Seminar 2. Cross-Validation. Pandas.

Valentin Malykh

MIPT

16/02/2016
Moscow

Table of content

Task of machine learning

X — set of objects
Y — set of labels
$y : X \to Y$ — target function

$\{x_1, \ldots, x_\ell\} \subset X$ — training sample
$y_i = y(x_i)$ — known answers

The goal is to find:
$a : X \to Y$ — algorithm or decision function approximating y on Y

Types of tasks

Classification

- $Y = \{-1, +1\}$ — two classes (binary)
- $Y = \{1, \ldots, M\}$ — multi-label classification (case A)
- $Y = \{0, 1\}^M$ — multi-label classification (case B)

Regression

- $Y = R$ or $Y = R^m$ — continuous space of Y

Ranking task

- Y — finite ordered space

Models and algorithms

Predictive model — parametric family of functions:

$$A = \{g(X, \theta | \theta \in \Theta)\},$$

where $g : X \times \Theta \to Y$ — some defined function,
Θ — set of allowable values of θ.

Learning algorithm is mapping $\mu : \{X \times Y\}^{\ell} \to A$, where $X = \{x_i, y_i\}_{i=1}^{\ell}$ and $a \in A$.

Loss function

$\mathcal{L}(a, x)$ — error value of algorithm $a \in A$ on object $x \in X$.

For classification:
$\mathcal{L}(a, x) = [a(x) \neq y(x)]$ — indicator

For regression:
$\mathcal{L}(a, x) = |a(x) - y(x)|$
$\mathcal{L}(a, x) = (a(x) - y(x))^2$

Empirical risk:

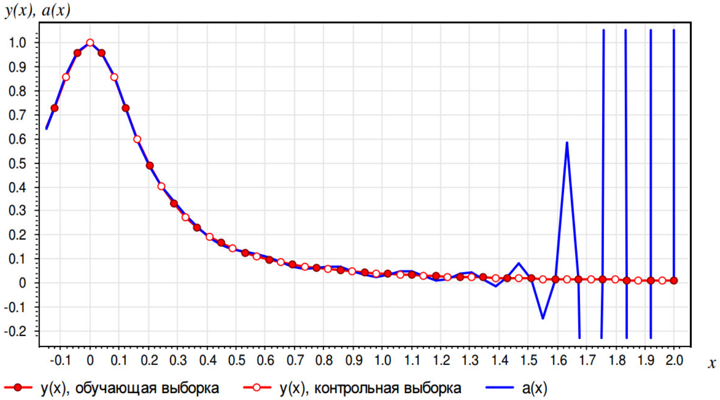
$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{l} \mathcal{L}(a, x_i).$$

Overfitting

- **Why?**
  — Redundant complexity of Θ
  — Finite size of training sample
- **How to detect?**
  — Split data on *training* and *test* sets
- **How to minimize?**
  — restrictions on θ
  — minimize theoretical estimation
  — minimize (carefully!!!) cross-validation estimations

Overfitting example

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



CV procedure

- Splitting $X = \{x_i, y_i\}_{i=1}^{l} = X_n^m \cup X_n^k$ on 2 parts in N different ways $(k + m = l)$
- For each $n \in \{1, \ldots, N\}$ train $a_n = \mu(X_n^m)$. Then calculate quality measure $Q_n = Q(a_n, X_n^k)$
- $CV(\mu, X^l) = \frac{1}{N} \sum_{n=1}^{N} Q(a_n, X_n^k)$

Types of CV

- **Complete CV**
  CV for all C_ℓ^k partitions for some k.
- **Random partition CV**
  CV for some number of random partitions from C_ℓ^k.
- **Leave-one-out CV (LOO)**
  Complete CV for $k = 1 \Rightarrow N = l$.

Types of CV

- **Hold-out CV**
  CV for one random partitions for some k, $N = 1$.
  Not the same with control on test set!!!

- q-**fold CV**
  For $k_1, \dots, k_q$, $k_1 + \dots + k_q = l$:
  $$X^l = X_1^{k_1} \cup \dots \cup X_q^{k_q}.$$
  Then $CV(\mu, X^l) = \frac{1}{q} \sum_{n=1}^{q} Q(\mu(X^l \setminus X_n^{k_n}), X_n^k)$

- $r \times q$-**fold CV**
  r iterations of q-fold CV.