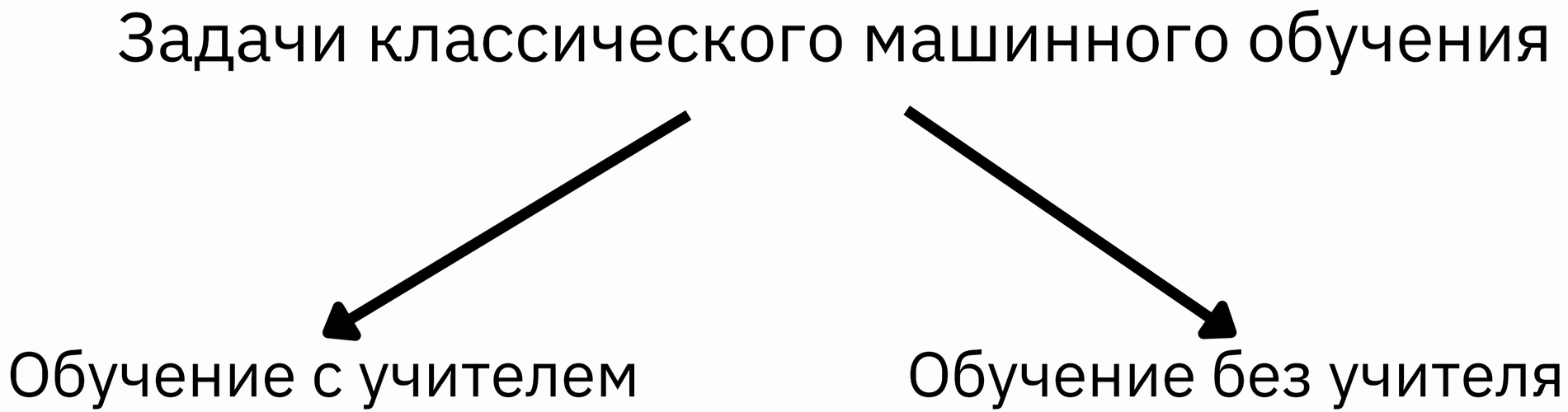
- Постановка задачи машинного обучения
- Задание признаков и ответов
- Модель машинного обучения
- Методы обучения моделей, эмпирический риск
- Недообучение и переобучение
- Оценки обобщающей способности

Задачи машинного обучения

Формулировка задачи МО

Сперва аналитик получает от заказчика реальную задачу. Для решения реальной задачи аналитик либо собирает, либо получает от заказчика данные. Поэтому наш основной объект исследования – данные. Далее аналитик сводит реальную задачу к математическим формулировкам. Таким образом, вместо реальной физической задачи решается математическая задача.

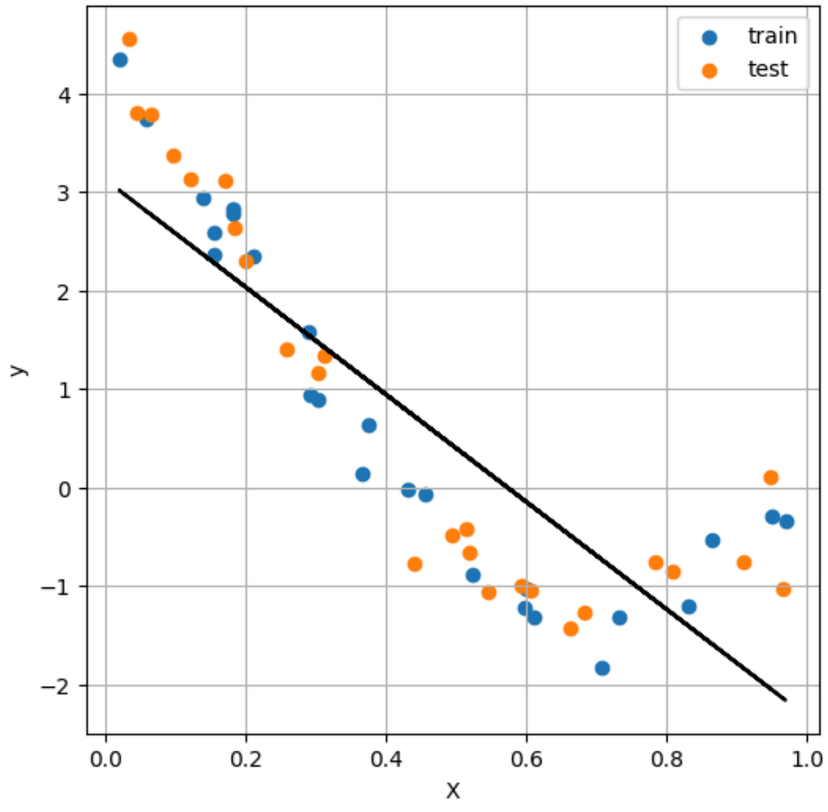
Задачи классического МО



Обучение с учителем

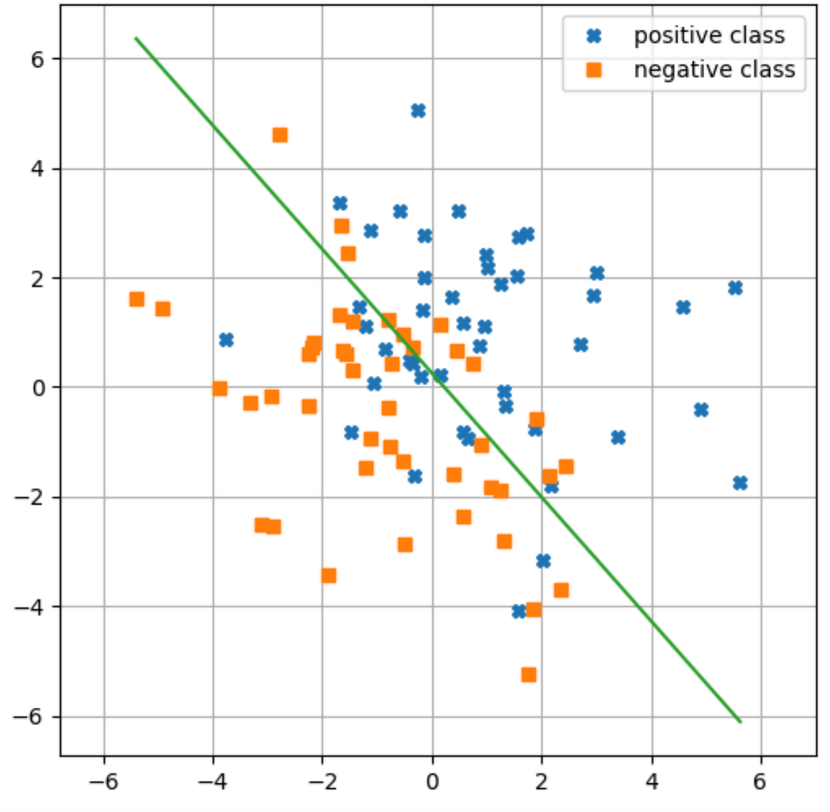
Регрессия

Прогнозирование числового значения



Классификация

Прогнозирование метки класса



Обучение без учителя

- *кластеризация* – разделение данных на группы;
- *поиск ассоциативных правил* – отыскание зависимых событий;
- *понижение размерности* – сжатие данных при разумной потере информации.

На то, к какой группе относится та или иная задача машинного обучения, влияют входные данные.

Постановка задачи

Сформулируем следующую математическую задачу. Пусть заданы два множества:

- $\mathbb{X}$ – пространство объектов;
- $\mathbb{Y}$ – множество ответов (предсказаний/оценок),

между этими множествами существует некоторая истинная зависимость

$$a : \mathbb{X} \to \mathbb{Y}.$$

Пусть дано подмножество $X = \{x_1, \ldots, x_n\} \subset \mathbb{X}$, которое мы будем называть *обучающей выборкой*, и подмножество $Y = \{y_1, \ldots, y_n\} = \{a(x_1), \ldots, a(x_n)\} \subset \mathbb{Y}$, которое мы будем называть *известными ответами*. Необходимо найти такую функцию $f : \mathbb{X} \to \mathbb{Y}$, что она приближает истинную зависимость a наилучшим образом на множестве $\mathbb{X}$.

Матрица объект-признак

Координаты вектора $x_i = (x_{i1}, \ldots, x_{im})$ называются признаками (фичами) объекта $x_i \in X$.
Таким образом, вводится понятие *матрицы «объект-признак»*:

$$X = (x_{ij})_{n \times m} = \begin{pmatrix} x_{11} & \ldots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nm} \end{pmatrix}$$

Варианты задания признаков

В итоге исходная обучающая выборка задается матрицей. Какие значения могут принимать элементы этой матрицы? Рассмотрим основные типы признаков. Пусть x_{ij} – j-ый признак i-ого объекта, $j = \overline{1, m}$. Тогда

- $x_{ij} \in \{0, 1\}$ – **бинарный** признак;
- $|x_{ij}| < \infty$ и определена операция «=» – **категориальный** признак;
- $|x_{ij}| < \infty$ и определены операции «=», «<», «>» – **порядковый** (ранговый) признак;
- $x_{ij} \in \mathbb{R}$ – **количественный** признак.

Тип задачи и множество Y

Как было сказано ранее, на тип решаемой задачи машинного обучения влияет вид входных данных. В частности, вид множества ответов как раз и определяет тип решаемой задачи. Для задач обучения с учителем

1. Классификация
   - $\mathbb{Y} = \{0, 1\}$ или $\mathbb{Y} = \{-1, 1\}$ – бинарная классификация;
   - $\mathbb{Y} = \{1, \ldots, L\}$ – классификация на L непересекающихся классов;
   - $\mathbb{Y} = \{0, 1\}^L$ – классификация на L пересекающихся классов.
2. Регрессия $\mathbb{Y} = \mathbb{R}$ или $\mathbb{Y} = \mathbb{R}^L$ – результат регрессии – это L числовых значений.

В задачах обучения без учителя множество объектов $\mathbb{Y}$ не задано, требуется проводить операции над пространством объектов.

Модель машинного обучения

Модель машинного обучения

Алгоритм, или модель, машинного обучения – это параметрическое семейство функций

$$\mathbb{F} = \{f(x, w) \mid w \in \mathbb{W}\}$$

где $f : \mathbb{X} \times \mathbb{W} \to \mathbb{Y}$ – это фиксированная функция, $\mathbb{W}$ – множество допустимых значений параметра w.

Таким образом, для решения задачи машинного обучения мы выбираем такую функцию $f \in \mathbb{F}$ и такие параметры $w \in \mathbb{W}$, что полученная модель f с параметрами w наилучшим образом приближает истинную зависимость на основании заданных множеств $\mathbb{X}$, $\mathbb{Y}$.

Семейство линейных моделей

Например, выделяют *семейство линейных моделей* с векторным параметром $w = (w_0, \ldots, w_m)$

- для задач регрессии $\mathbb{Y} = \mathbb{R}$

$$f(x, w) = w_0 + \sum_{j=1}^{m} w_j x_j;$$

- для задач бинарной классификации $\mathbb{Y} = \{-1, 1\}$

$$f(x, w) = \operatorname{sgn}\left(\sum_{j=1}^{m} w_j x_j\right).$$

Методы обучения моделей

Метод обучения

Метод обучения – это отображение вида

$$\mu : (\mathbb{X} \times \mathbb{Y}) \to \mathbb{W},$$

которое произвольной конечной выборке $(X \times Y)$ строит функцию $f(x, w)$, оптимизируя параметры модели $w \in \mathbb{W}$.

На **этапе обучения** с помощью метода обучения μ определяются конкретные значения параметров для модели из выбранного семейства.

На **этапе тестирования** с помощью построенной функции $f(x, w)$ для новых объектов $X' \subset \mathbb{X}$ мы получаем новые ответы $f(x', w)$, $x' \in X'$.

Эмпирический риск

Для оценки степени «наилучшести» выбранной модели определяются специальные функционалы, называемые функционалами качества.

Эмпирический риск – функционал качества модели f на конечной выборке $X \subset \mathbb{X}$ заданный следующим образом (иногда без множителя $1/n$)

$$L(f, X, y) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f, x_i, y_i),$$

где $\mathcal{L}(f, x, y)$, называемая *функцией потерь* (loss function), – это неотрицательная функция, пропорциональная величине ошибки алгоритма $f \in \mathbb{F}$ на объекте $x \in \mathbb{X}$, если верный ответ есть $y \in \mathbb{Y}$. То есть чем сильнее $f(x, w)$ отклоняется от правильных ответов $f(x)$, тем функции потерь больше.

Варианты функции потерь

Для задач классификации обычно принято использовать индикатор ошибочного прогноза

$$\mathcal{L}(f, x, y) = [f(x) \neq y].$$

Для задач регрессии обычно принято использовать

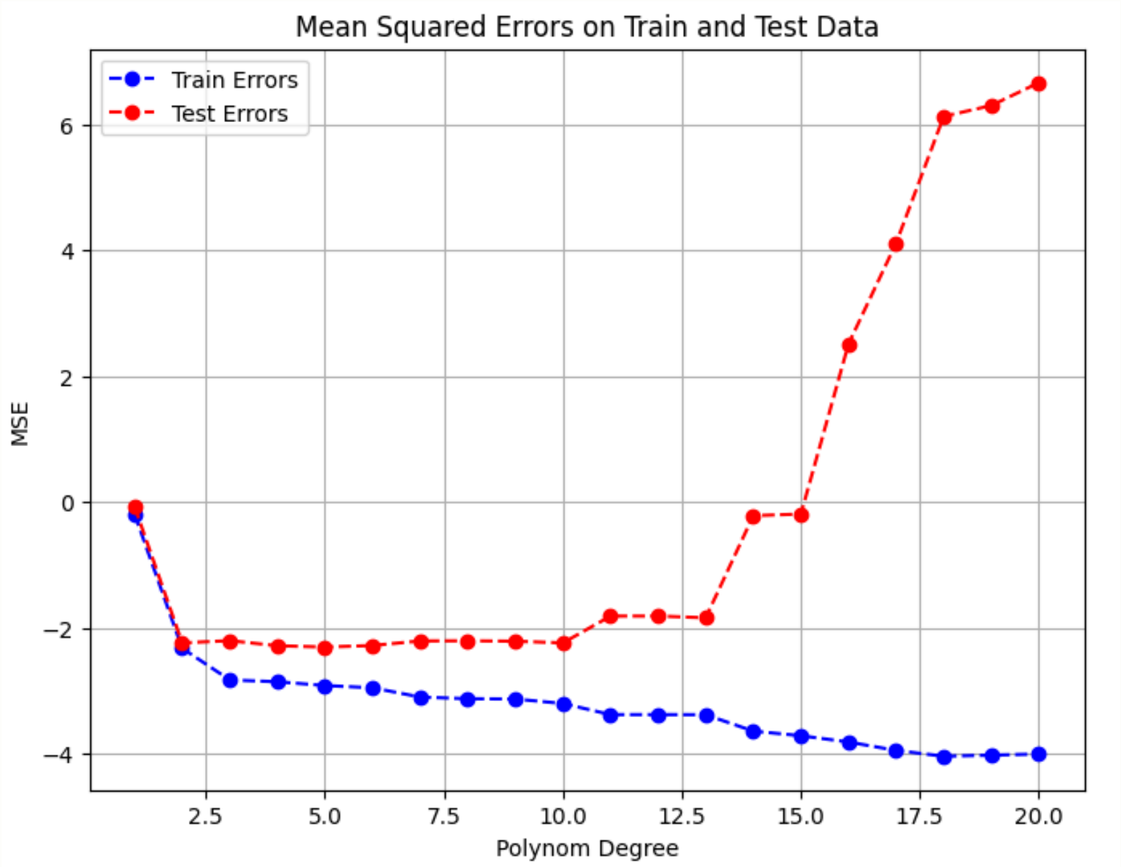
- $\mathcal{L}(f, x, y) = |f(x) - y|$ – абсолютную ошибку;
- $\mathcal{L}(f, x, y) = (f(x) - y)^2$ – квадрат ошибки;

Тогда, чтобы модель была наилучшей, она должна иметь минимальный эмпирический риск. Таким образом, задачу построения наилучшего алгоритма мы формулируем как задачу оптимизации

$$L(f, X, y) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f, x_i, y_i) \to \min_{f \in \mathbb{F}}.$$

Минимизация риска

При этом достижение минимума эмпирического риска на модели f не всегда гарантирует хорошее приближение истинной зависимости.



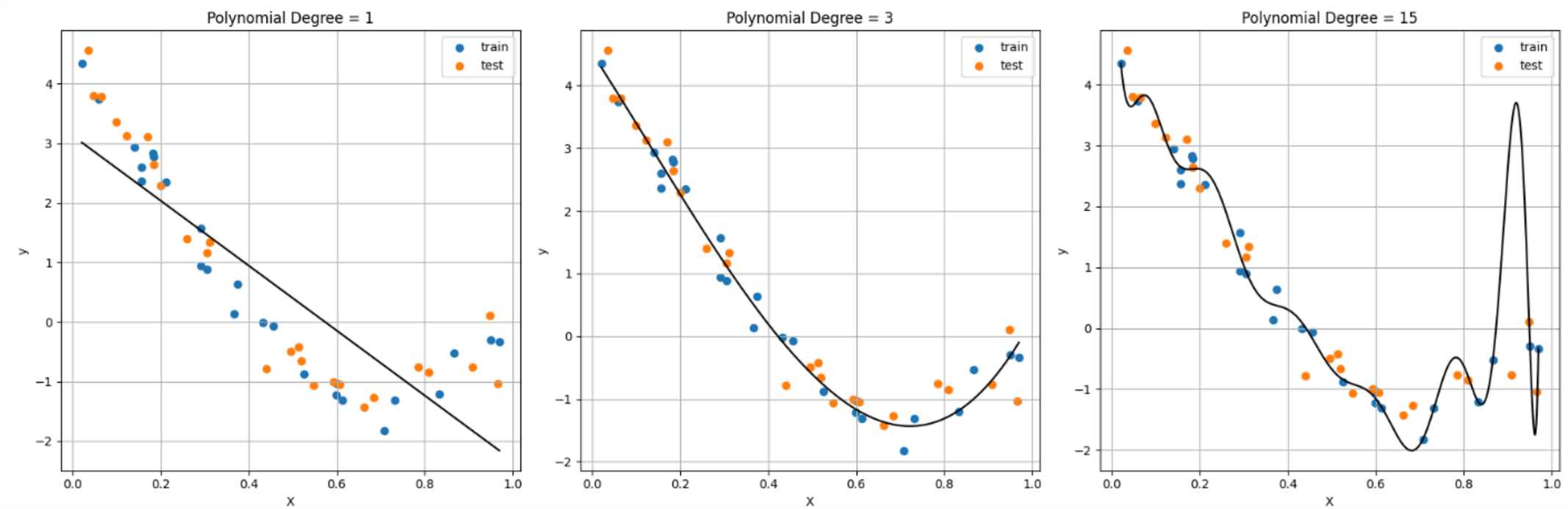
Недообучение и переобучение

Недообучение и переобучение

Недообучение

Оптимально

Переобучение



Недообучение и переобучение

- недообучена – данных слишком много / параметров недостаточно / модель простая, негибкая;
- переобучена – данных слишком мало / параметров слишком много / модель сложная, избыточно гибкая.

Из-за чего возникает переобучение?

- избыточная сложность пространства параметров W позволяет черезмерно точно подстроиться под обучающую выборку;

- выбор модели $f \in$ F производится по неполной информации X.

Переобучение есть всегда, когда оптимизация идет по конечной выборке.

Как обнаружить переобучение?

Эмпирически, путем разбиения выборки на train и test наборы.

https://pythonru.com/baza-znanij/sklearn-train-test-split

Как минимизировать переобучение?

- увеличить обучающую выборку;
- использовать класс более "простых" моделей, то есть выбирать их по обобщающей способности;
- накладывать ограничение на параметры w модели – регуляризация.

HO, LOO, CV

Обобщающая способность метода μ характеризуется величиной $L(\mu(X), X', y)$, где выборки X и X' получены из одного и того же неизвестного распределения.

Существуют эмпирические оценки обобщающей способности:

- Отложенная выборка (hold-out), $X = X_{train} \cup X_{test}$:

$$HO(\mu, X_{train}, X_{test}) = L(\mu(X_{train}), X_{test}, y) \to \min;$$

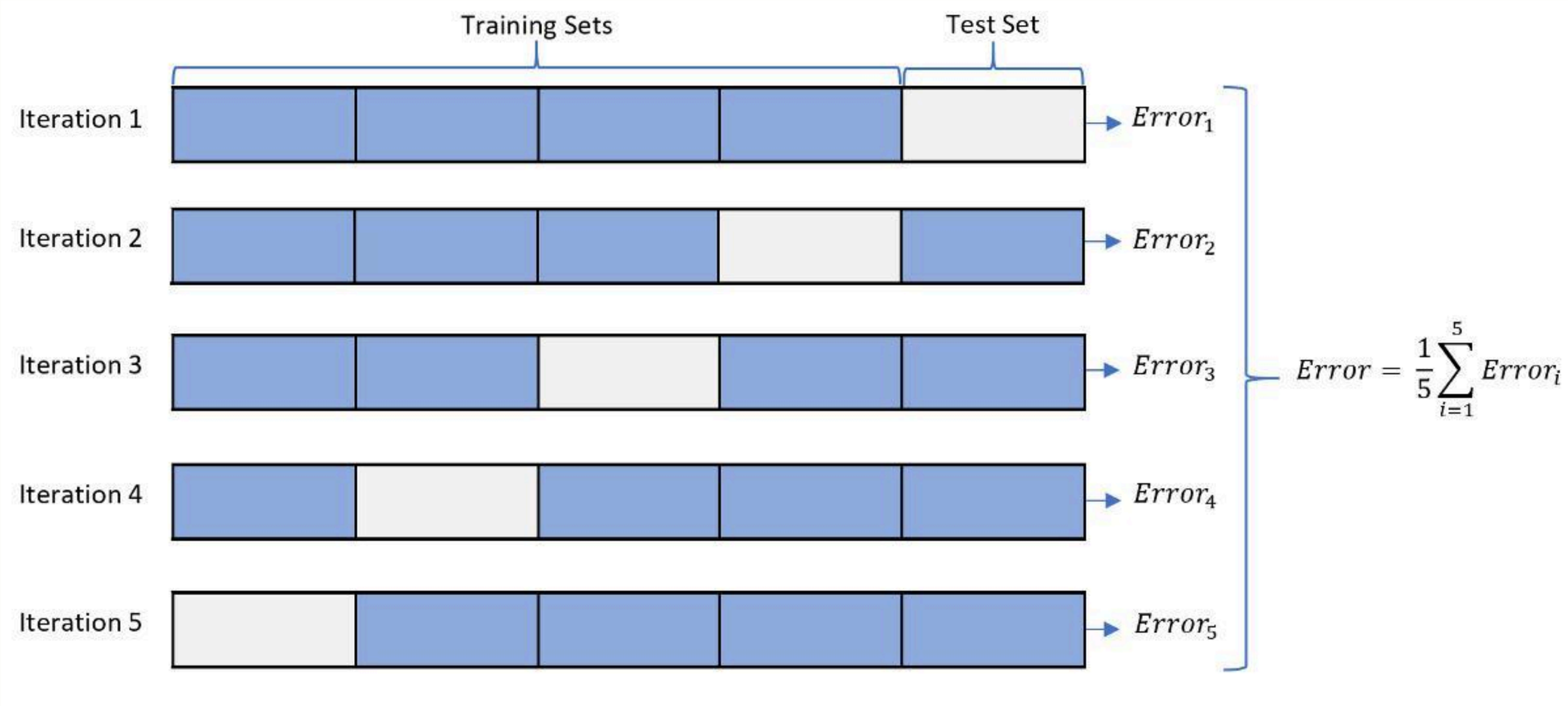
- Скользящий контроль (leave-one-out):

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

- Кросс-проверка (cross-validation), $X = X_1 \cup X_2 \cup \ldots \cup X_k$:

$$CV(\mu, X) = \frac{1}{k} \sum_{i=1}^{k} L(\mu(X \setminus X_k), X_k, y) \to \min\,.$$

Cross-validation



https://futureinapps.com/ru/blog/kross-validaciya-v-mashinnom-obuchenii-chast-1?pid=1kjAlrVFMI

Что сегодня рассмотрели?

- Обучение с учителем и без учителя
- Регрессия, классификация
- Матрица объект-признак
- Модели машинного обучения
- Эмпиричческий риск, функция потерь
- Недообучение и переобучение
- Hold-out, Leave-one-out, Cross-validation